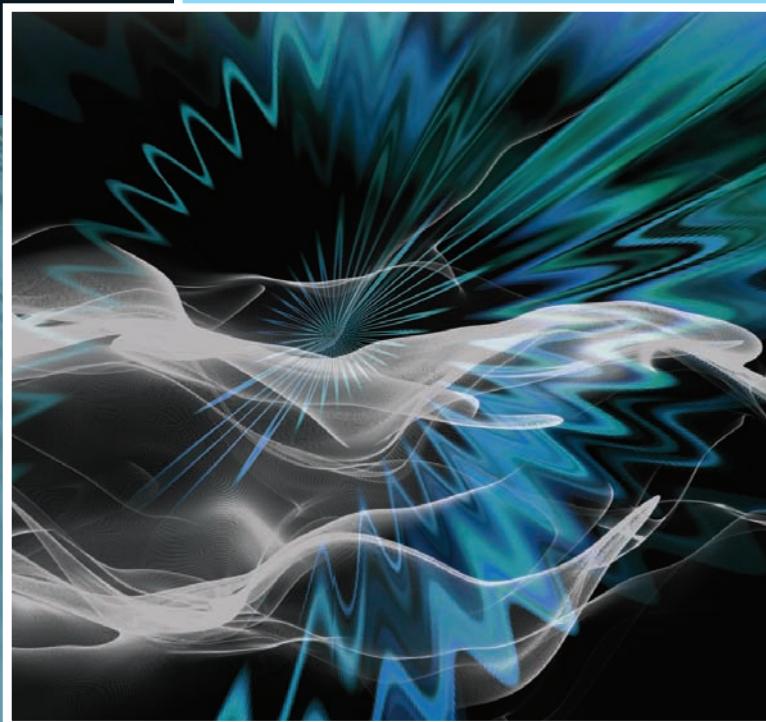


Circulation of this
edition outside the
Indian subcontinent is
UNAUTHORIZED

Electronic Communications



Fourth Edition

Dennis Roddy
John Coolen

Electronic Communications

This page is intentionally left blank.

Electronic Communications

Fourth Edition

Dennis Roddy
John Coolen
Lakehead University, Ontario

PEARSON

Copyright © 2014 Dorling Kindersley (India) Pvt. Ltd.

Licensees of Pearson Education in South Asia

No part of this eBook may be used or reproduced in any manner whatsoever without the publisher's prior written consent.

This eBook may or may not include all assets that were part of the print version. The publisher reserves the right to remove any material in this eBook at any time.

ISBN 9788177585582

eISBN 9789332538030

Head Office: A-8(A), Sector 62, Knowledge Boulevard, 7th Floor, NOIDA 201 309, India

Registered Office: 11 Local Shopping Centre, Panchsheel Park, New Delhi 110 017, India

Preface

The major additions to the fourth edition are new chapters on digital signals and digital communications. Many of the other chapters have been considerably expanded, and new problem sets have been added. Mathcad¹ has been introduced as a tool for problem solving, but the problems are formulated in such a way that these can be solved using other computer packages, or a good calculator. Although there are many powerful software packages available for circuit and system analysis, most are too highly specialized to be included here. A more general program such as Mathcad requires the user to be able to formulate the physical concepts in mathematical terms, which requires a good understanding of the underlying theory. Also, algebraic manipulation, with its attendant errors, can usually be avoided, as can simplifying assumptions which limit accuracy in certain instances.

The text is intended for use in the final year of technology programs in telecommunications. A one-term course covering the basics of communications systems might include material from chapters 1, 2, 3, 4, 8, 9, 10, 11, and 12. Material of a more difficult nature, such as that on the fast Fourier transform, could be omitted without breaking the continuity, and instructors who wished to include more on receiver principles might choose to make their own selection from these chapters and from chapters 5, 6, and 7. A second one-term course on transmission and propagation could be based on chapters 13, 14, 15, 16, with material dealing with systems being selected from the remaining chapters 17, 18, 19, and 20. There is more material in this book than can be reasonably covered in the final year of a technology program: The aim has been to make it useful as a reference text for technology graduates in the work force, as well as for students currently in college programs. It is also hoped that the book will provide a useful “bridging text” for those graduate technologists who continue with engineering degree studies.

The authors would like to thank the following reviewers for their valuable suggestions and input: Donald Stenz, Milwaukee School of Engineering; Alvis Evans, Tarrant County Jr. College; Allan Smith, Louisiana Tech University; Dr. Lester Johnson, Savannah State College; Warren Foxwell, DeVry Institute—Lombard; Donald Hill, RETs Electronics Institute; Shakti Chatterjee, DeVry Institute—Columbus; and Hassan Moghbelli, Purdue University—Calumet.

Dennis Roddy
John Coolen

The publishers would like to thank K. C. Raveendranathan, Professor and Head, Department of Electronics and Communication, Government Engineering College, Barton Hill, Trivandrum, for his valuable suggestions and inputs in enhancing the content of this book to suit the requirements of Indian universities.

¹ Mathcad is a trademark of Mathsoft Inc.

This page is intentionally left blank.

Contents

1 PASSIVE CIRCUITS, 1

- 1.1 Introduction, 1
- 1.2 Attenuator Pads, 1
- 1.3 Series Tuned Circuit, 9
- 1.4 Parallel Tuned Circuit, 15
- 1.5 Self-capacitance of a Coil, 17
- 1.6 Skin Effect, 19
- 1.7 Mutual Inductance, 20
- 1.8 High-frequency Transformers, 22
- 1.9 Tapped Inductor, 27
- 1.10 Capacitive Tap, 29
- 1.11 Maximum Power Transfer and Impedance Matching, 32
- 1.12 Low-frequency Transformers, 34
- 1.13 Passive Filters, 36
- Problems 46

2 WAVEFORM SPECTRA, 51

- 2.1 Introduction, 51
- 2.2 Sinusoidal Waveforms, 51
- 2.3 General Periodic Waveforms, 53
- 2.4 Trigonometric Fourier Series for a Periodic Waveform, 53
- 2.5 Fourier Coefficients, 55
- 2.6 Spectrum for the Trigonometric Fourier Series, 55
- 2.7 Rectangular Waves, 56
- 2.8 Sawtooth Waveform, 59
- 2.9 Pulse Train, 59
- 2.10 Some General Properties of Periodic Waveforms, 62
- 2.11 Exponential Fourier Series, 62
- 2.12 Approximate Formulas for the Fourier Coefficients, 64
- 2.13 Energy Signals and Fourier Transforms, 66

- 2.14 Fast Fourier Transform, 68
- 2.15 Inverse Fast Fourier Transform, 71
- 2.16 Filtering of Signals, 73
- 2.17 Power Signals, 74
- 2.18 Bandwidth Requirements for Analog Information Signals, 76
- Problems, 77

3 DIGITAL LINE WAVEFORMS, 82

- 3.1 Introduction, 82
- 3.2 Symbols, Binitis, Bits, and Bauds, 82
- 3.3 Functional Notation for Pulses, 84
- 3.4 Line Codes and Waveforms, 85
- 3.5 *M*-ary Encoding, 95
- 3.6 Intersymbol Interference, 96
- 3.7 Pulse Shaping, 96
- Problems, 101

4 NOISE, 105

- 4.1 Introduction, 105
- 4.2 Thermal Noise, 105
- 4.3 Shot Noise, 116
- 4.4 Partition Noise, 116
- 4.5 Low Frequency or Flicker Noise, 116
- 4.6 Burst Noise, 117
- 4.7 Avalanche Noise, 117
- 4.8 Bipolar Transistor Noise, 118
- 4.9 Field-effect Transistor Noise, 118
- 4.10 Equivalent Input Noise Generators and Comparison of BJTs and FETs, 118
- 4.11 Signal-to-noise Ratio, 120

4.12	S/N Ratio of a Tandem Connection, 120	6.6	Voltage-controlled Oscillators (VCOs), 186
4.13	Noise Factor, 122	6.7	Stability, 191
4.14	Amplifier Input Noise in Terms of F , 124	6.8	Frequency Synthesizers, 193 Problems 196
4.15	Noise Factor of Amplifiers in Cascade, 124	7	RECEIVERS, 198
4.16	Noise Factor and Equivalent Input Noise Generators, 126	7.1	Introduction, 198
4.17	Noise Factor of a Lossy Network, 127	7.2	Superheterodyne Receivers, 198
4.18	Noise Temperature, 128	7.3	Tuning Range, 200
4.19	Measurement of Noise Temperature and Noise Factor, 129	7.4	Tracking, 201
4.20	Narrowband Band-pass Noise, 130 Problems, 132	7.5	Sensitivity and Gain, 205
5	TUNED SMALL-SIGNAL AMPLIFIERS, MIXERS, AND ACTIVE FILTERS, 136	7.6	Image Rejection, 206
5.1	Introduction, 136	7.7	Spurious Responses, 208
5.2	The Hybrid- π Equivalent Circuit for the BJT, 136	7.8	Adjacent Channel Selectivity, 210
5.3	Short-circuit Current Gain for the BJT, 138	7.9	Automatic Gain Control (AGC), 212
5.4	Common-emitter (CE) Amplifier, 141	7.10	Double Conversion, 214
5.5	Stability and Neutralization, 150	7.11	Electronically Tuned Receivers (ETRs), 216
5.6	Common-base Amplifier, 151	7.12	Integrated-circuit Receivers, 218 Problems 221
5.7	Available Power Gain, 153	8	AMPLITUDE MODULATION, 223
5.8	Cascode Amplifier, 154	8.1	Introduction, 223
5.9	Hybrid- π Equivalent Circuit for an FET, 155	8.2	Amplitude Modulation, 224
5.10	Mixer Circuits, 158	8.3	Amplitude Modulation Index, 225
5.11	Active Filters, 163 Problems, 169	8.4	Modulation Index for Sinusoidal AM, 228
6	OSCILLATORS, 173	8.5	Frequency Spectrum for Sinusoidal AM, 228
6.1	Introduction, 173	8.6	Average Power for Sinusoidal AM, 231
6.2	Amplification and Positive Feedback, 173	8.7	Effective Voltage and Current for Sinusoidal AM, 232
6.3	<i>RC</i> Phase Shift Oscillators, 175	8.8	Nonsinusoidal Modulation, 233
6.4	<i>LC</i> Oscillators, 178	8.9	Double-sideband Suppressed Carrier (DSBSC) Modulation, 235
6.5	Crystal Oscillators, 185	8.10	Amplitude Modulator Circuits, 236
		8.11	Amplitude Demodulator Circuits, 240
		8.12	Amplitude-modulated Transmitters, 244
		8.13	AM Receivers, 247
		8.14	Noise in AM Systems, 252 Problems, 257

9	SINGLE-SIDEBAND MODULATION, 262		
9.1	Introduction, 262	11.3	Pulse Code Modulation (PCM) 341
9.2	Single-sideband Principles, 262	11.4	Pulse Frequency Modulation (PFM), 356
9.3	Balanced Modulators, 264	11.5	Pulse Time Modulation (PTM), 357
9.4	SSB Generation, 267	11.6	Pulse Position Modulation (PPM), 357
9.5	SSB Reception, 271	11.7	Pulse Width Modulation (PWM), 358
9.6	Modified SSB Systems, 273		Problems, 359
9.7	Signal-to-noise Ratio for SSB, 278		
9.8	Companded Single Sideband, 280		
	Problems, 280		
10	ANGLE MODULATION, 283	12	DIGITAL COMMUNICATIONS, 361
10.1	Introduction, 283	12.1	Introduction, 361
10.2	Frequency Modulation, 283	12.2	Synchronization, 362
10.3	Sinusoidal FM, 285	12.3	Asynchronous Transmission, 362
10.4	Frequency Spectrum for Sinusoidal FM, 287	12.4	Probability of Bit Error in Baseband Transmission, 364
10.5	Average Power in Sinusoidal FM, 291	12.5	Matched Filter, 368
10.6	Non-sinusoidal Modulation: Deviation Ratio, 292	12.6	Optimum Terminal Filters, 372
10.7	Measurement of Modulation Index for Sinusoidal FM, 293	12.7	Bit-timing Recovery, 372
10.8	Phase Modulation, 293	12.8	Eye Diagrams, 374
10.9	Equivalence between PM and FM, 294	12.9	Digital Carrier Systems, 375
10.10	Sinusoidal Phase Modulation, 296	12.10	Carrier Recovery Circuits, 386
10.11	Digital Phase Modulation, 297	12.11	Differential Phase Shift Keying (DPSK), 388
10.12	Angle Modulator Circuits, 297	12.12	Hard and Soft Decision Decoders, 390
10.13	FM Transmitters, 305	12.13	Error Control Coding, 390
10.14	Angle Modulation Detectors, 309		Problems, 403
10.15	Automatic Frequency Control, 318		
10.16	Amplitude Limiters, 319	13	TRANSMISSION LINES AND CABLES, 407
10.17	Noise in FM Systems, 320	13.1	Introduction, 407
10.18	Pre-emphasis and De-emphasis 324	13.2	Primary Line Constants, 408
10.19	FM Broadcast Receivers, 325	13.3	Phase Velocity and Line Wavelength, 409
10.20	FM Stereo Receivers, 328	13.4	Characteristic Impedance, 410
	Problems, 330	13.5	Propagation Coefficient, 412
11	PULSE MODULATION, 336	13.6	Phase and Group Velocities, 415
11.1	Introduction, 336	13.7	Standing Waves, 417
11.2	Pulse Amplitude Modulation (PAM), 336	13.8	Lossless Lines at Radio Frequencies, 419
		13.9	Voltage Standing-wave Ratio, 420
		13.10	Slotted-line Measurements at Radio Frequencies, 421

13.11	Transmission Lines as Circuit Elements, 424	16.10	Hertzian Dipole, 518
13.12	Smith Chart, 428	16.11	Half-wave Dipole, 520
13.13	Time-domain Reflectometry, 438	16.12	Vertical Antennas, 523
13.14	Telephone Lines and Cables, 440	16.13	Folded Elements, 526
13.15	Radio-frequency Lines, 443	16.14	Loop and Ferrite-rod Receiving Antennas, 527
13.16	Microstrip Transmission Lines, 443	16.15	Nonresonant Antennas, 529
13.17	Use of Mathcad in Transmission Line Calculations, 446	16.16	Driven Arrays, 530
	Problems, 450	16.17	Parasitic Arrays, 534
14	WAVEGUIDES, 453	16.18	VHF–UHF Antennas, 536
14.1	Introduction, 453	16.19	Microwave Antennas, 538 Problems, 545
14.2	Rectangular Waveguides, 453		
14.3	Other Modes, 464	17	TELEPHONE SYSTEMS, 548
	Problems, 467	17.1	Wire Telephony, 548
15	RADIO-WAVE PROPAGATION, 468	17.2	Public Telephone Network, 561 Problems, 573
15.1	Introduction, 468	18	FACSIMILE AND TELEVISION, 576
15.2	Propagation in Free Space, 468	18.1	Introduction, 576
15.3	Tropospheric Propagation, 473	18.2	Facsimile Transmission, 576
15.4	Ionospheric Propagation, 482	18.3	Television, 593
15.5	Surface Wave, 493	18.4	Television Signal, 606
15.6	Low Frequency Propagation and Very Low Frequency Propagation, 495	18.5	Television Receivers, 608
15.7	Extremely Low Frequency Propagation, 498	18.6	Television Transmitters, 612
15.8	Summary of Radio-wave Propagation, 503	18.7	High-definition Television, 614 Problems, 618
	Problems, 503	19	SATELLITE COMMUNICATIONS, 620
16	ANTENNAS, 505	19.1	Introduction, 620
16.1	Introduction, 505	19.2	Kepler's First Law, 620
16.2	Antenna Equivalent Circuits, 505	19.3	Kepler's Second Law, 621
16.3	Coordinate System, 509	19.4	Kepler's Third Law, 622
16.4	Radiation Fields, 510	19.5	Orbits, 622
16.5	Polarization, 510	19.6	Geostationary Orbit, 623
16.6	Isotropic Radiator, 512	19.7	Power Systems, 624
16.7	Power Gain of an Antenna, 513	19.8	Attitude Control, 624
16.8	Effective Area of an Antenna, 515	19.9	Satellite Station Keeping, 626
16.9	Effective Length of an Antenna, 516	19.10	Antenna Look Angles, 627
		19.11	Limits of Visibility, 635
		19.12	Frequency Plans and Polarization, 637
		19.13	Transponders, 638

19.14	Uplink Power Budget Calculations, 642	20.3	Losses in Fibers, 668
19.15	Downlink Power Budget Calculations, 646	20.4	Dispersion, 673
19.16	Overall Link Budget Calculations, 647	20.5	Light Sources for Fiber Optics, 682
19.17	Digital Carrier Transmission, 648	20.6	Photodetectors, 691
19.18	Multiple-access Methods, 649 Problems, 650	20.7	Connectors and Splices, 694
		20.8	Fiber-optic Communication Link, 698 Problems, 701
20	FIBER-OPTIC COMMUNICATIONS, 654		
20.1	Introduction, 654	A	Logarithmic Units, 704
20.2	Principles of Light Transmission in a Fiber, 654	B	The Transverse Electromagnetic Wave, 709
			INDEX, 713

APPENDIX

A	Logarithmic Units, 704
B	The Transverse Electromagnetic Wave, 709

This page is intentionally left blank.



Passive Circuits

1.1 Introduction

A *passive electric network* is defined in the *IEEE Standard Dictionary of Electrical and Electronic Terms* as an electric network containing no source of energy. Passive networks contain resistors, inductors, and capacitors connected in various ways. The properties of the network are independent of the energy sources energizing the network.

In this chapter, circuits that are of particular relevance to electronic communications are examined. It is assumed that the student has a thorough understanding of ac and dc circuit theory.

An objective of the chapter is to explain how circuits function, and for this reason specific computer packages for circuit analysis have been avoided. Even so, it should be recognized that the computer approach to circuit analysis generally eliminates much algebraic manipulation, which is a common source of errors.

Examples and problems are set up so that they also can be solved without the aid of specific computer packages or programs, but the student will find that a programmable calculator or a personal computer, along with a versatile program such as Mathcad/MATLAB, is an extremely useful tool for problem solving.

1.2 Attenuator Pads

An *attenuator pad* is a resistive network that is used to introduce a fixed amount of attenuation between a source and a load. Referring to Fig. 1.2.1, let I_{LO} represent the load current without the network inserted, and I_L the load current with the network inserted; then the *insertion loss* of the network is defined as the ratio I_L/I_{LO} .

In addition to providing attenuation of the signal, the pad usually has to provide input and output matching. Again referring to Fig. 1.2.1, this means that the input resistance R_{IN} must be equal to the source resistance R_S , and the output resistance R_{OUT} must be equal to the load resistance R_L . In evaluating R_{IN} the load must be connected, and in evaluating R_{OUT} the source must be connected.

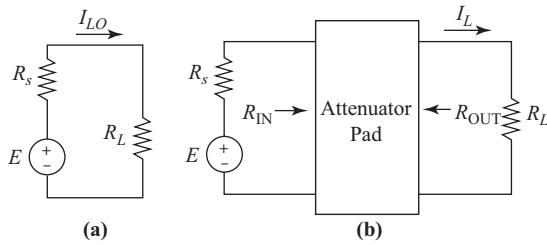


Figure 1.2.1 Source and load (a) connected directly and (b) connected through an attenuator pad.

From Fig. 1.2.1(a) the reference load current I_{LO} is seen to be given by

$$I_{LO} = \frac{E}{R_S + R_L} \quad (1.2.1)$$

The quantities I_L , R_{IN} , and R_{OUT} shown in Fig. 1.2.1(b) have to be evaluated for specific attenuator circuits. Commonly employed circuits are the T-attenuator and the pi-attenuator, analyzed in the following sections.

The T-Attenuator

The network resistors that make up a T-attenuator are shown as R_1 , R_2 , and R_3 in Fig. 1.2.2. The name T-attenuator arises because the circuit is configured like the letter T. Applying Kirchhoff's voltage law to the loop consisting of the source and R_1 and R_2 yields

$$E = I_1 \cdot (R_S + R_1 + R_3) - I_L \cdot R_3 \quad (1.2.2)$$

Applying Kirchhoff's voltage law to the loop consisting of R_2 , R_L , and R_3 yields

$$0 = -I_1 \cdot R_3 + I_L \cdot (R_2 + R_3 + R_L) \quad (1.2.3)$$

Equations (1.2.2) and (1.2.3) may be solved for I_L to give

$$I_L = \frac{E \cdot R_3}{(R_S + R_1 + R_3) \cdot (R_2 + R_3 + R_L) - R_3^2} \quad (1.2.4)$$

Combining Eqs. (1.2.1) and (1.2.4) gives for the insertion loss

$$\begin{aligned} I_L &= \frac{I_L}{I_{LO}} \\ &= \frac{R_3 \cdot (R_S + R_L)}{(R_S + R_1 + R_3) \cdot (R_2 + R_3 + R_L) - R_3^2} \end{aligned} \quad (1.2.5)$$

The insertion loss (IL) is usually quoted in decibels. This will be denoted by IL dB, which by definition is

$$\text{IL dB} = -20 \log_{10} \text{IL} \quad (1.2.6)$$

The negative sign is to show that attenuation occurs; that is, the insertion loss will come out as a positive number of decibels.

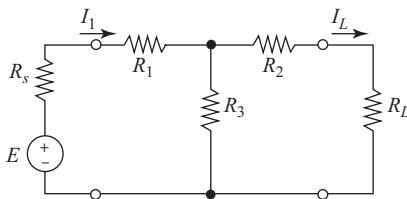


Figure 1.2.2 T-attenuator.

The equations for input resistance and output resistance are found from inspection of Fig. 1.2.2. For the input resistance, the source is removed and the load left connected. The input resistance is that seen looking to the right into the network:

$$R_{IN} = R_1 + \frac{R_3 \cdot (R_2 + R_L)}{(R_2 + R_3 + R_L)} \quad (1.2.7)$$

For the output resistance, the load is removed and the source left connected (with the source emf at zero). The output resistance is that seen looking to the left into the network:

$$R_{OUT} = R_2 + \frac{R_3 \cdot (R_1 + R_S)}{(R_1 + R_3 + R_S)} \quad (1.2.8)$$

Given the problem of designing an attenuator to meet specified values of input resistance, output resistance, and insertion loss, Eqs. (1.2.5), (1.2.7), and (1.2.8) can be solved for R_1 , R_2 , and R_3 . Because the insertion loss is usually quoted in decibels, it must first be converted to a current ratio using Eq. (1.2.6) before being substituted into the equations.

Unfortunately, explicit expressions for R_1 , R_2 , and R_3 cannot be obtained, and numerical methods must be employed. Furthermore, not all combinations of input resistance, output resistance, and insertion loss can be met in a given attenuator design, and where the design fails, some of the resistor values will come out negative.

EXAMPLE 1.2.1

Determine the *resistor* values for a T-attenuator that must provide 14-dB attenuation between a $75\text{-}\Omega$ source and a $50\text{-}\Omega$ load. The attenuator must also provide input and output matching.

SOLUTION From Eq. (1.2.6), the insertion loss is

$$IL = 10^{-0.7} = 0.2$$

The three equations to be solved are then

$$0.2 = \frac{R_3 \cdot (R_S + R_L)}{(R_S + R_1 + R_3) \cdot (R_2 + R_3 + R_L) - R_3^2}$$

$$75 = R_1 + \frac{R_3 \cdot (R_2 + R_L)}{(R_2 + R_3 + R_L)}$$

$$50 = R_2 + \frac{R_3 \cdot (R_1 + R_S)}{(R_1 + R_3 + R_S)}$$

Numerical methods must be employed, for example, using a computer or programmable calculator. The results obtained using Mathcad are $R_1 = 56\Omega$, $R_2 = 29\Omega$, and $R_3 = 25\Omega$, respectively, the values being rounded off to integer values.

Exercise 1.2.1 Repeat the preceding calculations for an insertion loss of 1 dB.

(Ans. $R_1 = 106\Omega$, $R_2 = -79\Omega$, and $R_3 = 450\Omega$.)

In the preceding exercise, the negative value for R_2 shows that the attenuator is physically unrealizable.

Although the attenuator design in general is best achieved by computational methods, as shown by the preceding results, there is one particular case of practical importance that allows for an analytical solution. This is when the attenuator has to provide a specified attenuation between a matched source and load. This results in a symmetrical attenuator in which $R_1 = R_2$ and $R_{IN} = R_{OUT}$. Let $R_1 = R_2 = R$ and $R_{IN} = R_{OUT} = R_o$. Then either Eq. (1.2.7) or (1.2.8) gives

$$R_o = R + \frac{R_3(R + R_o)}{R_3 + R + R_o} \quad (1.2.9)$$

From this, we have

$$\frac{R_o - R}{R + R_o} = \frac{R_3}{R_3 + R + R_o} \quad (1.2.10)$$

From Eq. (1.2.1), $I_{LO} = E/2R_o$, and because $R_{IN} = R_o$, the input current to the attenuator is also equal to I_{LO} . Applying the current divider rule to the center node gives $I_L = I_{LO}R_3/(R_3 + R + R_o)$, and therefore the insertion loss IL is

$$IL = \frac{R_3}{R_3 + R + R_o} \quad (1.2.11)$$

The right-hand side of this is seen to be the same as that for Eq. (1.2.10), and therefore

$$\frac{R_o - R}{R + R_o} = IL \quad (1.2.12)$$

$$\therefore R = R_o \frac{1 - IL}{1 + IL} \quad (1.2.13)$$

This allows the resistor R to be determined for a given insertion loss. Once R is known, R_3 can be determined from Eq. (1.2.11), and it is left as an exercise for the student to show that

$$R_3 = \frac{2R_o(IL)}{1 - (IL)^2} \quad (1.2.14)$$

EXAMPLE 1.2.2

A T-type attenuator is required to provide a 6-dB insertion loss and to match 50- Ω input and output. Find the resistor values.

SOLUTION From Eq. (1.2.6), 6 dB gives an insertion loss ratio of 0.5:1. Therefore, from Eq. (1.2.13),

$$R = 50 \times \frac{1 - 0.5}{1 + 0.5} = 16.67 \Omega$$

From Eq. (1.2.14),

$$R_3 = \frac{100 \times 0.5}{1 - 0.5^2} = 66.67 \Omega$$

The Pi-Attenuator

The network resistors making up the pi-attenuator are shown as R_A , R_B , and R_C in Figure 1.2.3. The name pi-attenuator arises because the circuit is configured like the Greek letter π . A direct analysis of the circuit may be carried out to find the input resistance, output resistance, and insertion loss in terms of the network resistors. Alternatively, Eqs. (1.2.5), (1.2.7), and (1.2.8) may be transformed through *duality* to give, for the pi-network,

$$\text{IL} = G_C \cdot \frac{G_S + G_L}{(G_S + G_A + G_C) \cdot (G_B + G_C + G_L) - G_C^2} \quad (1.2.15)$$

$$G_{\text{IN}} = G_A + \frac{G_C \cdot (G_B + G_L)}{G_B + G_C + G_L} \quad (1.2.16)$$

$$G_{\text{OUT}} = G_B + \frac{G_C \cdot (G_A + G_S)}{G_A + G_C + G_S} \quad (1.2.17)$$

In these equations, conductance G is the reciprocal of resistance R ; thus $G_C = 1/R_C$, and so on. Although the equations are most easily solved in terms of conductance, resistor values will usually be specified. The computations therefore generally involve the extra steps of converting source resistance and load resistance into conductance values and, once the network conductances are found, converting these back into resistor values.

Alternatively, a T-attenuator may be designed to meet the specified values of insertion loss, input resistance, and output resistance, and the resulting R_1 , R_2 , and R_3 values converted to R_A , R_B , and R_C values using the Y- Δ transformation.

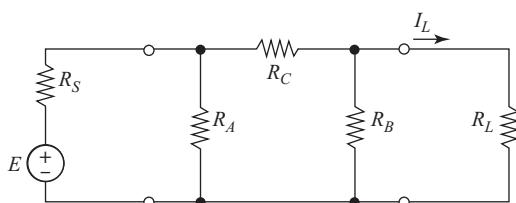


Figure 1.2.3 Pi-attenuator.

The equations obtained using the Y- Δ transformation are

$$R_A = \frac{R_1 R_2 + R_1 R_3 + R_2 R_3}{R_2} \quad (1.2.18)$$

$$R_B = R_A \frac{R_2}{R_1} \quad (1.2.19)$$

$$R_C = R_A \frac{R_2}{R_3} \quad (1.2.20)$$

There is no particular advantage favoring the π -configuration or the T-configuration, except in specific situations where one type may yield more practical values of resistors than the other. It also follows that if one type is physically unrealizable so also will be the other.

As with the T-attenuator, an analytical solution for the specific case of equal input and output resistances may be obtained. Denoting the insertion loss (as a current ratio, less than unity) as IL, and $R_S = R_L = R_o$, the resulting equations are

$$R_A = R_B = R_o \cdot \frac{1 + IL}{1 - IL} \quad (1.2.21)$$

$$R_C = R_o \cdot \frac{1 - (IL)^2}{2 \cdot (IL)} \quad (1.2.22)$$

EXAMPLE 1.2.3

A pi-attenuator is required to provide a 6-dB insertion loss and to match 50- Ω input and output. Find the resistor values.

SOLUTION From Eq. (1.2.6), an insertion loss of 6 dB gives $IL = 0.5$. From Eq. (1.2.21),

$$R_A = R_B = 50 \times \frac{1 + 0.5}{1 - 0.5} = 150 \Omega$$

From Eq. (1.2.22),

$$R_C = 50 \times \frac{1 - 0.5^2}{2 \times 0.5} = 37.5 \Omega$$

The L-Attenuator

The T- and pi-attenuators described so far are made up of three resistors. The value of each resistor can be chosen independently of the others, thus enabling the three design criteria of input resistance, output resistance, and insertion loss to be met. In many situations the only function of the pad is to provide matching between source and load, and although attenuation will be introduced, this may not be a critical design parameter. This allows a simpler type of pad to be designed, requiring only two resistors; it is known as an L-pad because the network configuration resembles an inverted letter L.

Figure 1.2.4 shows the L-attenuator, and it will be seen that this can be derived from either the T- or the pi-attenuator simply by the removal of one of the resistors. For convenience, it is assumed that the T-network

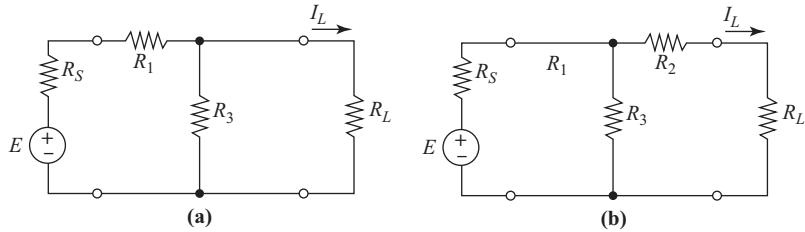


Figure 1.2.4 L-attenuator for (a) $R_S > R_L$ and (b) $R_S < R_L$.

forms the basis, so the resistor labels are the same as those used in the T-network, and the T-network equations can be used to evaluate these. Exactly the same resistor values would result by using the pi-network as the basis. As shown in Fig. 1.2.4, different configurations are required depending on whether $R_S > R_L$ or $R_S < R_L$.

Figure 1.2.4(a) shows the circuit required for the condition $R_S > R_L$. Inspection of the circuit shows that

$$R_S = R_{\text{in}} = R_1 + \frac{R_3 R_L}{R_3 + R_L}$$

So

$$R_S - R_1 = \frac{R_3 R_L}{R_3 + R_L}$$

and therefore

$$\frac{1}{R_S - R_1} = \frac{1}{R_3} + \frac{1}{R_L} \quad (1.2.23)$$

Also,

$$R_L = R_{\text{OUT}} = \frac{R_3(R_1 + R_S)}{R_3 + R_1 + R_S}$$

which gives

$$\frac{1}{R_L} = \frac{1}{R_3} + \frac{1}{R_1 + R_S} \quad (1.2.24)$$

The $1/R_3$ can be eliminated from Eqs. (1.2.23) and (1.2.24), giving

$$\therefore \frac{1}{R_S - R_1} + \frac{1}{R_S + R_1} = \frac{2}{R_L}$$

$$\therefore \frac{2R_S}{R_S^2 - R_1^2} = \frac{2}{R_L}$$

$$\therefore R_1^2 = R_S^2 - R_S R_L$$

or

$$R_1 = \sqrt{R_S(R_S - R_L)} \quad (1.2.25)$$

Adding Eqs. (1.2.23) and (1.2.24) and simplifying for R_3 gives

$$R_3 = \frac{R_S^2 - R_1^2}{R_1} \quad (1.2.26)$$

With $R_2 = 0$, Eq. (1.2.5) gives

$$IL = \frac{R_3 \cdot (R_S + R_L)}{(R_S + R_1 + R_3) \cdot (R_3 + R_L) - R_3^2} \quad (1.2.27)$$

These equations are for the situation shown in Fig. 1.2.4(a), where the source resistance is greater than the load resistance.

EXAMPLE 1.2.4

Design an L-attenuator to match a $75\text{-}\Omega$ source to a $50\text{-}\Omega$ load, and determine the insertion loss.

SOLUTION From Eq. (1.2.25),

$$R_1 = \sqrt{75 \times (75 - 50)} = 43.3 \Omega$$

From Eq. (1.2.26),

$$R_3 = \frac{75^2 - 43.3^2}{43.3} = 86.6 \Omega$$

From Eq. (1.2.27),

$$IL = \frac{86.6 \times (75 + 50)}{(75 + 43.3 + 86.6) \times (86.6 + 50) - 86.6^2} = 0.528$$

In decibels, this is $-20 \log 0.528 = 5.54 \text{ dB}$.

For the condition $R_S < R_L$, Fig. 1.2.4(b) applies. Inspection of the circuit yields

$$R_{IN} = \frac{R_3 \cdot (R_2 + R_L)}{R_2 + R_3 + R_L} \quad (1.2.28)$$

$$R_{OUT} = R_2 \frac{R_3 \cdot R_S}{R_3 + R_S} \quad (1.2.29)$$

These equations may be solved to yield

$$R_2 = \sqrt{R_L \cdot (R_L - R_S)} \quad (1.2.30)$$

$$R_3 = \frac{R_L^2 - R_2^2}{R_2} \quad (1.2.31)$$

Equation (1.2.30) is used to find R_2 , and then Eq. (1.2.31) to find R_3 . The insertion loss is obtained from Eq. (1.2.5), with $R_1 = 0$, as

$$\text{IL} = \frac{R_3 \cdot (R_S + R_L)}{(R_S + R_3) \cdot (R_2 + R_3 + R_L) - R_3^2} \quad (1.2.32)$$

EXAMPLE 1.2.5

Design an L-attenuator to match a $10\text{-}\Omega$ source to a $50\text{-}\Omega$ load, and determine the insertion loss.

SOLUTION From Eq. (1.2.30),

$$R_2 = \sqrt{50 \times (50 - 10)} = \mathbf{44.72\ \Omega}$$

and from Eq. (1.2.31)

$$R_3 = \frac{50^2 - 44.72^2}{44.72} = \mathbf{11.18\ \Omega}$$

From Eq. (1.2.32),

$$\text{IL} = \frac{11.18 \times (10 + 50)}{(10 + 11.18) \times (44.72 + 11.18 + 50) - 11.18^2} = 0.318$$

In decibels, this is $-20 \log 0.318 = \mathbf{9.95\ dB}$.

1.3 Series Tuned Circuit

Impedance of a Series Tuned Circuit

The series tuned circuit consists of a coil connected in series with a capacitor, as shown in Figure 1.3.1. Resistance r must be included since in a practical circuit there will always be resistance, mostly that of the coil.

Denoting by X the total reactance of the circuit, equal to $\omega L - 1/\omega C$, the impedance is given by

$$\begin{aligned} Z_s &= r + jX \\ &= r + j\left(\omega L - \frac{1}{\omega C}\right) \end{aligned} \quad (1.3.1)$$

The magnitude of the impedance is

$$|Z_s| = \sqrt{r^2 + X^2} \quad (1.3.2)$$

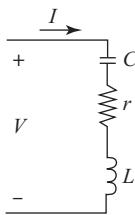


Figure 1.3.1 Series tuned circuit.

The phase angle of the impedance is

$$\phi_s = \arctan \frac{X}{r} \quad (1.3.3)$$

An examination of the impedance equation shows that at high frequencies such that $\omega L > 1/\omega C$ the inductive term dominates and X is positive. At low frequencies such that $\omega L < 1/\omega C$ the capacitive term dominates and X is negative. Figure 1.3.2 shows the impedance plot for a series circuit for which $C = 57 \text{ pF}$, $L = 263 \mu\text{H}$, and $r = 21.5 \Omega$.

Since X can vary from positive to negative, there must exist a frequency at which it is zero. This frequency is known as the *series resonant frequency*.

Series Resonant Frequency

Series resonance occurs when the reactive part of the impedance is zero or, equivalently, the phase angle is zero, as shown by Eq. (1.3.3). The magnitude of the impedance is a minimum at resonance, equal to r , from Eq. (1.3.2).

Denoting the series resonant frequency as $\omega_{so} = 2\pi f_{so}$, then for resonance

$$\omega_{so}L - \frac{1}{\omega_{so}C} = 0$$

from which

$$f_{so} = \frac{1}{2\pi\sqrt{LC}} \quad (1.3.4)$$

Equation (1.3.4) shows that by adjustment of either L or C (or both) the circuit can be brought into resonance with the applied frequency, a process known as *tuning*, and the circuit is also referred to as a *series tuned circuit*. The usefulness of the series tuned circuit is that it permits signals at one frequency to be selected in preference to those at other frequencies, a property referred to as *frequency selectivity*.

Series Q-Factor

The *Q-factor* (which stands for *quality factor*) can be defined as the ratio of inductive reactance at resonance to resistance in a tuned circuit. (The concept was originally applied to coils to indicate that a high reactance relative to resistance was desirable.) Normally, any series resistance associated with the capacitor in the tuned circuit is negligible, but, if significant, it is included in the total series resistance. The *Q-factor* can therefore be expressed as

$$Q_s = \frac{\omega_{so}L}{r} \quad (1.3.5)$$

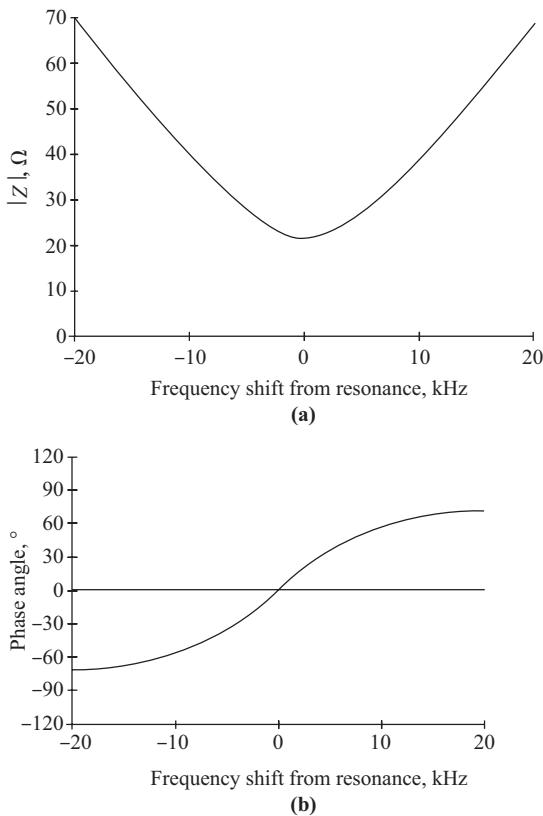


Figure 1.3.2 Impedance magnitude and phase angle as functions of frequency for $L = 263 \mu\text{H}$, $C = 57 \text{ pF}$, and $r = 21.5 \Omega$.

Since $\omega_{so}L = 1/\omega_{so}C$, the Q -factor can also be expressed as

$$Q_s = \frac{1}{\omega_{so}C_r} \quad (1.3.6)$$

The subscript s signifies series Q -factor. The Q -factor is an important parameter used in specifying the behavior of tuned circuits, so much so that instruments known as Q -meters are routinely used to measure Q . The Q -meter allows the Q -factor of a coil to be measured at a specific frequency and tuning capacitance.

Assuming that Q -meter measurements yield ω_{so} , C and Q , then L and r are easily found from Eqs. (1.3.5) and (1.3.6). Also, by combining Eqs. (1.3.4), (1.3.5), and (1.3.6), it is easily shown that

$$Q_s = \frac{1}{r} \cdot \sqrt{\frac{L}{C}} \quad (1.3.7)$$

The significance of Eq. (1.3.7) is that it shows Q_s is constant to the extent that L , C , and r are constant. This holds reasonably well for frequencies about resonance (but see Sections 1.5 and 1.6 for a discussion of ways in which Q may vary with frequency).

The Q -factor is also referred to as the *voltage magnification factor* because it gives the ratio of reactive voltage magnitude to applied voltage at resonance. This follows because the current at resonance is V/r , where

V is the applied voltage, and the magnitude of the voltage across L is $(V/r) \cdot \omega_{so}L = VQ$, and across C it is $(V/r) \cdot 1/\omega_{so}C = VQ$.

The magnitude of either reactive voltage is seen to be Q times the applied voltage, and this can reach comparatively high levels. Although the total reactive voltage at resonance is zero, it is possible to make use of the voltage magnification by coupling into the inductive or capacitive voltage separately, and use is made of this in filters and coupled circuits, as described later.

Note that the voltage rating of the reactive elements must take into account the expected high voltage at resonance, and that the inductive voltage is not the same as the voltage across the inductor, which includes the voltage across r .

Impedance in Terms of Q

Equation (1.3.1) for impedance can be rewritten as

$$Z_s = r \left(1 + j \left(\frac{\omega L}{r} - \frac{1}{\omega C_r} \right) \right)$$

L and C can be eliminated through the use of Eqs. (1.3.5) and (1.3.6) to give

$$Z_s = r \left(1 + j \left(\frac{\omega}{\omega_{so}} - \frac{\omega_{so}}{\omega} \right) Q_s \right) \quad (1.3.8)$$

Denoting the frequency variable by y ,

$$y = \frac{\omega}{\omega_{so}} - \frac{\omega_{so}}{\omega} \quad (1.3.9)$$

allows the impedance to be expressed as

$$Z_s = r(1 + jyQ_s) \quad (1.3.10)$$

$$|Z_s| = r\sqrt{1 + jyQ_s} \quad (1.3.11)$$

$$\phi_s = \tan^{-1} yQ_s \quad (1.3.12)$$

These impedance relationships enable the performance of the circuit to be readily gauged in terms of the Q -factor. The higher the Q -factor is, the greater the impedance magnitude at a given frequency off resonance and the sharper the phase change. The frequency selectivity of the circuit is also highly Q dependent.

Relative Response

The relative response of the circuit is the ratio of the current at any given frequency to the current at resonance. For a constant applied voltage V , the current in general is V/Z_s and at resonance it is V/r . Hence the relative response is

$$\begin{aligned} A_r &= \frac{r}{Z_s} \\ &= \frac{1}{1 + jyQ} \end{aligned} \quad (1.3.13)$$

The relative response determines the frequency selectivity of the circuit, which is how well it discriminates between wanted and unwanted signals. A measure of this is the -3dB bandwidth described in the following section.

Relative Response in Decibels

The relative response in decibels is the magnitude of A_r , expressed as a decibel voltage ratio:

$$\begin{aligned} A_r \text{ dB} &= 20 \log \frac{1}{\sqrt{1 + (yQ)^2}} \\ &= -10 \cdot \log(1 + (yQ)^2) \end{aligned} \quad (1.3.14)$$

The curve of Fig. 1.3.2 is plotted as a relative response curve in Fig. 1.3.3.

The -3dB Bandwidth

The main function of a tuned circuit is *frequency selection*, that is, the ability to select frequencies at or near resonance while rejecting other frequencies. A useful measure of the selectivity is the -3-dB bandwidth. This is the frequency band spanned by the -3-dB points on the resonance curve, as shown in Fig. 1.3.3.

At the -3-dB points, the magnitude of the relative response is $1/\sqrt{2}$, and comparing this with Eq. (1.3.13), it is seen that $2 = 1 + (y_3 Q_s)^2$ or $y_3 = \pm 1/Q_s$, where y_3 is y evaluated at the -3-dB frequencies. These are denoted as f'_3 and f''_3 in Fig. 1.3.3. This last expression written in full is

$$\frac{f_3}{f_{so}} - \frac{f_{so}}{f_3} = \pm \frac{1}{Q_s} \quad (1.3.15)$$

or

$$f_3^2 - f_{so}^2 = \pm \frac{f_{so} \cdot f_3}{Q_s}$$

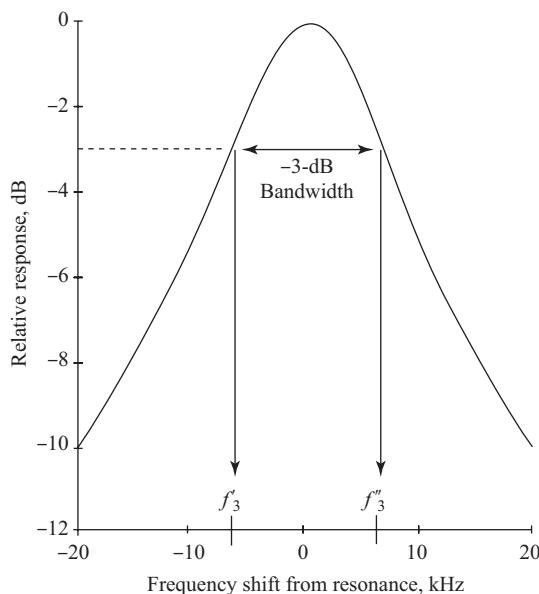


Figure 1.3.3 Relative response in decibels for a series tuned circuit.

Referring to Fig. 1.3.3, it is seen that f'_3 is less than f_{so} , and therefore

$$f'^2_3 - f^2_{so} = - \frac{f_{so} \cdot f'_3}{Q_s}$$

Also, f''_3 is seen to be greater than f_{so} , and therefore

$$f''^2_3 - f^2_{so} = + \frac{f_{so} \cdot f''_3}{Q_s}$$

Hence

$$f''^2_3 - f'^2_3 = \frac{f_{so}}{Q_s} (f''_3 + f'_3)$$

$$\therefore (f''_3 - f'_3)(f''_3 + f'_3) = \frac{f_{so}}{Q_s} (f''_3 + f'_3)$$

or

$$f''_3 - f'_3 = \frac{f_{so}}{Q_s} \quad (1.3.16)$$

However, as can be seen from Fig. 1.3.3, $f''_3 - f'_3$ is the -3 -dB bandwidth, and therefore

$$B_{3 \text{ dB}} = \frac{f_{so}}{Q_s} \quad (1.3.17)$$

Thus, for the series tuned circuit of Example 1.3.2 for which $Q = 100$ and $f_{so} = 1.3 \text{ MHz}$, the -3 -dB bandwidth is 13 kHz.

Series Tuned Wavetrap

One function of a tuned circuit is to select a signal at a wanted frequency while rejecting signals at other frequencies. One way in which this may be achieved is to use the tuned circuit as a *wavetrap*, meaning simply that it traps signals at the resonant wavelength. Figure 1.3.4 shows a simple wavetrap circuit. By tuning the series tuned circuit to resonate at the unwanted frequency, the unwanted signal will be shunted away from the load resistor R_L .

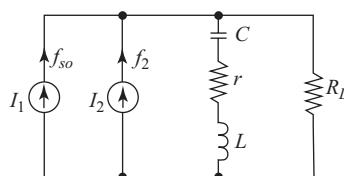


Figure 1.3.4 Simple wavetrap circuit.

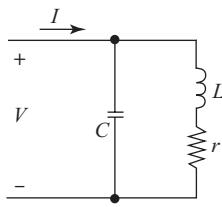


Figure 1.4.1 Parallel tuned circuit.

1.4 Parallel Tuned Circuit

The parallel tuned circuit is shown in Fig. 1.4.1. The inductor has inductance L and resistance r . The capacitor has capacitance C and is assumed to have negligible resistance. This represents quite accurately most parallel tuned circuits. As will be shown, the resonant frequency and the Q -factor of the parallel tuned circuit are, for all practical purposes, equal to those of the series tuned circuit, but the impedance is the inverse, being very high at resonance and decreasing as the frequency departs from the resonant value.

Impedance of a Parallel Tuned Circuit

Denoting the capacitive branch impedance by Z_C , and the inductive branch by Z_L , then from Fig. 1.4.1 the parallel impedance is given by

$$Z_p = \frac{Z_L Z_C}{Z_L + Z_C} \quad (1.4.1)$$

Now, $Z_C = 1/j\omega C$ and $Z_L = r + j\omega L \approx j\omega L$. The approximation introduced here is that the inductive reactance will be very much greater than the resistance at high frequencies, which is normally the case. It will also be seen that the denominator is equal to the impedance of the same components connected in series, or $Z_s = Z_L + Z_C$, which from Eq. (1.3.10) is $Z_s = r(1 + jyQ_s)$. Combining these expressions gives, to a very close approximation:

$$\begin{aligned} Z_p &= \frac{L/C}{r(1 + jyQ_s)} \\ &= \frac{R_D}{(1 + jyQ_s)} \end{aligned} \quad (1.4.2)$$

where R_D is known as the *dynamic impedance*:

$$R_D = \frac{L}{Cr} \quad (1.4.3)$$

The parallel impedance is seen to equal the dynamic impedance when the complex term in the denominator is equal to unity. This corresponds to the resonant condition for the parallel tuned circuit. The subscript D , for “dynamic,” is used to emphasize that the expression applies only for alternating currents at resonance, and the symbol R is to show that at resonance the impedance is purely resistive.

Before examining the resonant conditions in more detail, it is left as an exercise for the student to show that, in terms of Q -factor, alternative forms of the equation for dynamic impedance are

$$\begin{aligned} R_D &= \omega_o L \cdot Q \\ &= \frac{Q}{\omega_o C} \\ &= Q^2 \cdot r \end{aligned} \quad (1.4.4)$$

In these expressions the single subscript o is used for resonant frequency, and the subscript s is dropped from Q , for reasons which will be apparent shortly.

The form $Q/\omega_o C$ is particularly useful because each of the quantities involved can be obtained directly from Q -meter measurements. The form $Q^2 \cdot r$ is interesting in that it shows clearly the relationship between the series and parallel impedances at resonance. For example, if $Q = 100$ and $r = 20 \Omega$, then when connected as a series circuit the impedance at resonance would be 20Ω purely resistive, while connected as a parallel circuit the impedance at resonance would be $200 \text{ k}\Omega$ purely resistive. Note again the distinction, however, that the 20Ω is the physical resistance of the coil, which opposes direct as well as alternating currents (but see Section 1.6), while the $200 \text{ k}\Omega$ is a “dynamic resistance” applicable only to alternating current at resonance.

In summary, it is seen that the parallel circuit offers a high impedance, and the series circuit a low impedance at resonance, and the parallel impedance varies with frequency as the inverse of the series impedance. Furthermore, it may be shown that if I_o is the input current at resonance to a parallel circuit, the magnitude of the current in the capacitive branch is $I_o Q$ and in the inductive branch $\equiv I_o Q$. Thus, whereas the series resonant circuit exhibits voltage magnification, the parallel resonant circuit exhibits current magnification.

Parallel Resonant Frequency and Q -Factor

Parallel resonance occurs when the reactive part of the impedance is zero. This requires the imaginary term $jyQs$ in the impedance equation to be equal to zero. Since this is the same term as occurs in the series impedance equation, the resonant frequency must be the same for both circuits, and the subscript s can be dropped. Furthermore, by defining the Q -factor as $Q = \omega_o L/r = 1/\omega_o Cr$, the Q -factor can be used for these ratios wherever they appear in equations, whether for series or parallel circuits, and no subscript is required.

It will be recalled that the expression for parallel impedance involved the approximation $\gamma \ll |\omega L|$, which is true for most cases of practical interest, and this is a required condition for the simplified relationships between parallel and series circuits to hold true.

Relative Response of the Parallel Tuned Circuit

When the parallel tuned circuit is fed from a constant current source I , the voltage in general is given by $V = I \cdot Z_p = I \cdot R_D/(1 + jyQ)$ and at resonance by $V_o = I \cdot R_D$. The relative response A_r is the ratio V/V_o and is seen therefore to be given by

$$A_r = \frac{1}{1 + jyQ} \quad (1.4.5)$$

This shows that the relative response for the parallel tuned circuit is identical to that for the series tuned circuit, given by Eq. (1.3.13).

It also follows that the -3-dB bandwidth will be given by the same expression as for the series tuned circuit, or $B_{3 \text{ dB}} = f_o/Q$. Note carefully, however, that the relative response for the parallel circuit is defined in terms of voltages, while that for the series circuit is defined in terms of currents.

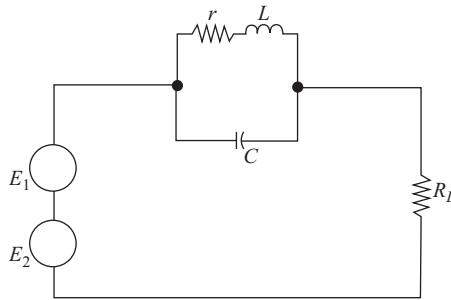


Figure 1.4.2 Parallel tuned wavetrap.

A Simple Parallel Tuned Wavetrap

The tuned circuit in Fig. 1.3.4 may be connected as a parallel tuned circuit, and tuned to resonate at the wanted frequency. In this way it becomes a *wavetrap* for the unwanted frequency.

The parallel tuned circuit may also be used as a wavetrap, as shown in Fig. 1.4.2. In this situation the wanted and unwanted signals appear as emfs in series. The parallel tuned wavetrap is connected in series with the load and is tuned to resonate at the unwanted frequency, so it presents a high impedance to this current component (recall that the series circuit was connected in parallel with the load).

1.5 Self-capacitance of a Coil

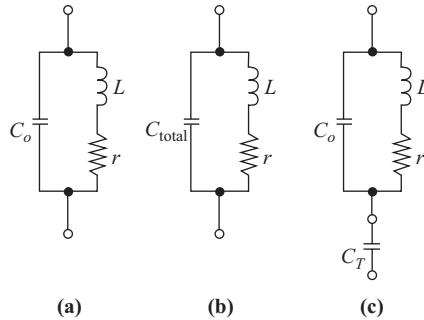
In addition to resistance, an inductor has capacitance distributed between the turns of the winding. A reasonably accurate circuit representation using lumped (as distinct from distributed) components is shown in Fig. 1.5.1(a). This figure shows that the coil in fact appears as a parallel tuned circuit, which can be represented as an impedance $Z_{\text{eff}} = r_{\text{eff}} + j\omega L_{\text{eff}}$ and which can be evaluated by the methods described in Section 1.4. The coil has a self-resonant frequency given by $\omega_{sr} = 1/\sqrt{LC_o}$, and for the coil to behave as an inductor with series resistance and inductance, the frequency of operation must be below the self-resonant frequency of the coil.

A useful approximate expression may be derived for L_{eff} for frequencies such that $|yQ|^2 \gg 1$ (note that $y = \omega/\omega_{sr} - \omega_{sr}L/\omega$ in this context). For typical circuits this would include frequencies up to about 90% of f_{sr} . Considering the coil as a parallel tuned circuit, then Eq. (1.4.2) can be used to represent this:

$$\begin{aligned}
 Z_p &= \frac{R_{Dsr}}{(1 + jyQ_{sr})} \\
 &= \frac{R_{Dsr}}{1 + (yQ_{sr})^2} (1 - jyQ_{sr}) \\
 &\approx \frac{R_{Dsr}}{(yQ_{sr})^2} (1 - jyQ_{sr})
 \end{aligned} \tag{1.5.1}$$

The subscript sr is used to denote the coil self-resonant conditions. From this it is seen that $r_{\text{eff}} = R_{Dsr}/(yQ_{sr})^2$. Recalling that $R_{Dsr} = \theta_{sr}^2 r$, this may be substituted to give

$$r_{\text{eff}} = \frac{r}{y^2} \tag{1.5.2}$$



C_T = tuning capacitance

$$C_{\text{total}} = C_o + C_T$$

Figure 1.5.1 (a) Coil with self-capacitance C_o . (b) In a parallel tuned circuit C_o is absorbed in the total tuning capacitance. (c) In a series tuned circuit, C_o appears separate from C_T .

The effective inductance is obtained by equating $\omega L_{\text{eff}} = -R_{Dsr}/yQ_{sr}$. Again recalling that $R_{Dsr} = Q_{sr}\omega_{sr}L$ gives

$$\omega L_{\text{eff}} = -\frac{\omega_{sr}L}{y} \quad (1.5.3)$$

After simplifying, this gives

$$L_{\text{eff}} = \frac{L}{1 - (\omega/\omega_{sr})^2} \quad (1.5.4)$$

The effective Q -factor of the coil at angular frequency ω may be defined as $Q_{\text{eff}} = \omega L_{\text{eff}}/r_{\text{eff}}$. Using Eqs. (1.5.2) and (1.5.4) gives, after simplifying,

$$Q_{\text{eff}} = Q \left\{ 1 - \left(\frac{\omega}{\omega_{sr}} \right)^2 \right\} \quad (1.5.5)$$

Care must be taken in how the effective circuit values are used. If the coil forms part of a parallel tuned circuit, as shown in Fig. 1.5.1(b), then C_o is simply absorbed in the total tuning capacitance and the circuit behaves as a normal parallel tuned circuit, with the total tuning capacitance, which includes C_o , resonating with the actual inductance L . The operating frequency must be below the self-resonant frequency of the coil.

When the coil is used as part of a series tuned circuit, as shown in Fig. 1.5.1(c), then L_{eff} is the inductance to which the external capacitor must be tuned for resonance, and Q_{eff} will be the effective Q -factor, assuming that the losses in the capacitor are negligible. The bandwidth of the series circuit, assuming Q_{eff} remains sensibly constant over the bandwidth range, is given by

$$B_{3 \text{ dB eff}} = \frac{f_o}{Q_{\text{eff}}} \quad (1.5.6)$$

It is worth noting that most Q -meters measure Q_{eff} .

EXAMPLE 1.5.1

A coil has a series resistance of 5Ω , a self-capacitance of 7 pF , and an inductance of $1 \mu\text{H}$. Determine the effective inductance and effective Q -factor when the coil forms part of a series tuned circuit resonant at 25 MHz .

SOLUTION The self-resonant frequency of the coil is

$$f_{sr} = \frac{1}{2\pi\sqrt{10^{-6} \times 7 \times 10^{-12}}} = 60 \text{ MHz}$$

The Q -factor of the coil, excluding self-capacitive effects, is

$$Q = \frac{2\pi \times 25 \times 10^6 \times 10^{-6}}{5} = 31.4$$

Hence

$$L_{\text{eff}} = \frac{10^{-6}}{1 - (25/60)^2} = 1.21 \mu\text{H}$$

and

$$Q_{\text{eff}} = Q \left(1 - \left(\frac{25}{60} \right)^2 \right) = 26$$

1.6 Skin Effect

The self-induced emf in a conductor resulting from the rate of change of flux linkages opposes the current flow that gives rise to the flux (Lenz's law). Normally it is assumed that all the flux links with all the conductor. However, the actual flux linkages increase toward the core of the conductor, since the magnetic flux within the conductor only links with the inner section; in Fig 1.6.1(a), for example, flux line ϕ_1 links with the complete conductor, while flux line ϕ_2 links only with the section of radius a . The self-induced emf is greatest at the center of the conductor, which experiences the greatest flux linkages, and becomes less toward the outer circumference. This results in the current density being least at the center and increasing toward the outer circumference, since the induced emf opposes the current flow (Lenz's law). The lower current density at the center results in lower magnetic flux there, which tends to offset the effect producing the nonuniform distribution, and, in this way, equilibrium conditions are established. The overall effect, however, is that the current tends to flow near the surface of the conductor, this being referred to as the *skin effect*. Because the current is confined to a smaller cross section of the conductor, the apparent resistance of the conductor increases. The increase is more noticeable for thick conductors and at high frequencies (where the rate of change of flux linkages is high). Equally important is the fact that the resistance becomes dependent on frequency.

With coils, a special type of wire called *Litzendraht wire* (*Litz wire*, for short) is often used to reduce skin effect. Litz wire is made up of strands insulated from each other and wound in such a way that each strand changes position between the center and outer edge over the length of the wire [Fig. 1.6.1(b)]. In this manner, each strand, on average, has equal induced emfs, so that over the complete cross section (made up of the many cross sections of the individual strands), the current density tends to be uniform.

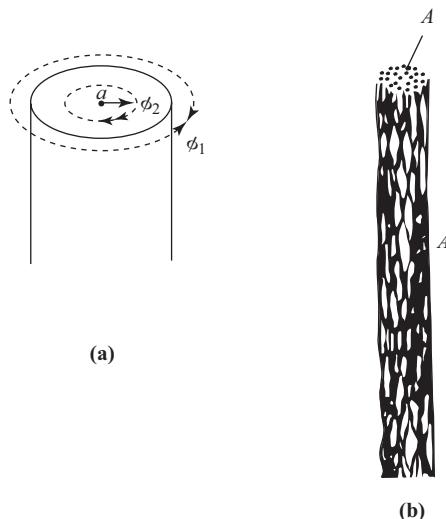


Figure 1.6.1 (a) Magnetic flux linkage in a conductor. (b) Litzendraht wire.

1.7 Mutual Inductance

Reaction between inductive circuits that are physically isolated can occur as a result of common magnetic flux linkage. This effect can be taken into account by means of a mutual inductance M . For a harmonically varying current I_1 in an inductance L_1 that is magnetically coupled to an inductance L_2 , the induced emf in L_2 is given by

$$E_2 = \pm j\omega M I_1 \quad (1.7.1)$$

The sign to be used depends on the physical disposition of the coils, the situation for the positive sign (meaning E_2 leads I_1 by 90°) being shown in Fig. 1.7.1(a).

A useful equivalent circuit is shown in Fig. 1.7.1(b), but this is valid only for ac conditions; it shows a dc path where none may exist in the actual circuit.

The situation where E_2 lags I_1 by 90° is shown in Fig. 1.7.2. In practice the coils may be enclosed in a screening can, shown dotted, and the internal connections are not visible. Thus the output may be reversed in phase from that of the previous situation, as shown. An equivalent circuit for this situation is shown in Fig. 1.7.2(b).

On circuit diagrams the phase relationships are sometimes identified by dots, this being referred to as the *dot notation method*. Thus, for Fig. 1.7.1, if the primary dot is placed at terminal 1, the secondary dot would be placed at terminal 2. For Fig. 1.7.2, with the primary dot at terminal 1, the secondary dot would be placed at terminal 2'.

From Fig. 1.7.2(a), it is seen that with terminals 1' and 2' joined together, the total number of turns in a given direction of winding is increased, and therefore the total inductance between terminals 1 and 2 should be *at least* $L_1 + L_2$. Mutual inductance, if present, must therefore add to this, and from Fig. 1.7.2(b), the inductance between terminals 1 and 2 is seen to be $L_1 + L_2 + 2M$. This is the maximum value of inductance that can be achieved with L_1 and L_2 in series.

For Fig. 1.7.1(a), with terminals 1' and 2' joined together, the magnetizing force of L_2 opposes that of L_1 , because the windings are in opposite directions. As a result, the mutual inductance M opposes the self-inductance

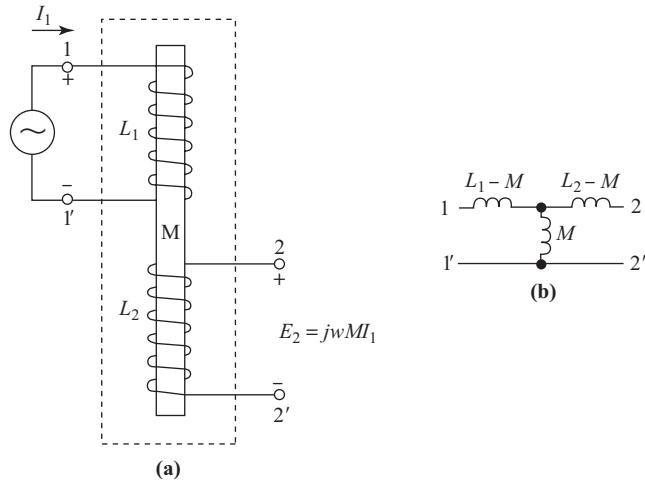


Figure 1.7.1 (a) Mutual-inductive coupling in which the secondary voltage leads the primary current by 90° . (b) Equivalent circuit for (a).

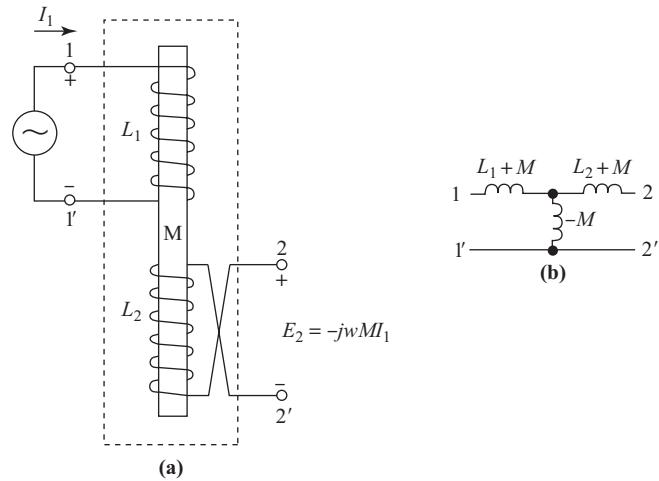


Figure 1.7.2 (a) Mutual-inductive coupling in which the secondary voltage lags the primary current by 90° . (b) Equivalent circuit for (a).

in each case. From Fig. 1.7.1(b), the total inductance between terminals 1 and 2 is seen to be $L_1 + L_2 - 2M$. Thus the total series inductance of two coils connected in series is given by

$$L_s = L_1 + L_2 \pm 2M \quad (1.7.2)$$

A similar but rather more involved argument shows that for two inductors in parallel the total inductance is given by

$$L_p = \frac{L_1 L_2 - M^2}{L_1 + L_2 \pm 2M} \quad (1.7.3)$$

By utilizing mutual inductance between two coils, the effective inductance can be altered in steps, by making the appropriate connections, from a minimum L_{\min} to a maximum L_{\max} , where

$$L_{\min} = \frac{L_1 L_2 - M^2}{L_1 + L_2 + 2M} \quad (1.7.4)$$

$$L_{\max} = L_1 + L_2 + 2M \quad (1.7.5)$$

Furthermore, by making M variable, for example, by physically altering the spacing between the coils, a continuously variable inductance is obtained.

Note that M cannot be identified as a physical winding in the sense that L_1 and L_2 can, and it has to be determined by measurement. In practice, it may prove easier to determine what is termed the *coefficient of coupling* k , where

$$M = k\sqrt{L_1 L_2} \quad (1.7.6)$$

1.8 High-frequency Transformers

Mutual inductive coupling forms the basis of most high-frequency transformers. The circuit for a high-frequency transformer is shown in Fig. 1.8.1(a), where, in addition to the inductive elements, the series resistance of each coil is also shown. The notation has also been changed slightly; the subscript p signifies “primary” and s signifies “secondary.” The primary, the secondary, or both, may be tuned. C_p is the primary tuning capacitance, and C_s the secondary tuning capacitance.

The equivalent circuit, based on the equivalent circuit of Fig. 1.7.1(b), is used here, which, it will be recalled, means that the secondary induced voltage $j\omega M I_p$ leads the primary coil current I_p by 90° . In general, it seldom matters if this voltage leads or lags the current by 90° .

The following notation is used in the equivalent circuit:

$$Z_p = r_p + j\omega L_p \quad (1.8.1)$$

$$Z_s = r_s + j\omega L_s \quad (1.8.2)$$

$$Z_m = j\omega M \quad (1.8.3)$$

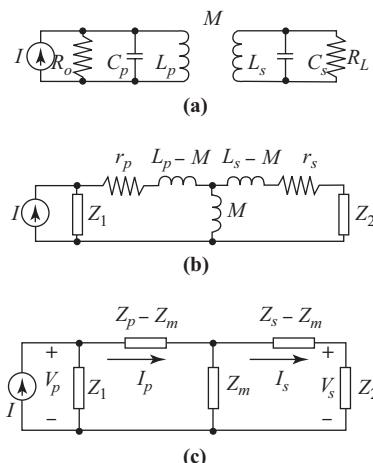


Figure 1.8.1 (a) High-frequency transformer circuit. (b) Equivalent circuit for (a). (c) Block schematic for (b).

The load resistance is in parallel with the secondary tuning capacitor, and the parallel impedance Z_2 is

$$Z_2 = \frac{R_L}{1 + j\omega C_s R_L} \quad (1.8.4)$$

Likewise, at the primary, the source resistance R_o appears in parallel with C_p , and the impedance Z_1 is

$$Z_1 = \frac{R_o}{1 + j\omega C_p R_o} \quad (1.8.5)$$

Transfer Impedance

The transfer impedance Z_T is the ratio of output voltage V_s to input current I . To find V_s , note that current I divides at the input junction and the current divider rule may be employed to find I_p . Current I_p divides at the center junction and again the current divider rule may be employed to find I_s . The analysis is straightforward, and the result is

$$I_s = \frac{IZ_1 Z_m}{(Z_p + Z_1)(Z_s + Z_2) - Z_m^2} \quad (1.8.6)$$

The secondary voltage is $I_s Z_2$, and hence the transfer impedance is

$$\begin{aligned} Z_T &= \frac{V_s}{I} \\ &= \frac{Z_1 Z_2 Z_m}{(Z_p + Z_1)(Z_s + Z_2) - Z_m^2} \end{aligned} \quad (1.8.7)$$

It will be observed that Z_T takes into account the damping effects of the source resistance R_o and load resistance R_L .

EXAMPLE 1.8.1

A high-frequency transformer has identical primary and secondary circuits for which $L_p = L_s = 150 \mu\text{H}$, $C_p = C_s = 470 \text{ pF}$, and the Q -factor for each circuit alone (that is, not coupled) is 85. The coefficient of coupling is 0.01. The load resistance is 5000Ω , and the constant current source feeding the transformer has an internal resistance of $75 \text{ k}\Omega$. Determine the transfer impedance at resonance.

SOLUTION The common resonant frequency is

$$\omega_o = \frac{1}{\sqrt{150 \times 10^{-6} \times 470 \times 10^{-12}}} = 3.77 \text{ Mrad/sec}$$

$$Z_2 = \frac{5000}{1 + j3.77 \times 10^6 \times 470 \times 10^{-12} \times 5000} = 63 - j558 \Omega$$

$$Z_1 = \frac{75,000}{1 + j3.77 \times 10^6 \times 470 \times 10^{-12} \times 75,000} = 4.3 - j565 \Omega$$

At resonance, $Q = \omega_o L / r$ and hence $r = \omega_o L / Q$, so

$$Z_p = Z_s = \omega_o L \left(\frac{1}{Q} + j \right) = 6.65 + j565 \Omega$$

$$Z_M = j3.77 \times 10^6 \times .01 \times 150 \times 10^{-6} = j5.65 \text{ ohms}$$

Denote the denominator by $\Delta = (Z_p + Z_1)(Z_s + Z_2) - Z_M^2 \cong 791 + j80 \Omega^2$.

Hence the transfer impedance is

$$Z_r = \frac{Z_1 Z_2 Z_M}{\Delta} = 43.8 - j2.25 \times 10^3 \Omega$$

This example shows that at resonance the transfer impedance is almost entirely capacitive; that is, the output voltage will lag the current. For a 1-mA input current, the output voltage will be approximately $-j2.25$ V.

Synchronously Tuned Circuits

Where the primary and secondary are tuned separately to the same resonant frequency, the transformer is referred to as a *synchronously tuned transformer*. Because of the mutual coupling between the circuits, each circuit, primary and secondary, will detune the other to some extent, and this may result in two peaks occurring in the overall frequency response curve. Whether or not two peaks occur depends on the degree of coupling, and, more specifically, the shape of the response curve depends on the product $k\sqrt{Q_p Q_s}$.

In general, the primary circuit constants will differ from those of the secondary. To illustrate the important features of the synchronously tuned transformer, identical tuned circuits will be assumed; thus $L_p = L_s = L$, $Q_p = Q_s = Q$, and $C_p = C_s = C$.

Assuming that the primary is fed from a constant current source, the variation of output voltage with frequency is given by the transfer impedance Z_T . The magnitude of Z_T in decibels relative to 1Ω ($20 \cdot \log |Z_T|$) is shown in Fig. 1.8.2 for three different values of kQ , where the peak output for the $kQ = 1$ curve is used as the zero reference for the decibel scale (the computations for these curves were carried out using Mathcad).

For $kQ = 1$, the transformer is said to be *critically coupled*; for $kQ < 1$, it is *undercoupled*; and for $kQ > 1$, it is *overcoupled*.

The double-humped curve of Fig. 1.8.2 is a feature of the overcoupled circuit. This is often combined with critically coupled (or slightly undercoupled) circuits to obtain a composite response that is flat along the top and that has sides that fall away sharply. Figure 1.8.3 illustrates the composite response of the three curves $kQ = 0.5, 1$, and 2 . The curves are plotted in decibels relative to the resonant (or mid-frequency) response against the frequency variable f/f_o . It should be noted that the overall, or composite, curve is obtained simply by adding the decibel response curves for the individual circuits.

-3-dB Bandwidth of Synchronously Tuned Circuits

The -3 -dB bandwidth also depends on the degree of coupling. To find the bandwidth in any particular case, it is easiest to solve the transfer impedance equation. For a constant source current I , the output voltage V_s will drop by 3 dB at the frequencies where the magnitude of the transfer impedance is $1/\sqrt{2}$ times its resonant value. Denoting by $|Z_T(f_o)|$ the magnitude of the transfer impedance at resonance and by $|Z_T(f)|$ the value at frequency f , then the roots of the equation $|Z_T(f_o)|/\sqrt{2} - |Z_T(f)| = 0$ gives the -3 -dB frequencies, and the bandwidth is the difference between these. This equation can be rearranged as

$$|Z_T(f_o)| - \sqrt{2} |Z_T(f)| = 0 \quad (1.8.8)$$

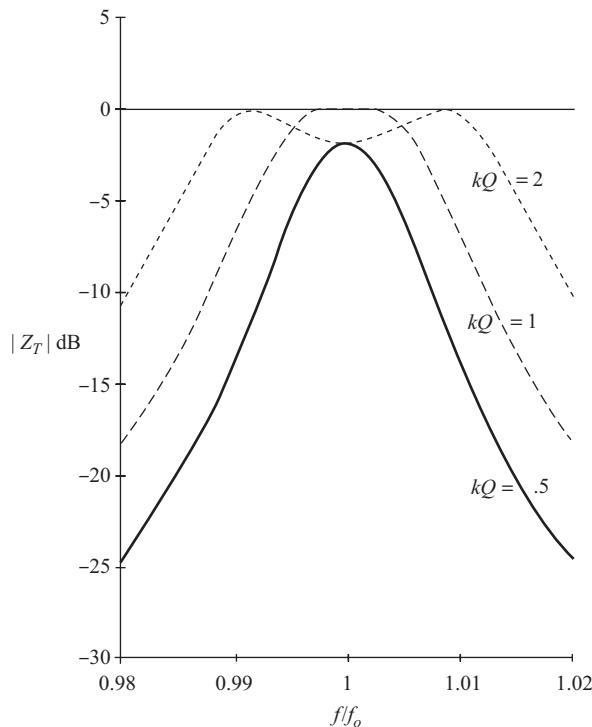


Figure 1.8.2 Frequency response curves for coupled circuits.

This equation is best solved by numerical methods. With identical primary and secondary circuits and for $kQ = 1$, it will be found that the magnitude of the voltage transfer function is equal to unity, and the -3-dB bandwidth is $\sqrt{2}$ times the bandwidth of either circuit alone (see Problem 1.39). The -60-dB bandwidth of the critically coupled circuit is reduced by a factor of about $\sqrt{10^3}$ compared to the corresponding bandwidth of either single tuned circuit.

An alternative approach to finding bandwidth is to plot the response curves and estimate the bandwidths from these.

Voltage Transfer Function

Another useful function is the *voltage transfer function*, which is the ratio of load voltage V_s to input voltage V_p . The load voltage is $V_s = IZ_T$, and from Fig. 1.8.1(c) the input voltage is $V_p = IZ_{IN}$, where Z_{IN} is the input impedance seen by the source I (and note that the internal impedance of the source is included in Z_{IN}). Combining the impedances in Fig. 1.8.1(c) in the normal series-parallel manner and simplifying yields

$$Z_{IN} = Z_1 \cdot \frac{Z_p(Z_s + Z_2) - Z_M^2}{(Z_p + Z_1)(Z_s + Z_2) - Z_M^2} \quad (1.8.9)$$

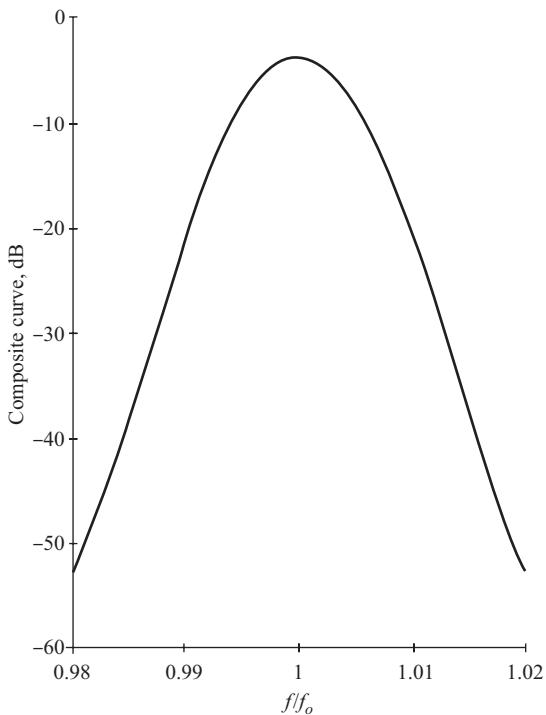


Figure 1.8.3 Composite response curve for the combined $kQ = 0.5, 1$, and 2 transformers.

Denoting the voltage transfer function as VTF, then

$$\begin{aligned}
 \text{VTF} &= \frac{V_s}{V_p} \\
 &= \frac{Z_T}{Z_{\text{IN}}} \\
 &= \frac{Z_2 \cdot Z_M}{Z_p(Z_s + Z_2) - Z_M^2} \tag{1.8.10}
 \end{aligned}$$

Exercise 1.8.1 Calculate the VTF for the transformer specified in Example 1.8.1. (Ans. $(1.86 - j 80 \times 10^{-3})$.)

Untuned Primary and Untuned Secondary Circuits

Two other commonly encountered configurations are the *untuned primary tuned secondary* and the *tuned primary untuned secondary*, shown in Fig. 1.8.4. The procedure for analyzing these circuits follows the general procedure given above. In the first case, the C_p term is omitted from the input admittance given by Y_1 , and in the second case, the C_s term is omitted from the admittance Y_2 . The analysis of amplifier circuits using these circuit configurations is covered in Chapter 5.

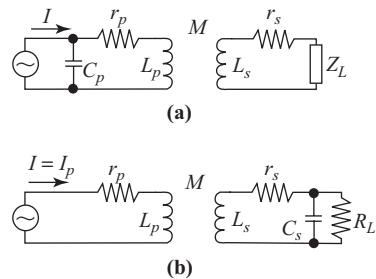


Figure 1.8.4 (a) Tuned primary untuned secondary. (b) Untuned primary tuned secondary coupling.

1.9 Tapped Inductor

The tapped inductor circuit is shown in Fig. 1.9.1(a), where it will be seen that the load is connected to a tapping point on the inductor. Mutual inductive coupling exists between the two sections of the coil, while the total inductance L_p usually forms part of a tuned circuit, completed by the primary tuning capacitor C_p , as shown. The tapped inductor is a coupling method that is widely used to reduce the damping effect of a load or a source on the Q -factor of a tuned circuit. Both tapping arrangements may be used together as shown in Fig. 1.9.1(c).

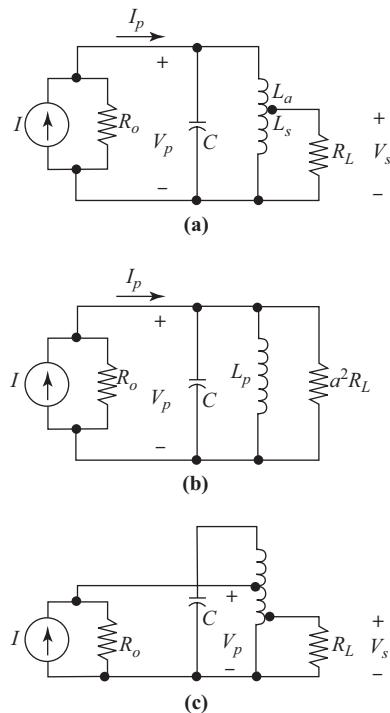


Figure 1.9.1 (a) Tapped inductor used to reduce the loading effect from the load. (b) Equivalent circuit for ideal coupling. (c) Source and load tapped down.

The primary inductance is given by $L_p = L_a + L_s + 2M$, where L_a is the self-inductance of the upper section of the tapped inductor uncoupled to L_s , L_s the self-inductance of the lower section uncoupled to L_a , and $M = k\sqrt{L_a L_s}$. Thus, M is the mutual inductance between L_a and L_s (and not between primary L_p and secondary L_s).

Consider first the ideal arrangement, where the coil resistance may be assumed zero, and the coefficient of coupling k between the two sections of the coil is unity. With the secondary load removed, the primary current consists of a magnetizing component $I_m = V_p/j\omega L_p$. With the secondary load connected, the magnetomotive force (mmf) of the secondary must be balanced by an additional mmf from the primary, which requires an additional component of primary current I'_p such that $N_p I'_p = N_s I_s$. Here, N_p is the number of primary turns, which is equal to the *total* number of turns on the coil, and N_s is the number of turns on the secondary L_s . Denoting the transformation ratio by

$$a = \frac{N_p}{N_s} \quad (1.9.1)$$

it is seen that the total primary current is given by

$$\begin{aligned} I_p &= I_m + I'_p \\ &= \frac{V_p}{j\omega L_p} + \frac{I_s}{a} \end{aligned} \quad (1.9.2)$$

Also, because the same total flux links with all the coil, the induced emf per turn will be the same for both primary and secondary, and hence

$$\begin{aligned} \frac{V_p}{V_s} &= \frac{N_p}{N_s} \\ &= a \end{aligned} \quad (1.9.3)$$

The load impedance is given by $Z_L = V_s/I_s$, and substituting this along with Eq. (1.9.3) in Eq. (1.9.2) gives

$$\begin{aligned} I_p &= \frac{V_p}{j\omega L_p} + \frac{V_p}{a^2 Z_L} \\ &= V_p \left(\frac{1}{j\omega L_p} + \frac{1}{a^2 Z_L} \right) \end{aligned} \quad (1.9.4)$$

This shows that the load impedance Z_L is transformed to an equivalent impedance $a^2 Z_L$ in parallel with the inductance L_p .

The equivalent circuit for a load impedance consisting of a resistor R_L is shown in Fig. 1.9.1(b). At resonance, the capacitor C_p resonates with the ideal inductor L_p . The dynamic impedance of this combination is infinite, so the effective load impedance referred to the primary is $a^2 R_L$, and a comparatively low value of R_L can be transformed by a factor a^2 to a higher value.

The voltage transfer function for this ideal tapped-coil, as given by Eq. (1.9.3), is equal to $1/a$. When the circuit is fed at resonance from a constant current source I and internal resistance R_o , then $V_p = IR_p$

where R_p is equal to $a^2 R_L$ in parallel with R_o . The secondary, or load, voltage is $V_s = V_p/a = IR_p/a$ and the transfer impedance at resonance is

$$\begin{aligned} Z_T &= \frac{V_s}{I} \\ &= \frac{R_p}{a} \end{aligned} \quad (1.9.5)$$

When $R_o \gg a^2 R_L$, $Z_T \approx a R_L$. The significance of this is that with the tapped-tuned circuit as a load for an amplifier, the voltage gain is increased by a factor a , even though the inductor behaves as a step-down transformer of ratio $1/a$.

This would suggest that very high gains could be achieved by making a as large as possible, by tapping well down the coil. However, by making a very large, the condition $R_o \gg a^2 R_L$ no longer applies (see Problem 1.42). Also, in practice the finite dynamic impedance of the tuned circuit would limit gain, and more importantly, the coupling coefficient can never be unity in practice. This is especially true when the tapping point is close to the ends of a coil (as required for large a), where leakage flux is quite high. Thus the voltage gain cannot be increased indefinitely by increasing a .

The coefficient of coupling k , as well as the inductances L_a and L_s , varies in a complicated way with the position of the tap on the coil, and design information is usually presented in the form of curves for specified coil sizes and k values. In practice, it is sometimes assumed that inductance is proportional to the square of the number of turns to get an approximate idea of the values involved, but the assumption does not yield accurate results in most cases, and the best procedure usually is to determine the optimum tapping point experimentally.

1.10 Capacitive Tap

As an alternative to using an inductive tap, a capacitive tap may be used as shown in Fig. 1.10.1(a). Again, the idea is to reduce the loading effect of R_L on the tuned circuit so that reasonable selectivity is maintained.

Consider first the capacitive branch, and denote the conductance of the load as $G_L = 1/R_L$. The admittance of C_2 and G_L in parallel is

$$Y_2 = j\omega C_2 + G_L \quad (1.10.1)$$

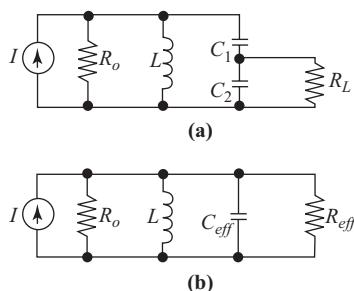


Figure 1.10.1 (a) Capacitive tap. (b) Equivalent circuit.

The admittance of C_1 is

$$Y_1 = j\omega C_1$$

and therefore the total admittance of the capacitive branch is

$$Y_C = \frac{Y_1 Y_2}{Y_1 + Y_2} \quad (1.10.2)$$

Substituting the full expressions for Y_1 and Y_2 and rationalizing gives

$$Y_C = \frac{G_L(\omega C_1)^2}{G_L^2 + \omega^2(C_1 + C_2)^2} + j \frac{\omega C_1(G_L^2 + \omega^2 C_2(C_1 + C_2))}{G_L^2 + \omega^2(C_1 + C_2)^2} \quad (1.10.3)$$

For the load resistance not to affect the tuning capacitance, it is necessary that the susceptance of C_2 be sufficiently greater than G_L so that it carries most of the current in the capacitive branch. This allows the approximation $(\omega(C_1 + C_2))^2 \gg G_L^2$ to be made in the denominator, and the expression for Y_C becomes, after simplifying,

$$Y_C \approx \left(\frac{C_1}{C_1 + C_2} \right)^2 G_L + j\omega \frac{C_1 C_2}{C_1 + C_2} \quad (1.10.4)$$

It is seen therefore that the effective tuning capacitance is C_1 in series with C_2 .

The transformation ratio for the capacitive tap may be defined as

$$a = \frac{C_1 + C_2}{C_1} \quad (1.10.5)$$

Thus, the conductance in shunt across the tuned circuit is G_L/a^2 or, in terms of load resistance, the effective parallel resistance is $a^2 R_L$.

For the condition that $(\omega C_2)^2 \gg G_L^2$, the voltage across the load is that provided by the capacitive tap, $V_L \approx V_p/a$. The voltage across the circuit is $V_p = IZ_p$, where $Z_p = R_o \parallel a^2 R_L \parallel Z_{TL}$ is the parallel combination of impedances seen by the constant current source I , and Z_{TC} is the impedance of the tuned circuit alone. Thus the transfer impedance for the capacitive tapped circuit is

$$\begin{aligned} Z_T &= \frac{V_L}{I} \\ &= \frac{V_p}{a} \cdot \frac{Z_p}{V_p} \\ &= \frac{Z_p}{a} \end{aligned} \quad (1.10.6)$$

At resonance, the impedance of the tuned circuit alone is R_D , the dynamic impedance, and on the assumption that R_D and R_o are both much greater than $a^2 R_L$, the transfer impedance becomes

$$Z_T \approx aR_L \quad (1.10.7)$$

Now the output voltage for a constant current I is $IZ_T = aIR_L$. If the load R_L is connected directly across the tuned circuit, the output voltage will be IR_L , again on the assumption that R_o and R_D are very much greater than R_L . Hence it would appear that the tap provides a voltage gain of a . However, a cannot be increased indefinitely. It must be kept in mind that the analysis is only valid for the condition that $a^2 R_L \ll (R_o \text{ and } R_D)$, and increasing a indefinitely eventually invalidates this condition.

On the assumption that R_o and R_D are very much greater than $a^2 R_L$, the effective impedance at resonance is $a^2 R_L$, and this can be interpreted as the effective dynamic resistance of the tuned circuit. From the definition of dynamic resistance, the effective Q -factor of the circuit is given by $Q_{\text{eff}} = \omega_o C_s a^2 R_L$ and the -3 -dB bandwidth is

$$\begin{aligned} B_{-3 \text{ dB}} &= \frac{f_o}{Q} \\ &= \frac{f_o}{2\pi f_o C_s a^2 R_L} \\ &= \frac{1}{2\pi C_3 a^2 R_L} \\ &= \frac{1}{2\pi C_2 a R_L} \end{aligned} \quad (1.10.8)$$

EXAMPLE 1.10.1

For the circuit of Fig. 1.10.1(a), $C_1 = 70 \text{ pF}$, $C_2 = 150 \text{ pF}$, and the load resistance is 200Ω . The Q -factor of the undamped circuit is 150 and the resonant frequency is 27 MHz. Determine the voltage gain at resonance achieved by using the capacitive tap and the -3 -dB bandwidth. The current source has an internal resistance of $40 \text{ k}\Omega$.

SOLUTION $\omega_o = 2\pi \times 27 \times 10^6 = 169.6 \text{ Mrad/s}$

and $G_L = \frac{1}{200} = 5 \text{ mS}$.

Checking the approximation used in the denominator,

$$\frac{[\omega_o(C_1 + C_2)]^2}{G_L^2} = 55.7$$

Hence the approximation is valid and

$$a = \frac{C_1 + C_2}{C_1} = 3.14$$

The effective load across the tuned circuit is

$$R_{\text{eff}} = a^2 R_L = 1.98 \text{ k}\Omega$$

It will be seen that this is much less than the internal resistance of the source. The tuning capacitance is

$$C_s = \frac{C_1 C_2}{C_1 + C_2} = 47.7 \text{ pF}$$

and

$$R_D = \frac{Q}{\omega_o C_s} = 18.5 \text{ k}\Omega$$

This is also much greater than R_{eff} , and so the -3 -dB bandwidth is given by

$$B_{-3 \text{ dB}} = \frac{1}{2\pi C_2 a R_L} = 1.69 \text{ MHz}$$

1.11 Maximum Power Transfer and Impedance Matching

When a signal source is required to deliver power to a load, the power transfer should be a maximum. The average power delivered to the load Z_L in Fig. 1.11.1(a) is

$$\begin{aligned} P_L &= V_R I \\ &= \frac{ER_L}{\sqrt{(R_s + R_L)^2 + (X_s + X_L)^2}} \frac{E}{\sqrt{(R_s + R_L)^2 + (X_s + X_L)^2}} \\ &= \frac{E^2 R_L}{(R_s + R_L)^2 + (X_s + X_L)^2} \end{aligned} \quad (1.11.1)$$

Looking first at the effect of X_L on P_L , it will be seen that by making $X_L = -X_s$, P_L will be at a maximum given by

$$P_L = \frac{E^2 R_L}{(R_s + R_L)^2} \quad (1.11.2)$$

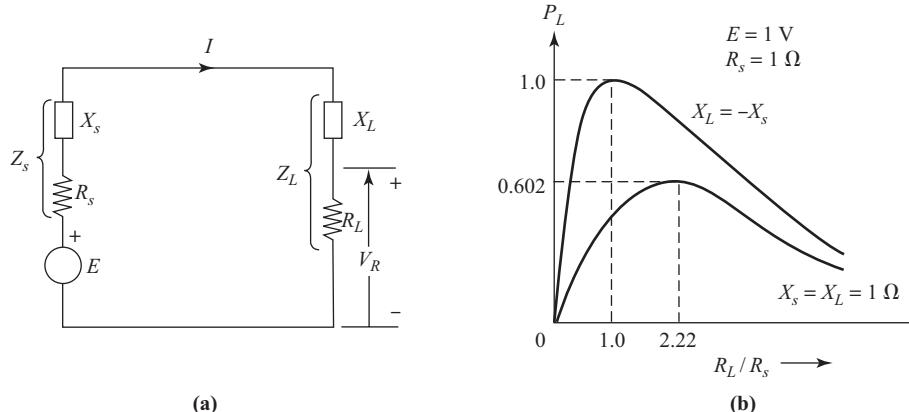


Figure 1.11.1 (a) Signal source delivering power to load. (b) Power as a function of load impedance.

R_L can now be varied to maximize this expression, and the best value is obtained by equating the differential coefficient of P_L with respect to R_L to zero:

$$\frac{dP_L}{dR_L} = \frac{E^2[(R_s + R_L) - 2R_L]}{(R_s + R_L)^3} = 0$$

from which

$$R_s = R_L \quad (1.11.3)$$

Combining this with the condition on X_L , it is seen that for maximum power transfer the load impedance must be the complex conjugate of the source impedance, or

$$R_L + jX_L = R_s - jX_s \quad (1.11.4)$$

This is known as a *conjugate match*, and although it results in maximum power transfer, it is good at only one frequency (that which makes $jX_L = -jX_s$).

Substituting $R_L = R_s$ in Eq. (1.11.2) gives the maximum available average power:

$$\begin{aligned} P_L &= \frac{E^2 R_s}{(R_s + R_s)^2} \\ &= \frac{E^2}{4R_s} \end{aligned} \quad (1.11.5)$$

Alternatively, if the constant current generator representation of a source is used, then

$$\begin{aligned} P_L &= \frac{(IR_s)^2}{4R_s} \\ &= \frac{I^2 R_s}{4} \end{aligned} \quad (1.11.6)$$

Equation (1.11.6) gives the maximum available power in terms of the constant current I and internal resistance R_s .

Equation (1.11.6) may be written in terms of source conductance $G_s = 1/R_s$ as

$$P_L = \frac{I^2}{4G_s} \quad (1.11.7)$$

When this is compared with Eq. (1.11.5), the duality of the voltage and current representations becomes apparent.

Where operation is required over a range of frequencies, the best value for Z_L is that which produces a true *reflectionless match*, which is

$$Z_L = Z_s \quad (1.11.8)$$

or

$$R_L + jX_L = R_s + jX_s$$

Reflectionless matching is not as efficient as conjugate matching, as illustrated in Fig. 1.11.1(b). However, it does give a fairly broad maximum, and there are other considerations that make it the usual choice when signal transmission over lines is required, as described in Chapter 13.

1.12 Low-frequency Transformers

In the *ideal* low-frequency transformer, all the magnetic flux set up by the primary ampere-turns links with the secondary winding (or secondary windings, as there may be more than one). Because of the very close coupling between primary and secondary ($k \approx 1$), the turns ratio N_p/N_s proves to be a more useful parameter than mutual inductance M in describing the low-frequency transformer. Also, in the ideal case, the voltage drops in primary and secondary windings can be neglected, as can the power loss in the magnetic core.

Under these conditions, the applied primary voltage V_p is equal to the back emf induced in the primary winding of N_p turns, which by Faraday's law of magnetic induction gives

$$V_p = N_p \frac{d\phi}{dt} \quad (1.12.1)$$

Likewise, ignoring the voltage drop in the secondary winding, the emf induced in the secondary E_s will be equal to the secondary terminal voltage V_s , so

$$V_s \cong E_s = N_s \frac{d\phi}{dt} \quad (1.12.2)$$

It follows, therefore, that

$$\begin{aligned} \frac{V_p}{V_s} &= \frac{N_p}{N_s} \\ &= a \end{aligned} \quad (1.12.3)$$

where $a = N_p/N_s$ is the turns ratio.

When the secondary is loaded such that it draws a current I_s , the secondary ampere turns $N_s I_s$ must balance the primary ampere turns $N_p I_p$ (otherwise, the imbalance would result in a change in induced current in such a direction as to restore balance). It follows, therefore, that

$$\begin{aligned} \frac{I_p}{I_s} &= \frac{N_s}{N_p} \\ &= \frac{1}{a} \end{aligned} \quad (1.12.4)$$

Clearly, from Eqs. (1.12.3) and (1.12.4) $V_p I_p = V_s I_s$, which is to be expected for an ideal transformer.

The load Z_L connected to the secondary may be referred to the primary Z'_L in the following way. The secondary load [Fig. 1.12.1(a)] is

$$Z_L = \frac{V_s}{I_s} \quad (1.12.5)$$

The load, as seen from the primary terminals, is

$$Z'_L = \frac{V_p}{I_p} \quad (1.12.6)$$

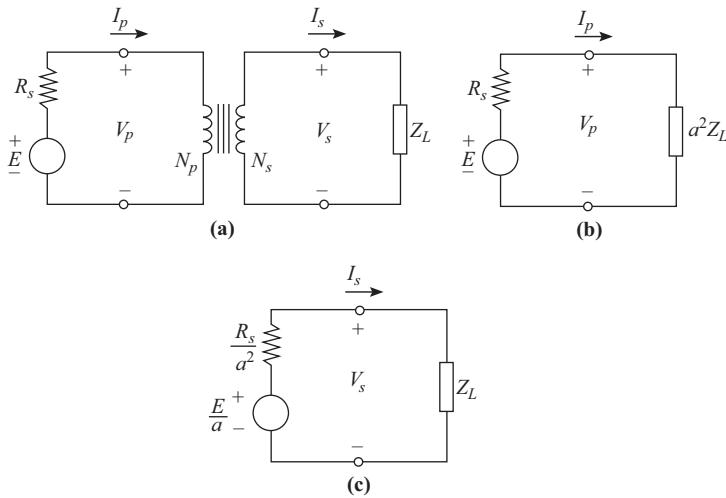


Figure 1.12.1 Ideal low-frequency transformer (a) circuit, (b) circuit referred to primary, and (c) circuit referred to secondary.

Substituting for V_p and I_p from Eqs. (1.12.3) and (1.12.4) and using the relationship in Eq. (1.12.5), Eq. (1.12.6) for Z'_L may be transformed [Fig. 1.12.1(b)] to

$$Z'_L = a^2 Z_L \quad (1.12.7)$$

Although based on the ideal transformer, this relationship is usually sufficiently accurate for and proves to be very useful in practical calculations.

By similar arguments, a voltage generator source of emf E and internal resistance R_s may be transferred to the secondary so that the load appears to be fed from a source of emf E/a and internal resistance R_s/a^2 [Fig. 1.12.1(c)].

The low-frequency circuit model for a practical transformer is shown in Fig. 1.12.2(a). In the practical transformer, a small primary current is required to set up the magnetic flux in the core; this can be accounted for by showing an inductance L_c in parallel with the ideal primary. There will also be eddy current and hysteresis power losses in the core. These losses are dependent on the primary voltage and independent of current and can be represented by a resistance R_c in parallel with the ideal primary winding. Each winding will have resistance, represented by r_p for the primary and r_s for the secondary. Each winding that carries current will produce a certain amount of magnetic flux that does not link with the other windings; this is known as *leakage flux*, and its effect is represented by the inductances L_p and L_s .

The self-capacitance of the primary winding is represented by C_p and of the secondary winding by C_s . In addition, there will be capacitive coupling between the windings, represented by C_{ps} .

The voltage-gain/frequency response for the transformer is shown in Fig. 1.12.2(b). At low frequencies, gain fall-off occurs because of L_c . At mid-frequencies, the reactive elements can be ignored and the response curve is reasonably flat. At the higher frequencies, a series resonance occurs, which results in a peak in the response curve. Beyond this the capacitive shunting effect results in gain fall-off.

It must be kept in mind that the equivalent circuit is valid for ac signals only and cannot be used for a dc analysis.

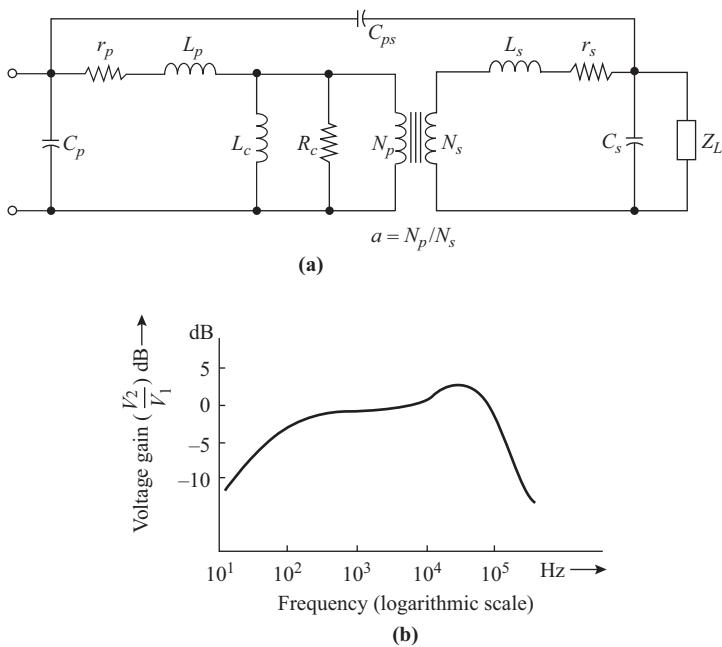


Figure 1.12.2 (a) Equivalent circuit for a low-frequency transformer. (b) Frequency response.

1.13 Passive Filters

Filter Transfer Function

Filtering of signals in telecommunications is necessary in order to select the desired signal from the range of signals transmitted and also to minimize the effects of noise and interference on the wanted signal. Electrical filters may be constructed using resistors and capacitors, resistors and inductors, or all three types of components, but it will be noticed that at least one reactive type of component must be present. The resonant circuits described in Sections 1.3 and 1.4 and also the tuned transformers described in Section 1.8 are all examples of filters. Many applications in telecommunications require filters with very sharply defined frequency characteristics, and the filter circuits are much more complex than simple tuned circuits. Most complex filters use all three types of components: inductors, capacitors, and resistors. The inductors tend to be large and costly, and these are now being replaced in many filter designs by electronic circuits that utilize operational amplifiers along with capacitors and resistors. Such filters are known as *active filters*. Active filters have many advantages over passive filters, the chief ones being that they are small in size, lightweight, and less expensive and offer more flexibility in filter design. The disadvantages are that they require external power supplies and are more sensitive to environmental changes, such as changes in temperature.

Filter design is a very extensive topic, embracing active filters, passive filters, and digital filters, and in this section only a brief introduction to passive filters will be given. In addition to passive filters designed using electrical components, various other types are available that utilize some form of electromechanical coupling. These include piezoelectric filters and electromechanical filters.

A filter will alter both the amplitude and the phase of the sinusoidal signal passing through it. For audio applications, the effect on phase is seldom significant. Filters are classified by the general shape of the amplitude-frequency response into *low-pass filters*, *high-pass filters*, *band-pass filters*, and *band-stop filters*. These

designations are also used for digital and video filtering, but the effect on phase is also very important in these applications. A further designation is the *all-pass filter*, which affects only the phase, and not the amplitude, of the signal.

The names given refer to the shape of the amplitude part of the filter transfer function. The filter transfer function is defined as the ratio of output voltage to input voltage (current, but not power, could be used instead of voltage) for a sinusoidal input. Thus, if the input to the filter is a sine wave having an amplitude of x and a phase angle of θ_x , the output will also be a sine wave, but with a different amplitude and phase angle in general. Let y represent the output amplitude and θ_y the output phase; then the filter transfer function is

$$H(f) = |H(f)| \angle \theta = \frac{y}{x} \angle \theta_y - \theta_x \quad (1.13.1)$$

The modulus or amplitude-frequency part of the transfer function is $|H(f)|$, and this is sketched in Fig. 1.13.1(a) for the various kinds of filters listed above. The low-pass filter (LPF) is seen to be characterized by a passband of frequencies extending from zero up to some cut-off frequency f_c . Ideally, the response should drop to zero beyond the cut-off, but in practice there is a transition region leading to the edge of the stopband at f_s . The stopband is the region above f_s where the transmission through the filter is ideally zero. Again, in practice there will be a finite attenuation in the stopband and, also, ripple may be present in both the passband and stopband, as shown in Fig. 1.13.1(a).

The high-pass filter (HPF) characteristic is shown in Fig. 1.13.1(b). Here, the stopband is from zero up to some frequency f_s , the transition region from f_s up to the cut-off frequency f_c , and the passband from f_c onward. As with the LPF, ripple may appear in both the stopband and the passband.

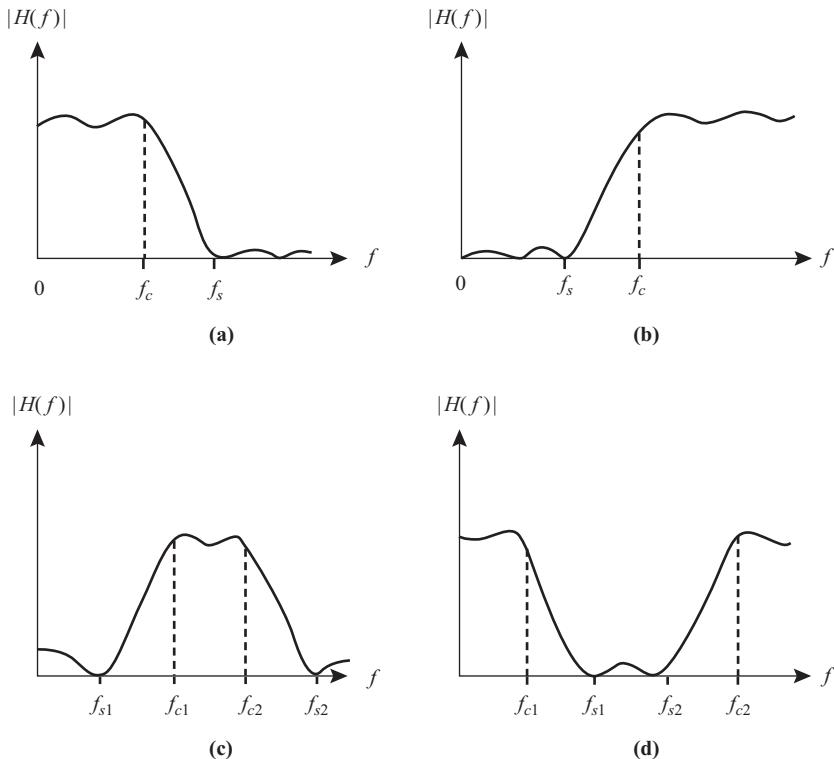


Figure 1.13.1 Amplitude response for basic filter designations: (a) low-pass, (b) high-pass, (c) band-pass, and (d) band-stop.

The band-pass filter (BPF) characteristic is shown in Fig. 1.13.1(c). The passband is seen to be defined by two cut-off frequencies, a lower one at f_{c1} and an upper one at f_{c2} . There is a lower transition region leading to a lower stopband frequency limit f_{s1} . The lower stopband is from zero up to f_{s1} . At the other end, the upper transition region leads from f_{c2} to f_{s2} , and then the upper stopband extends from f_{s2} upward. The coupled tuned circuit response shown in Fig. 1.8.3 is an example of a band-pass response.

The band-stop filter (BSF), or band-reject filter, response is shown in Fig. 1.13.1(d). This has a lower passband extending from zero to f_{c1} , a lower transition region extending from f_{c1} to f_{s1} , a stopband extending from f_{s1} to f_{s2} , and then an upper transition region extending from f_{s2} to f_{c2} and an upper passband extending upward from f_{c2} .

A number of well-established filter designs are available, each design emphasizing some particular aspect of the response characteristic. Although these designs apply to all the categories mentioned previously, they will be illustrated here only with reference to the low-pass filter. In the following sections the response curves are normalized such that the maximum value is unity.

Butterworth response. The modulus of the Butterworth response is given by

$$|H(f)| = \frac{1}{\sqrt{1 + (f/f_c)^{2m}}} \quad (1.13.2)$$

This gives what is termed a maximally flat response. The response is sketched in Fig. 1.13.2(a). The order of the filter is m , an integer, and the filter response approaches more closely to the ideal as m increases.

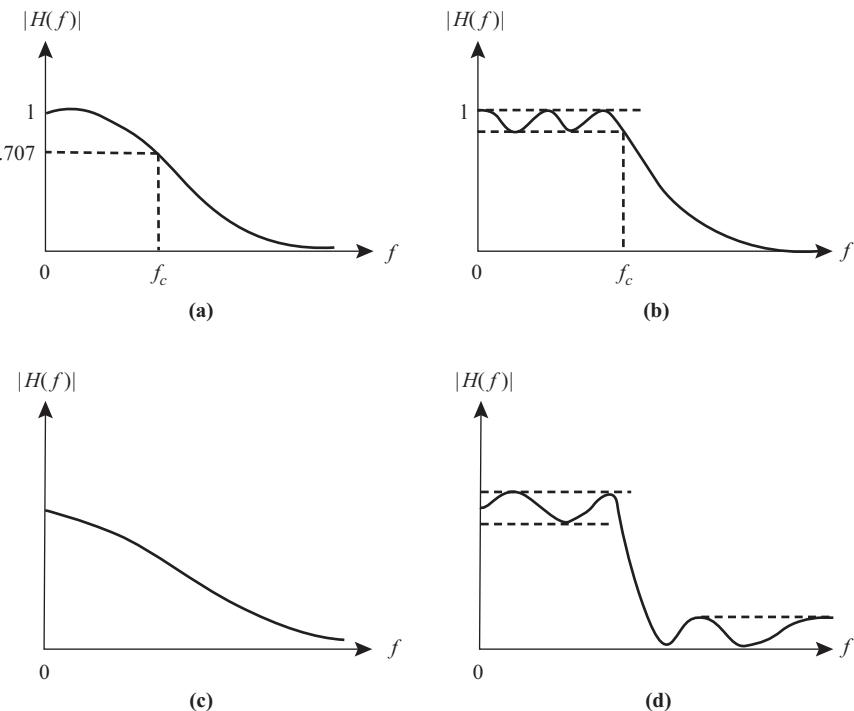


Figure 1.13.2 Sketches of the amplitude/frequency responses of several types of low-pass filters: (a) Butterworth; (b) Chebyshev; (c) maximally flat time delay (MFTD); (d) Cauer, or elliptic.

Whatever the order of the filter, it will be seen from Eq. (1.13.2) that at the cut-off frequency $f = f_c$ the response is reduced by $1/\sqrt{2}$, or -3 dB. Thus, at the cut-off frequency the response is not abruptly “cut off.” The simple RC low-pass filter is an example of a first-order Butterworth filter.

Chebyshev (or Tchebycheff) response. The Chebyshev response is given by

$$|H(f)| = \frac{1}{\sqrt{1 + \varepsilon^2 C_m^2(f/f_c)}} \quad (1.13.3)$$

Here, $C_m(f/f_c)$ is a function known as a Chebyshev polynomial, which for $-1 \leq f/f_c \leq 1$ is given by $\cos(m \cos^{-1}(f/f_c))$. This is a rather formidable expression, but it can be seen that, for the range of f/f_c specified, the Chebyshev polynomial oscillates between ± 1 . This produces an equiripple response in the passband, and the coefficient ε can therefore be chosen to make the ripple as small as desired. The order of this filter response is also m , and this controls the sharpness of the transition region. The Chebyshev response is sketched in Fig. 1.13.2(b). It will be noticed that the cut-off frequency in this case defines the ripple passband. For $f/f_c > 1$, the Chebyshev polynomial is given by $\cosh(m \cosh^{-1}(f/f_c))$.

Maximally flat time delay response. This type of filter is designed not for sharp cut-off, but to provide a good approximation to a constant time delay or, equivalently, a linear phase-frequency response. In other words, the phase response, rather than the amplitude response, is of more importance. Such filters are required when handling video waveforms and pulses. The amplitude response is a monotonically decreasing function of frequency, meaning that it always decreases as frequency increases from zero, as sketched in Fig. 1.13.2(c).

The Cauer (or elliptic) filter. The filter response $H(f)$ can be written generally as the ratio of two polynomials in frequency, $N(f)/D(f)$. For the filters described so far, the numerator $N(f)$ is made constant, and the filter response is shaped by the frequency dependence of the denominator $D(f)$. In the Cauer filter, both the numerator and denominator are made to be frequency dependent, and although this leads to a more complicated filter design, the Cauer filter has the sharpest transition region from passband to stopband. Often, in telephony applications a sharp transition band is the most important requirement. The term *elliptic filter* is also widely used for this type of filter and comes about because the response can be expressed in terms of a mathematical function known as an elliptic function. The amplitude response of the Cauer filter has ripple in both the passband and the stopband, as sketched in Fig. 1.13.2(d).

LC Filters

A low-pass LC filter network is shown in Fig. 1.13.3(a). This can be considered as being made up from two end sections and an intermediate section as shown in Fig. 1.13.3(b). The filter can be extended by increasing the number of intermediate sections, leaving the end sections unchanged. Extending the filter in this way sharpens the cut-off or transition region of the response.

The filter can also be constructed from π sections as shown in Fig. 1.13.3(c), this being similar to the T and π equivalences used with attenuators Figs. 1.12.2 and 1.12.3. It is left as an exercise for the student to develop the intermediate section and end sections equivalent for the π -network.

A high-pass filter network is shown in Fig. 1.13.3(d), and this can also be made up from an intermediate T section and two end sections. A π -type equivalent circuit can also be constructed.

A quick way of determining whether a filter is low-pass or high-pass is to examine the transmission paths at dc and at very high frequencies. For the circuit of Fig. 1.13.3(a), a dc path exists between input and output, and high-frequency transmission will be blocked by the L_1 inductors. Therefore, the filter is low-pass.

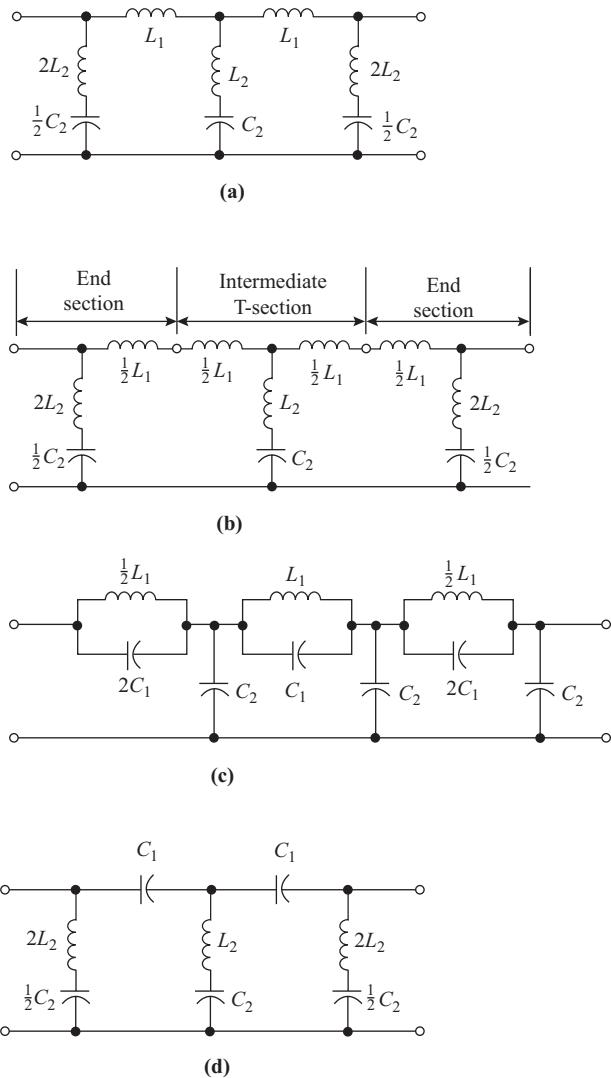


Figure 1.13.3 (a) Low-pass filter circuit. (b) Equivalent circuit consisting of two end sections and one intermediate T-section. (c) Low-pass filter circuit with an intermediate π -section. (d) High-pass filter circuit with an intermediate T-section.

For Fig. 1.13.3(c), a dc path also exists between input and output, while at high frequencies the C_2 capacitors will shunt the signal to ground. Therefore, this is also a low-pass filter. For Fig. 1.13.3(d), there is no dc path between input and output, and at high frequencies (above the series resonance of L_2C_2) the C_1 capacitors allow signal transmission, and therefore this is a high-pass filter.

Piezoelectric Crystal Filters

Piezoelectric crystals exhibit the property that, when an electric potential is applied across the faces of the crystal, it physically bends or deforms. Conversely, when the same crystal is mechanically deformed by pressure, an electric potential is developed between the crystal faces. The crystal also exhibits the phenomenon of

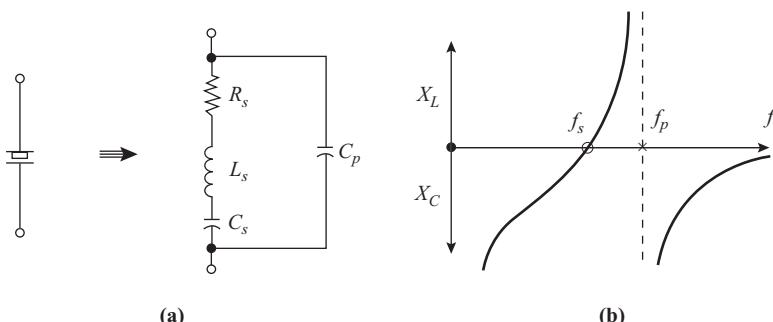


Figure 1.13.4 Quartz piezoelectric crystal: (a) the graphic symbol and the equivalent circuit of a quartz crystal; (b) variation of crystal terminal reactance with frequency.

mechanical resonance when it is excited with an alternating potential of the correct frequency. The frequency of mechanical resonance is determined by the size and shape of the crystal sample in question and can be controlled over several orders of magnitude, from about 20 kHz to about 50 MHz, with considerable precision. In form, the packaged crystal is a slice of crystal cut in such a way as to give the desired mechanical resonant frequency, with electrodes deposited on opposite sides so that a capacitive device is made.

Electrically, the mechanical resonance of this device makes the crystal look like a very high Q series resonant circuit, with a capacitor in parallel with it. This capacitor causes a second parallel resonance, which occurs at a frequency that is very close to the mechanical resonant point. The reactance of a quartz crystal is plotted in Fig. 1.13.4 and shows that, for low frequencies up to the series mechanical resonance, the crystal is capacitive. For frequencies between the series resonant and parallel resonant points, the reactance is inductive, and for frequencies above the parallel resonance, the reactance is again capacitive. At series resonance $X_{Ls} = X_{Cs}$ and the reactance is zero, and at parallel resonance $X_{Ls} = (X_{Cs} \text{ ser. } X_{Cp})$ and the reactance is infinite. The resonant frequencies of the crystal are very well defined and very stable, provided that the operating temperature is kept constant, making it very well suited as the high- Q resonant circuit that controls the operating frequency of oscillator circuits.

The reactance characteristic of the quartz crystal is changed radically by placing an inductance in parallel with it. The series resonant frequency remains unchanged, but the parallel resonant frequency is moved higher, so the separation between the two is greater than is the case for the crystal by itself.

Placing an inductor in series with the crystal has similar drastic effects on the reactance characteristic. In this case, however, the parallel resonant frequency remains unchanged, while the series resonant frequency is caused to move lower and a second series resonant frequency is created.

The frequency separation between the series and parallel resonant frequencies of the crystal itself is small, on the order of a few hundred hertz at most for a 1-MHz crystal. Frequency spreading by means of series or parallel inductors can increase this separation to a few thousand hertz, making it possible to use the crystals as band-pass filter elements for IF amplifiers and for sideband separation.

The crystal gate shown in Fig. 1.13.5(a) is a narrow-band sharp-cut-off filter circuit that makes use of the reactance characteristic of the crystal itself. It has been used for separating the sidebands in single sideband (SSB) circuits. When the capacitance of C_2 is relatively large, a high-pass sharp-cut-off filter with the characteristics of Fig. 1.13.5(b) is formed. At the frequency f_∞ , the reactance of the crystal is capacitive and equal in magnitude to that of the capacitor C_2 , so the signal fed to the output through the crystal is equal in magnitude and opposite in phase to that fed through C_2 , causing a complete cancellation at the output. At frequency f_0 the reactances are again equal in magnitude, but this time the crystal is inductive. The signal through the crystal is

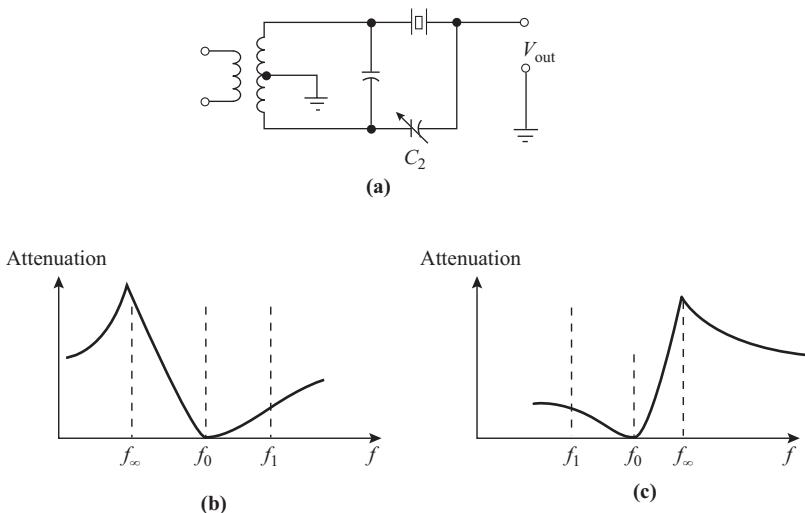


Figure 1.13.5 Crystal gate: (a) the circuit; (b) plot of attenuation versus frequency for the case where X_{C2} is small, giving a high-pass characteristic; (c) plot of attenuation versus frequency for the case where X_{C2} is large, giving a low-pass characteristic.

shifted by 90° , while that through the capacitor is shifted by -90° , so both arrive at the output in phase with each other. For frequencies above f_0 up to f_1 , just below the parallel resonant frequency of the crystal, the attenuation remains low as the signal is propagated through the capacitor C_2 . Severe phase-shift distortion occurs near f_p , so the usable passband is only between f_0 and f_1 . The cut-off beyond f_1 is quite gentle, but f_0 and f_∞ are only separated by a few hundred hertz, providing a sharp lower cut-off. Frequency shifting by a series inductor can be used to increase the passband to usable widths.

When the reactance of C_2 is made considerably higher, a low-pass filter with the characteristic of Fig. 1.13.5(c) results. Again, at f_0 the crystal reactance is equal to that of the capacitor and inductive, so the signals arrive at the output in phase with each other. At f_∞ the crystal reactance is equal to that of the capacitor and capacitive, resulting in complete signal cancellation, this time at a frequency higher than f_0 (and f_p). For frequencies below f_0 down to f_1 near f_s , the attenuation is low, determined by X_{C2} . Phase-shift distortion near f_s prevents use of frequencies below this, and again the usable band-pass can be increased by using a series inductor.

The crystal gate is inexpensive to build and uncritical in its adjustment, making it attractive, but it suffers from the disadvantage of providing a very narrow usable passband width. The crystal lattice filter is a more complicated circuit, but it provides bandwidths of a few hundred hertz to several tens of kilohertz and essentially flat response characteristics within the passband. Furthermore, sharp cut-off can be provided on both the upper and lower edges of the passband.

Figure 1.13.6(a) shows the circuit of a full-lattice filter, and Fig. 1.13.6(b) shows a half-lattice filter. The attenuation characteristics of both are the same and are shown in Fig. 1.13.6(c). The crystals used in the lattice are matched pairs, so crystals CR_1 and CR_2 are identical, and CR_3 and CR_4 are identical, but different from CR_1 and CR_2 . The crystals are chosen so that the series resonant frequency of CR_3 and CR_4 coincides with the parallel resonant frequency of CR_1 and CR_2 . The inductances of the coils in the input and output circuits are effectively in parallel with the crystals and act to space out the separation between series and parallel resonance and provide the second parallel resonance of each crystal.

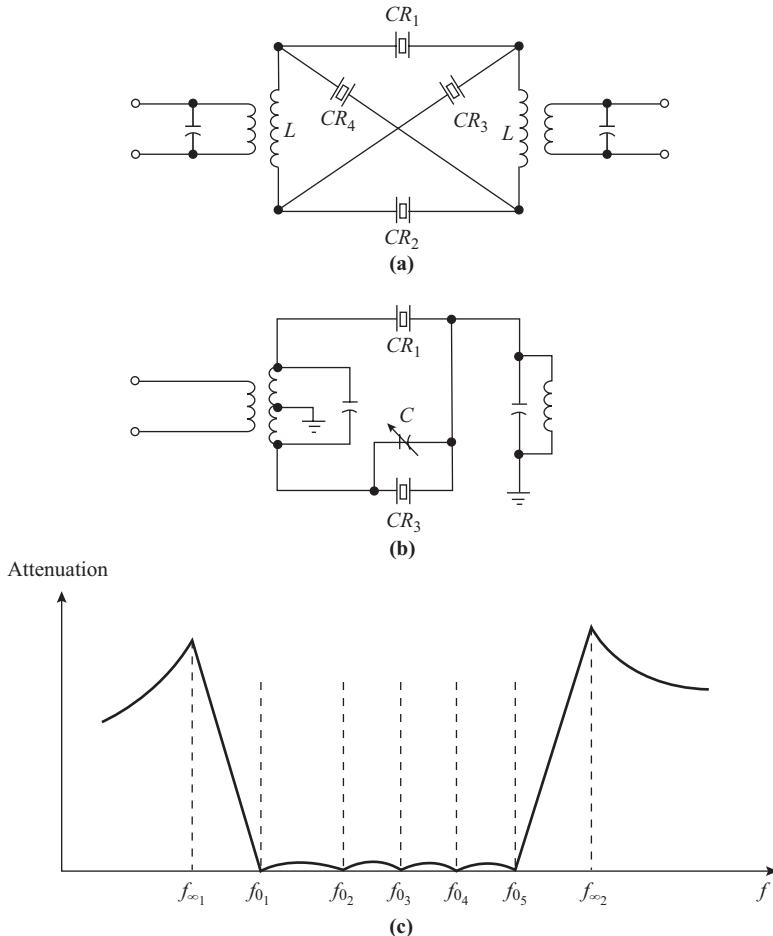


Figure 1.13.6 (a) Full-lattice crystal filter circuit. (b) Half-lattice crystal circuit. (c) Attenuation versus frequency for lattice filter.

At f_{∞_1} the reactances of X_1 and X_3 are both inductive and equal, so the in-phase and antiphase signals fed to the output cancel, providing infinite attenuation. This is very near the parallel resonant frequency of X_1 , and just past this frequency f_{0_1} occurs, where again the crystal reactances are equal, but X_1 is capacitive and X_3 is inductive, so the signals arrive at the output in phase with each other. f_{∞_1} and f_{0_1} delineate the lower transition band of the filter.

At f_{∞_2} the reactances X_1 and X_3 are equal but capacitive, so cancellation again takes place. This occurs just above and very near the upper parallel resonant frequency of X_3 . Just below f_{∞_2} , the reactances are again equal, but opposite, so the signals are again in phase, at f_{0_5} . Frequencies f_{∞_1} and f_{0_5} delineate the upper transition band of the filter.

At f_{0_3} the reactances X_1 and X_3 are again equal and opposite, so the signals arrive in phase. At frequency f_{0_2} , crystal 1 has zero reactance, shorting the input to the output. At frequency f_{0_4} , crystal 3 has zero reactance and again the output is connected directly to the input. Between these frequencies, the signal is fed to the output with very low attenuation values, providing the band-pass of the filter. Outside the band, attenuation is

relatively high and can be improved by providing additional gradual cut-off filtering in tandem, such as with a normal IF amplifier filter.

Surface Acoustic Wave Filters

The piezoelectric crystal described in the previous section depends for its operation on *bulk acoustic waves*, that is, mechanical vibrations that travel through the bulk of the solid. As the frequency of operation is increased, thinner crystals are required, and this sets an upper limit on frequency of about 50 MHz. It is also possible to set up *surface acoustic waves* (SAWs) on a solid, that is, mechanical vibrations that travel across the surface of the solid. Molecules on the surface actually follow an elliptical path that penetrates a short way into the bulk. In the case of piezoelectric material, a piezoelectric emf is generated at the surface, and this provides a means of coupling an electric signal into and out of the surface acoustic wave. The velocity of propagation of the surface acoustic wave is on the order of 3000 m/s. Since wavelength is related to frequency by $\lambda = v/f$, a frequency of, for example, 100 MHz will set up a surface wavelength of 30 μm ($1 \mu\text{m} = 10^{-6} \text{ m}$). The electrode structure on a SAW device requires spacings on the order of a wavelength, and thus very compact devices can be made. Since the action takes place on the surface, the bulk size can be chosen to provide mechanical strength without

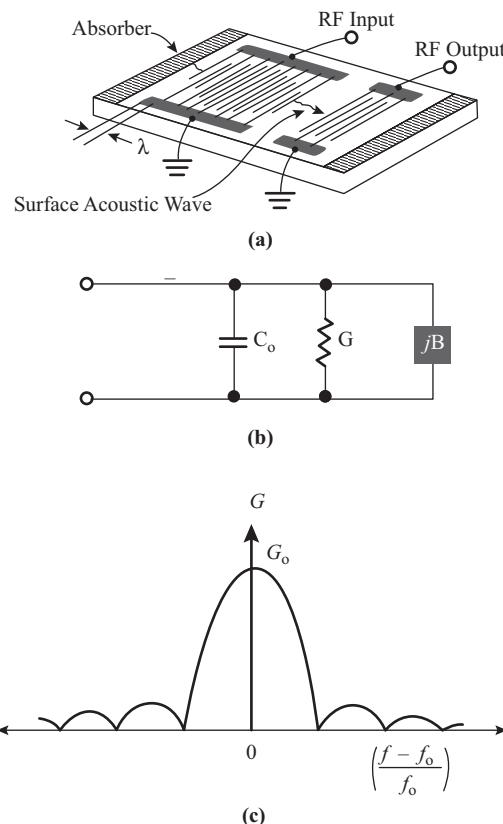


Figure 1.13.7 (a) Basic configuration for a surface acoustic wave delay line. The absorbing layer at each end reduces reflections from the edges of the substrate. (b) Equivalent circuit for an interdigital surface acoustic wave transducer (IDT). (c) Frequency response of the conductance component G . f_o is the center frequency of the filter. [(a) and (b) are courtesy of Waguish S. Ishak, H. Edward Karrer, and William R. Shreve, *Hewlett Packard Journal*, December 1981.]

interfering with the surface operation. The electrodes may be deposited on the surface using one of several well-established methods in production use for silicon integrated circuit fabrication.

Figure 1.13.7(a) shows the basic electrode configuration for a delay line filter. The electrode structure consists of interdigitated metallic stripes, with the spacing between stripes that are connected together being one wavelength λ long. This is the center wavelength, and the filter has a band-pass characteristic, the response falling off as the input frequency is shifted to either side of the center frequency. The actual shape of the response curve depends on the electrode configuration, and a range of amplitude and phase combinations is possible. However, the SAW filter characteristic is always band-pass.

The input and output electrode structures are seen to be similar and are referred to as *interdigital transducers* (IDTs). The surface acoustic wave generated by the input coupling travels out in both directions. Thus

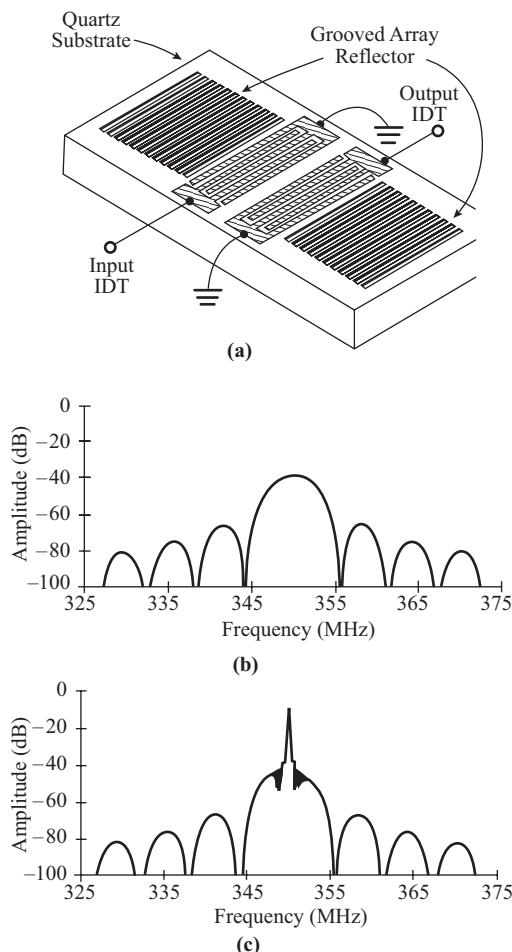


Figure 1.13.8 (a) Two-port surface-acoustic-wave resonator. The arrays of grooves at each end reflect the surface waves excited by the input IDT. The reflected waves constructively add at a frequency largely determined by the periodicity of the grooves. (b) Frequency response of a 350-MHz SAW delay line. By adding grating reflectors at each end, a resonant peak is obtained at the center frequency, shown in (c). (Courtesy, Peter S. Cross and Scott S. Elliott, *Hewlett Packard Journal*, December 1981.)

only part of the total surface acoustic wave reaches the output IDT. The outer part is absorbed or dissipated in an electrode placed at the edge for this purpose. This prevents reflected waves from occurring. The equivalent circuit for the IDT is shown in Fig. 1.13.7(b). The capacitance C_0 is fixed by the geometric structure and the dielectric constant of the substrate. The susceptance component jB is zero at the center frequency of the filter and shows a periodic variation with frequency. Just around the center frequency, this susceptance is inductive for higher frequencies and capacitive for lower frequencies. The conductance G is also a function of frequency, and this in fact largely determines the filter response. The variation of G with frequency is sketched in Fig. 1.13.7(c).

Resonators can also be constructed from SAW devices, one arrangement being shown in Fig. 1.13.8(a). The end absorbers are replaced by *grating reflectors*, which consist of an array of reflecting slots or grooves, the latter spaced $\lambda/2$ apart at the resonant frequency. Each groove reflects a small fraction of the incident surface acoustic wave, and the phasing is such that the individual reflections combine to give a peak at the output. The frequency response of a delay line filter using absorbers is shown in Fig. 1.13.8(b); that of one using grating reflectors is shown in Fig. 1.13.8(c).

Electromechanical Filters

A metal disk that is mechanically driven with an axially applied oscillating force will exhibit a resonant mode that is analogous to parallel resonance in an electrical circuit. Similarly, an axially driven rod will exhibit series resonance. When a series of disks and rods is interconnected to form a ladder network, and the resonances of the various components are carefully chosen, a band-pass filter with sharp cut-off characteristics results. Collins Radio has developed a mechanical filter of this type for use as sideband filters in communications receivers. These filters are electrically driven through magnetostrictive couplers and generally operate in the 100-kHz region. Bandwidths up to 5 kHz with very sharp cut-off characteristics are available. The unit is packaged in containers that are about 2 inches long and $\frac{1}{2}$ inch square and are arranged for printed-circuit-board mounting. The mechanical filter is pre-tuned at the factory, and no further adjustment is necessary or permitted, making it a very convenient component for receiver manufacture.

The ceramic filter is also a type of mechanical filter, but in this case the resonant components are in the form of disks of a piezoelectric ceramic such as barium titanate, arranged in a ladder configuration. The piezoelectric ceramic disks provide their own electromechanical conversion, and separate transducers are not necessary. In this respect the ceramic filter is more akin to the quartz crystal filter.

The ceramic filter is smaller than the mechanical filter by about half and has characteristics that are more nearly those of a complex crystal filter. They are also less complex and easier to manufacture than the mechanical filter and thus somewhat cheaper, which makes them very attractive for use in receivers. Ceramic filters are currently available in frequencies around 500 kHz and bandwidths ranging from 2 to 50 kHz.

PROBLEMS

- 1.1. Define the term *insertion loss* as applied to attenuators. A variable attenuator provides insertion loss in steps of 0.2 dB, ranging from 1 to 6 dB. Calculate the insertion loss as a current ratio at each step.
- 1.2. An emf source of internal resistance of $300\ \Omega$ is connected to a $75\text{-}\Omega$ load through a basic attenuator consisting of a $1000\text{-}\Omega$ resistor in series with the input and a $30\text{-}\Omega$ resistor in parallel with the output. Determine the insertion loss in decibels. Does this attenuator provide matching?
- 1.3. Determine the resistor values for a T-attenuator that must provide 10-dB insertion loss between a $75\text{-}\Omega$ source and a $50\text{-}\Omega$ load, while maintaining input and output matching.

- 1.4. Repeat Problem 1.3 for an insertion loss of 3 dB. Is this attenuator physically realizable?
- 1.5. A T-type attenuator has to provide 9-dB insertion loss between a $50\text{-}\Omega$ source and a $50\text{-}\Omega$ load. Determine the resistor values.
- 1.6. For the T-type attenuator of Example 1.2.1, show that the generator sees an effective load of $50\ \Omega$ and that the load is fed by a generator of equivalent internal resistance of $50\ \Omega$.
- 1.7. Design a 9-dB attenuator that has a symmetrical T configuration and that provides input and output matching for $600\ \Omega$.
- 1.8. Repeat Problem 1.7 for a symmetrical π -network configuration.
- 1.9. Repeat Problem 1.3 for a π network.
- 1.10. Repeat Problem 1.4 for a π network.
- 1.11. Repeat Problem 1.5 for a π network.
- 1.12. A symmetrical π -attenuator has resistor values $R_A = R_B = 144.4\ \Omega$ and $R_C = 106.7\ \Omega$. The attenuator is designed to work between a $75\text{-}\Omega$ load and a $75\text{-}\Omega$ source. Determine the insertion loss in decibels.
- 1.13. Derive Eqs. (1.2.30) and (1.2.31).
- 1.14. An L-attenuator is required to match a $300\text{-}\Omega$ source to a $175\text{-}\Omega$ load. Determine the resistor values and the insertion loss in decibels.
- 1.15. An L-attenuator is required to match a $75\text{-}\Omega$ source to a $330\text{-}\Omega$ load. Determine the resistor values and the insertion loss in decibels.
- 1.16. An inductor has a series resistance of $7\ \Omega$ and inductance of $75\ \mu\text{H}$. It forms part of a series tuned circuit that has a Q of 95. Determine the resonant frequency.
- 1.17. For a series tuned circuit, the resonant frequency is 1.3 MHz, the Q -factor is 100, and the tuning capacitance is 57 pF. Plot the impedance magnitude and phase angle as functions of frequency over a $\pm 20\text{-kHz}$ range about the resonant frequency.
- 1.18. A series tuned circuit has a Q of 130 and a tuning capacitance of 250 pF and is resonant at 450 kHz. Determine the impedance (a) at resonance, and (b) at frequencies $\pm 5\%$ off resonance.
- 1.19. Calculate the relative response of the circuit in Problem 1.18, as a ratio, and in decibels, at a frequency of 400 kHz.
- 1.20. Show that the variable introduced in Eq. (1.3.9) is given by $y \equiv \pm 2\Delta f/f_o$, where $\Delta f = |f - f_o|$, at frequencies close to resonance. Using this, plot the modulus and phase angle of the impedance of a series LRC circuit as a function of $\pm 2\Delta f/B_3$, where B_3 is the -3-dB band-width and $R = 1\ \Omega$.
- 1.21. Calculate the -3-dB bandwidth of the circuit in Problem 1.17.
- 1.22. A series tuned circuit is used as a wavetrap as shown in Fig. 1.3.4. Given that the resonant frequency of the circuit is 2 MHz and the Q -factor is 95, determine the rejection ratio of the unwanted signal relative to a wanted signal at 2.2 MHz. The load resistance can be assumed sufficiently high to be ignored.
- 1.23. Repeat Problem 1.22 for a load resistance of $1200\ \Omega$, where the tuning capacitance is 100 pF.
- 1.24. The inductor and capacitor in Problem 1.18 are connected in parallel. Find the parallel resonant frequency. Does this differ significantly from the series value?
- 1.25. A parallel resonant circuit is formed using a 100-pF capacitor and an inductor that has a series resistance of $40\ \Omega$ and inductance of 5 mH. Determine the impedance at resonance. Given that the resonant current is $3\ \mu\text{A}$, calculate the current through the capacitor.
- 1.26. A series tuned circuit has a Q -factor of 35. Assuming that all the losses are in the coil, given by $r = 2\ \Omega$, determine the resonant impedance of the same components connected as a parallel tuned circuit.

- 1.27. A parallel tuned circuit is resonant at 10.7 MHz with a capacitance of 230 pF and a Q -factor of 120. Determine the magnitude and phase angle of the circuit at a frequency of 10 MHz.
- 1.28. The circuit of Problem 1.22 is rearranged as a parallel wavetrap as shown in Fig. 1.4.2, the circuit being retuned to resonate at 2.2 MHz. The other values remain unaltered. Calculate the rejection ratio in this case.
- 1.29. Repeat Problem 1.22 with the tuned circuit connected for parallel resonance and a load resistance of 1200 Ω . The tuning capacitance is 100 pF.
- 1.30. A coil is self-resonant at 20 MHz and has a Q -factor of 200 (neglecting self-capacitance), which may be assumed constant. Calculate the effective Q -factor at frequencies of 5 and 10 MHz.
- 1.31. At frequencies well below the self-resonant frequency, a coil has a series resistance of 10 Ω and an inductance of 15 μH . The self-capacitance is 10 pF. The coil is used as an inductance in a series tuned circuit at a frequency of 1.3 MHz. Determine the effective inductance and the effective Q -factor.
- 1.32. The coil in problem 1.31 is used in a parallel tuned circuit, tuned to resonate at 5 MHz. Determine the value of the external tuning capacitor required, the effective inductance, and the effective Q -factor.
- 1.33. Explain briefly what is meant by skin effect and why it is undesirable. What steps may be taken to reduce skin effect in inductors?
- 1.34. Two windings are arranged such that the coefficient of coupling between them is 0.01. The self-inductances of the windings are 10 mH and 5 mH. Calculate the maximum and minimum inductance values obtainable from the combination.
- 1.35. The mutual inductance between two coils is 2 μH , the inductance of the secondary winding is 10 μH , and its resistance is negligible. A resistive load is placed across the secondary, the ratio of secondary inductive reactance to load resistance being unity. Given that the primary current is 1 mA, calculate the magnitude of the load current.
- 1.36. For a coupled circuit the Q -factor of the primary alone (secondary isolated) is 100 when tuned to resonate at 1 MHz, with a tuning capacitance of 200 pF. The self-inductance of the secondary is 0.13 mH, and its resistance may be neglected. The load resistance is 5 k Ω , and the coefficient of coupling is 0.2. Determine the effective load referred to the primary at the primary resonant frequency.
- 1.37. A parallel tuned circuit is resonant at 10.7 MHz with a tuning capacitor of 200 pF. The Q -factor is 150. The circuit is loosely coupled to a 50- Ω resistive load through an untuned mutual inductive coupling loop of self-inductance 0.1 μH and $k = 0.1$. Determine (a) the dynamic impedance of the circuit at resonance, and (b) the -3-dB bandwidth.
- 1.38. Explain briefly why an overcoupled tuned transformer should show two peaks. How can the response curves of tuned transformers be combined to produce a flat-topped response with sharply falling sides?
- 1.39. A synchronously tuned transformer has identical primary and secondary circuits, for which the following apply: $Q = 100$, $C = 200 \text{ pF}$, $f_o = 1 \text{ MHz}$, and $kQ = 1$. Determine the voltage transfer function at f_o and the -3-dB bandwidth. The damping effects of source and load resistances may be ignored.
- 1.40. For the circuit in Problem 1.39, plot the frequency response curve, in decibels, for the transfer impedance normalized to its value at 1 MHz. Use a frequency range of $\pm 2\%$ about 1 MHz.
- 1.41. The circuit of Problem 1.39 is fed from a constant current source of 40 μA and internal resistance 100 k Ω . Determine the frequency at which the output voltage has zero phase angle and its value at this frequency. Determine also the -3-dB bandwidth.
- 1.42. Show that for the circuit of Fig. 1.9.1(b), when a is made large the transfer impedance reduces to R_o/a . Discuss the significance of this.

- 1.43.** A tapped inductor has $L_a = 100 \mu\text{H}$, $L_s = 9 \mu\text{H}$, and $k = 1$ between sections. The load resistance connected to the tapping point is 300Ω . Calculate the effective parallel resistance and the effective inductance of the circuit. The coil series resistance may be ignored.
- 1.44.** Two 10-pF capacitors are connected in series to form the tuning capacitor of a 100-MHz parallel resonant circuit. The dynamic impedance of the circuit is $1 \text{ M}\Omega$. A $10\text{-k}\Omega$ load is connected across one of the capacitors. Determine (a) the effective dynamic impedance of the loaded circuit, and (b) the effective tuning capacity.
- 1.45.** For the circuit of Fig. 1.10.1 (a), $C_1 = 70 \text{ pF}$, $C_2 = 150 \text{ pF}$, and the load resistance is 2000Ω . Determine the inductance value needed to resonate the circuit at 27 MHz and the effective load impedance presented to the current source at resonance. Assume the dynamic impedance of the circuit is high enough to be ignored.
- 1.46.** Determine the voltage transfer function for the circuit in Problem 1.45.
- 1.47.** Determine the transfer impedance for the circuit of Problem 1.45 given that the source resistance can be neglected.
- 1.48.** Explain what is meant by a conjugate match. A signal source can be represented by an equivalent voltage generator of internal series impedance $(50 + j10) \Omega$ at 200 MHz . Calculate the series RC values of load required for conjugate matching. The internal emf of the source is $3 \mu\text{V}$. Determine the maximum average power delivered to the load.
- 1.49.** To function properly, a circuit has to be connected to a $600\text{-}\Omega$ signal source. The actual source is a $50\text{-}\Omega$ microphone. Calculate the turns ratio of the matching transformer required. State any assumptions made.
- 1.50.** Explain what is meant by the transfer function of a filter. The input voltage to a filter is $5 \sin \omega t$, and the corresponding output voltage is $2 \sin(\omega t + 40^\circ)$. What is the value of the transfer function at this frequency?
- 1.51.** A square wave having the same amplitude and frequency as the sine wave input is applied to the filter of Problem 1.50. Can the value of the transfer function as determined in Problem 1.50 be used to determine the output for the square wave input?
- 1.52.** A second-order low-pass Butterworth filter has a -3-dB frequency of 1000 Hz . Calculate the output amplitude from the filter when the input is a 1-V sinusoid at a frequency of (a) 100 Hz , and (b) 5000 Hz . (c) What is the transfer function magnitude, in decibels, in each case?
- 1.53.** On the same set of axes, plot the amplitude response for a first-order and a second-order low-pass Butterworth filter over the range from 0 to -30 dB . Use a logarithmic frequency scale normalized to the -3-dB value.
- 1.54.** Calculate the amplitude response for a first-order Chebyshev LPF for which $\epsilon = 0.25$ at frequencies of (a) $0.5f_c$, (b) f_c , and (c) $5f_c$.
- 1.55.** Explain how the crystal lattice filter acts to provide band-pass filtering.
- 1.56.** Using the *freqs* command in MATLAB, plot the frequency response of the filter whose transfer function is given by: $H(s) = \frac{1}{s^2 + 2s + 1}$. (Hint: $a=[1 2 1]$; $b=[1]$; $\text{freqs}(b,a)$;
- 1.57.** Study the Chebychev and Butterworth filters as implemented in MATLAB. Obtain the numerator/denominator polynomials of a 2^{nd} order Butterworth low pass filter. (Hint: $[b,a]=\text{buttap}(2)$; $\text{freqs}(b,a)$;
- 1.58.** Repeat the above exercise for a 3^{rd} order type-1 Chebychev high pass filter, when the ripple in the pass band is 0.1dB . (Hint: $[b,a]=\text{cheb1ap}(3,0.1)$; $\text{freqs}(b,a)$;

- 1.59.** Obtain the transfer function of the high pass filter circuit with an intermediate T-section, as shown in Chapter 1, Figure 1.13.3(d). Plot the frequency response using MATLAB.
- 1.60.** If the load impedance is $3+4j\Omega$, then what should be the source impedance so that maximum power transfer takes place?
- 1.61.** For an m^{th} order Butterworth low pass filter, the filter response is given by: $|H_m(f)| = \frac{1}{\sqrt{1 + (f/f_c)^{2m}}}$. Plot $|H_m(f)|$ for various values of m using MATLAB and show that $|H_m(f)|$ approaches ideal case when $m \rightarrow \infty$.
- 1.62.** Obtain the frequency response of the circuit shown in Figure P1.62 analytically. What is the order of the filter? Show that it is a low pass filter using MATLAB.

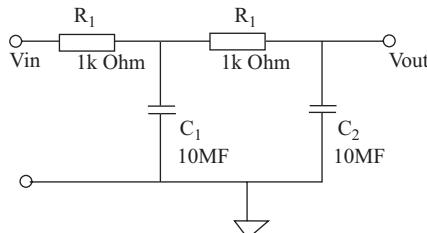


Figure P1.62

- 1.63.** Explore the MATLAB functions for transforming a low pass filter into an equivalent high pass filter.
- 1.64.** MATLAB can be used to obtain the mesh currents/voltages in an electrical circuit very conveniently. For example, if current–voltage relation in a circuit is given by $\mathbf{A} * \mathbf{i} = \mathbf{V}$, where \mathbf{i} is the mesh current vector ($= [i_1 \ i_2 \ i_3]^T$) and \mathbf{A} is the impedance matrix, $\begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}$ and \mathbf{V} is the mesh voltage vector, ($=[v_1 \ v_2 \ v_3]^T$), then, $\mathbf{i} = \text{inv}(\mathbf{A}) * \mathbf{V}$ or $\mathbf{i} = \mathbf{A} \mathbf{V}$. Use the above to find out the currents in circuits shown in Figure P1.64.

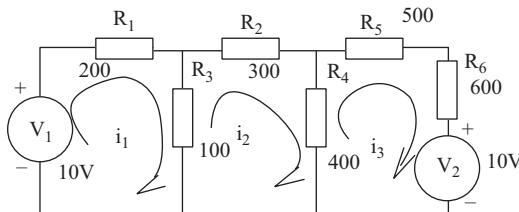


Figure P1.64

- 1.65.** Explore the *filter visualization tool (fvtool)* in MATLAB Ver 7.0.
- 1.66.** Explore the *filter design and analysis tool (fdatool)* in MATLAB Ver 7.0.



Waveform Spectra

2.1 Introduction

Wave motion is a familiar, everyday phenomenon, as observed in water waves, sound waves, heat waves, and so on. The idea of a wave implies some quantity that varies with distance and with time, and the waveform is generally taken to mean the graph of the quantity plotted as a function of either of these two variables. However, in this chapter, waveform will refer only to the time function.

The student will already be familiar with the sine and cosine wave functions encountered in steady-state ac analysis of circuits. The main properties of these trigonometric wave functions are reviewed in the following section in order to lay the groundwork for the study of waveform analysis. Waveform analysis consists of expanding a waveform (for example, a rectangular waveform or triangular waveform) into a trigonometric series, which forms the spectrum of the wave. A knowledge of the spectra of waveforms enables us to predict how the frequency response of a transmission system affects the waveform, as will be illustrated later in the text.

2.2 Sinusoidal Waveforms

A voltage waveform that is a sinusoidal function of time may be written as

$$v(t) = V_{\max} \sin 2\pi f_0 t \quad (2.2.1)$$

Here, the constants are V_{\max} , the peak value of the voltage, and f_0 , the frequency of the wave. The periodic time is $T_0 = 1/f_0$ as shown in Figure 2.2.1(a). The cosine wave is described by

$$v(t) = V_{\max} \cos 2\pi f_0 t \quad (2.2.2)$$

In analysis, using the cosine wave rather than the sine wave as a reference produces somewhat neater (but otherwise equivalent) results mathematically. The cosine wave is known as an even function because the

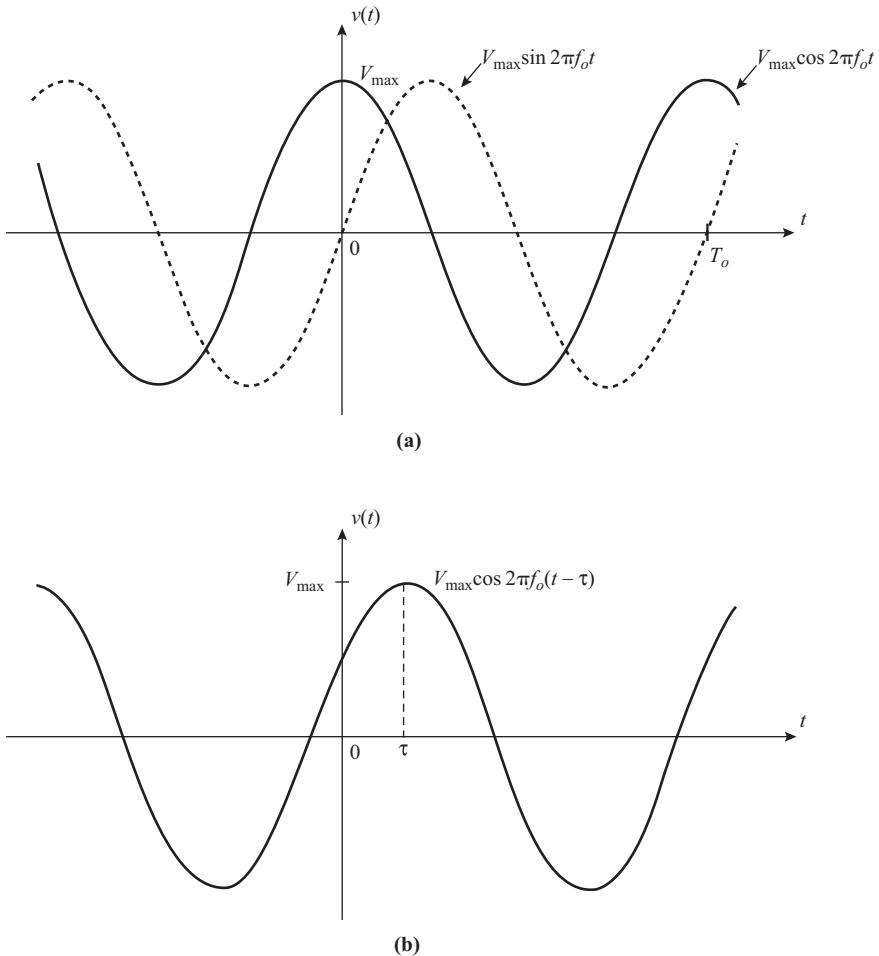


Figure 2.2.1 (a) Sine and cosine waves. (b) General cosinusoidal function.

curve is symmetrical about the vertical axis. [In mathematical notation, $v(-t) = v(t)$]. A sine wave is an odd function because the curve is skew-symmetrical about the vertical axis [or $v(-t) = -v(t)$].

More generally, the waveform can be as shown in Figure 2.2.1(b). Expressing this with reference to a cosine wave gives

$$v(t) = V_{\max} \cos 2\pi f_0(t - \tau) \quad (2.2.3)$$

In terms of a phase angle ϕ , the equation can be written as

$$v(t) = V_{\max} \cos (2\pi f_0 t + \phi) \quad (2.2.4)$$

It will be seen that this requires $\phi = -2\pi\tau/T_0$. A numerically negative value of ϕ results in a lagging phase angle, and a positive value a leading phase angle. For the sine wave, $\tau = T_0/4$, and so the sine wave lags the cosine wave by 90° . Physically, this means that the sine wave reaches its positive peak one-quarter cycle later than the cosine wave.

Sine and cosine waves are examples of periodic functions, which are described in the following section.

Exercise 2.2.1 Given that $\tau = -12.5 \mu\text{s}$ and $f_0 = 10^4 \text{ Hz}$, determine the angle of phase lead or lag in Eq. (2.2.4). (Ans. 45° lead.)

2.3 General Periodic Waveforms

A periodic function is one that repeats itself at regular intervals over the complete time domain, $-\infty \leq t \leq +\infty$. The periodic time is the time spanned by one repetition, and, for example, for the trigonometric functions described in the previous section the periodic time is denoted by T_0 between two successive peaks. The periodic time can be selected anywhere on the function; it is simply a matter of convenience to show it between peaks (or between zeros).

An example of a nonsinusoidal periodic waveform is shown in Fig. 2.3.1. Mathematically, a periodic function has the property that $v(t) = v(t + T_0)$, which simply states that for any time t the value at time $(t + T_0)$ is equal to the value at time t .

2.4 Trigonometric Fourier Series for a Periodic Waveform

As a preliminary example of a trigonometric expansion of a periodic wave, consider the series

$$v(t) = \sin(2\pi \cdot 100 \cdot t) + 0.3 \sin(2\pi \cdot 200 \cdot t) + 0.2 \sin(2\pi \cdot 300 \cdot t) \quad (2.4.1)$$

With time t in seconds, examination of the series shows that the first sinusoidal term has a frequency of 100 Hz. This is termed the *fundamental frequency*. The second sinusoidal term has a frequency of 200 Hz and is known as the *second harmonic*. The amplitude of the second harmonic is 0.3. The third sinusoidal term has a frequency of 300 Hz and is known as the *third harmonic*. Its amplitude is 0.2.

Exercise 2.4.1 Plot on a single graph the three trigonometric terms in Eq. (2.4.1), and on a separate graph plot the resultant wave. (Ans. The waveforms are plotted in Fig. 2.4.1, where it will be seen that the resultant is also periodic, with a periodic time of $\frac{1}{100} \text{ s.}$)

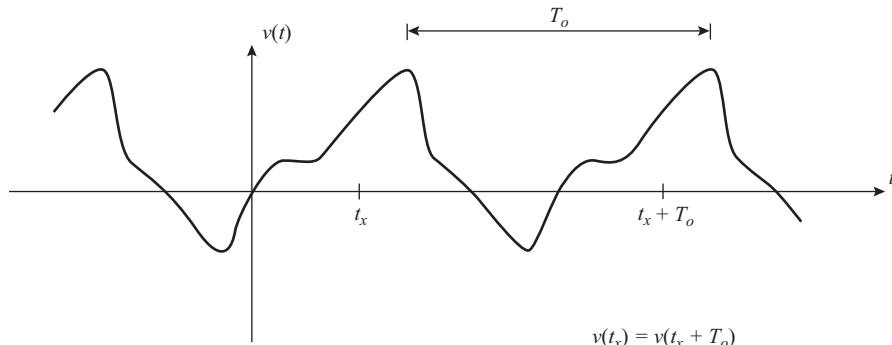


Figure 2.3.1 Nonsinusoidal periodic function.

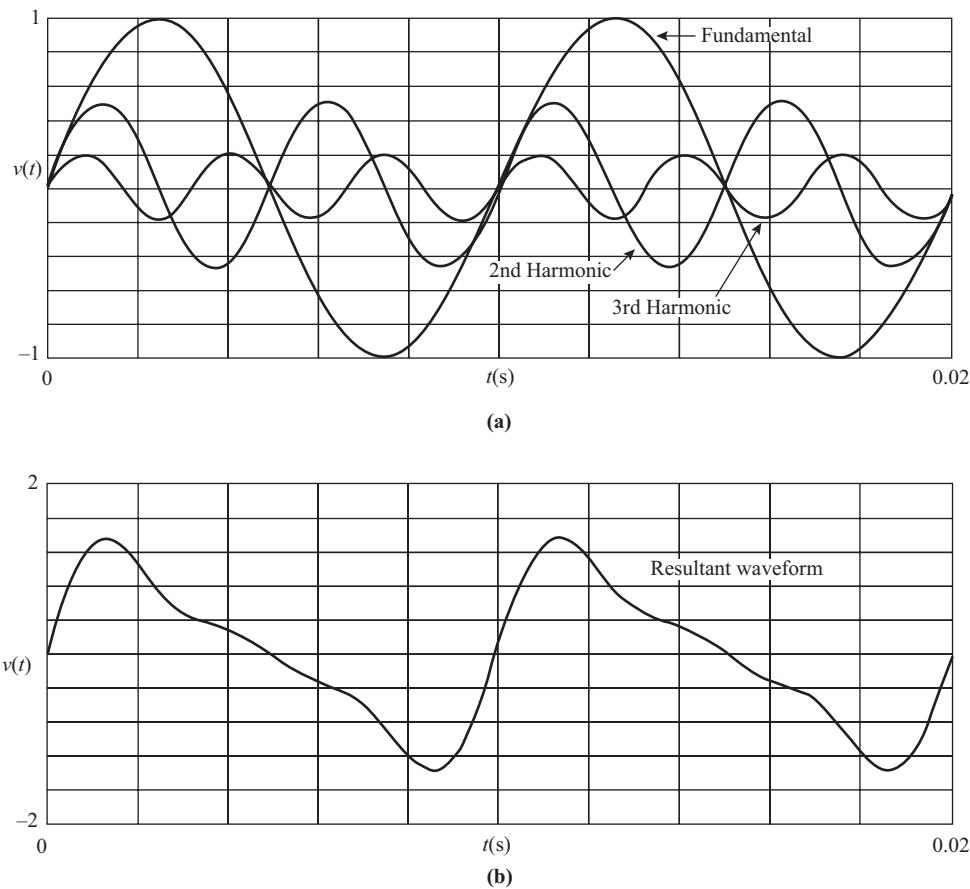


Figure 2.4.1 Waveform resulting from the series given in Eq. (2.4.1).

More generally, a periodic wave can be expanded in a trigonometric series containing sine and cosine terms. There may also be a constant term (representing the dc component in the case of voltage and current waveforms) so that the general trigonometric expansion may be written as

$$v(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos(2\pi n f_0 t) + \sum_{n=1}^{\infty} b_n \sin(2\pi n f_0 t) \quad (2.4.2)$$

The mean or dc component is represented by a_0 , and the harmonic number by n (the fundamental is taken to be the first harmonic). In theory, there may be an infinite number of harmonic terms, and so the summations allow for this possibility. The coefficients a_n and b_n represent the amplitudes or peak values of the trigonometric terms. Evaluating these coefficients is what waveform analysis is mostly about, and methods for doing so will be described later.

Although it is usually easier to evaluate the a_n and b_n coefficients separately, in use it is usually more informative to combine them in the following manner. Any two corresponding terms can be combined using the trigonometric identity as

$$a_n \cos(2\pi n f_0 t) + b_n \sin(2\pi n f_0 t) \equiv A_n \cos(2\pi n f_0 t + \phi_n) \quad (2.4.3)$$

For the identity to hold requires that

$$A_n = \sqrt{a_n^2 + b_n^2} \quad (2.4.4)$$

$$\phi_n = -\tan^{-1} \frac{b_n}{a_n} \quad (2.4.5)$$

A_n is taken as the positive value of the square root. Hence the original trigonometric series can be written as

$$v(t) = a_0 + \sum_{n=1}^{\infty} A_n \cos(2\pi n f_0 t + \phi_n) \quad (2.4.6)$$

The trigonometric series given by Eq. (2.4.6) is known as a Fourier series, named after its discoverer Joseph Fourier (1768–1830), a French mathematician.

2.5 Fourier Coefficients

Fourier showed that the coefficients can be found as follows:

$$a_0 = \frac{1}{T_0} \int_{T_0} v(t) dt \quad (2.5.1)$$

$$a_n = \frac{2}{T_0} \int_{T_0} v(t) \cos(2\pi n f_0 t) dt \quad (2.5.2)$$

$$b_n = \frac{2}{T_0} \int_{T_0} v(t) \sin(2\pi n f_0 t) dt \quad (2.5.3)$$

Here, the integral sign with the T_0 subscript means integration over one complete period. It will be observed that, to evaluate the coefficients using integration, the functional form of $v(t)$ must be known. Now there are many regular-type waveforms for which this is so and for which the Fourier coefficients have been evaluated and tabulated. Thus, where a waveform function is known and the Fourier coefficients are required, the student has the choice of evaluating the integrals or of referring to a handbook (for example the *Mathematical Handbook* by Murray R. Spiegel in the Schaum's outline series) for the results. Solving the integrals is a mathematical exercise, which, although useful, will not be followed here. Instead, some results for known waveforms will be presented after the idea of a spectrum has been introduced.

2.6 Spectrum for the Trigonometric Fourier Series

The right-hand side of the Fourier trigonometric series, Eq. (2.4.6), can be interpreted as a series of harmonic waves each of amplitude A_n and fixed phase angle ϕ_n and a dc component a_0 . These results can be presented graphically in what is known as a spectrum graph, as shown in Fig. 2.6.1.

Figure 2.6.1(a) shows the amplitude (or peak value) of each component and is known as the *amplitude spectrum* or *spectrum magnitude*. The phase angle of lead or lag is shown in Fig. 2.6.1(b). It will be seen that both amplitude and phase angle are plotted as functions of the harmonic frequencies, and hence all the information needed to specify the series is contained in the spectrum. The convention is that the cosine

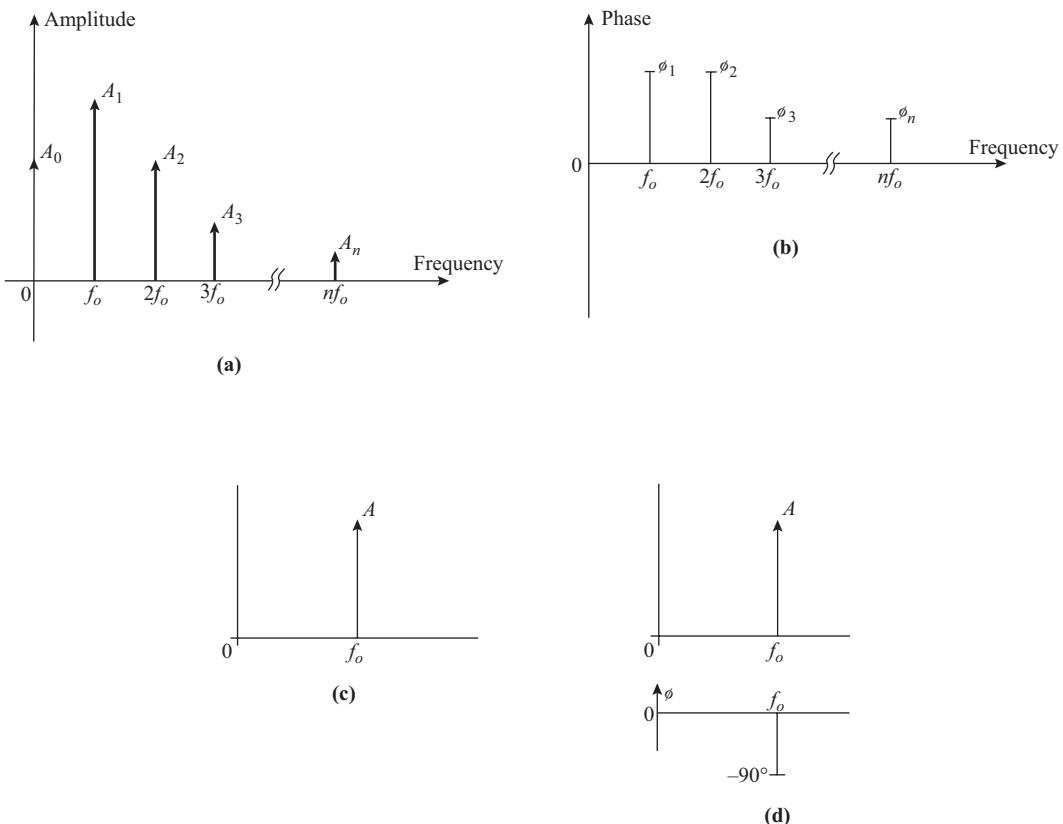


Figure 2.6.1 Spectrum for the series given in Eq. (2.4.6).

wave is used as reference, so, for example, the spectrum for $A \cos 2\pi f_0 t$ would be as shown in Fig. 2.6.1(c), and for $A \sin 2\pi f_0 t$ as in Fig. 2.6.1(d). The spectra for some known waveforms are described in the following sections.

2.7 Rectangular Waves

A rectangular voltage waveform (sometimes referred to as a square wave) is shown in Fig. 2.7.1(a). This is a periodic wave and is assumed to exist for all time, so the range for t is $-\infty \leq t \leq \infty$, just as for the sine and cosine waves. By choosing the zero time origin such as to make the waveform an even function, as shown in Fig. 2.7.1(a), only cosine terms appear in the trigonometric Fourier series.

Applying the equations given in Sections 2.4 and 2.5 results in the series

$$v(t) = \frac{4 \cdot V}{\pi} \left(\cos \omega_0 t - \frac{1}{3} \cos 3\omega_0 t + \frac{1}{5} \cos 5\omega_0 t - \frac{1}{7} \cos 7\omega_0 t + \dots \right) \quad (2.7.1)$$

Here, $\omega_0 = 2\pi f_0$ is the angular frequency in radians per second.

It will be seen that there is no dc component, and this should be apparent from the symmetry of the waveform about the time axis. The harmonic amplitudes are seen to be given by $|a_n| = 4V/(n\pi)$, where $n = 1, 3, 5, \dots$; that is, only odd harmonics are present.

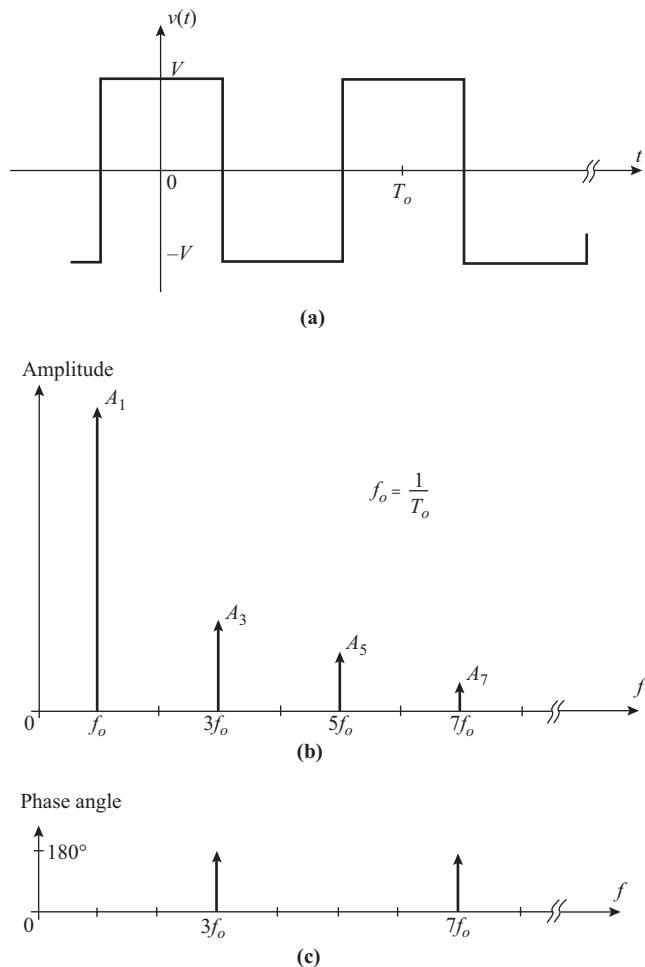


Figure 2.7.1 (a) Rectangular voltage waveform, (b) its magnitude spectrum, and (c) its phase spectrum.

Since $b_n = 0$, it follows from Eq. (2.4.4) that $A_n = |a_n|$, and therefore the magnitude of the spectrum is as shown in Fig. 2.7.1(b). Note that the negative sign associated with every other term in Eq. (2.7.1) is accounted for as a 180° phase shift, and so the phase angle part of the spectrum appears as shown in Fig. 2.7.1(c).

If the square wave is shifted so that it is not symmetrical about the horizontal (time) axis, then a dc component appears in the spectrum. This is simply the mean value of the voltage and is the value that would be read, for example, on a moving coil voltmeter. For example, if the waveform is moved up so that the lowest value is zero, as shown in Fig. 2.7.2(a), then by inspection the mean value is seen to be V .

The trigonometric Fourier series for this waveform is

$$v(t) = V + \frac{4 \cdot V}{\pi} \left(\cos \omega_0 t - \frac{1}{3} \cos 3\omega_0 t + \frac{1}{5} \cos 5\omega_0 t - \frac{1}{7} \cos 7\omega_0 t + \dots \right) \quad (2.7.2)$$

The spectrum is as shown in Fig. 2.7.2(b) and (c).

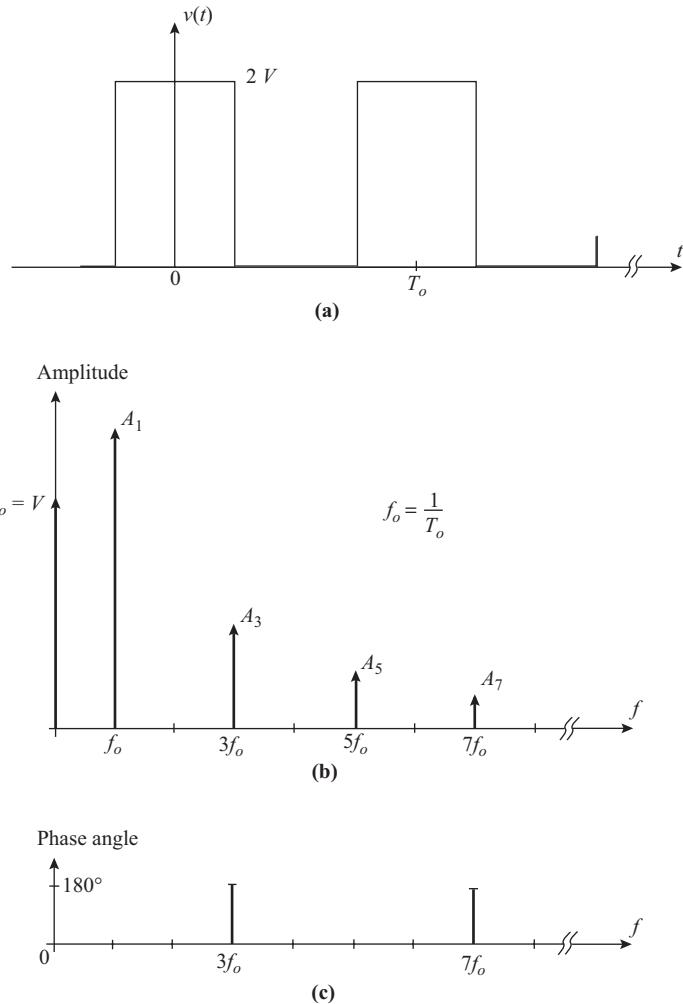


Figure 2.7.2 (a) Square wave with a dc component; (b) and (c) its spectrum.

A waveform can be reconstructed by graphically adding the trigonometric terms in the spectrum. (Adding components in this way to produce a waveform is known as *waveform synthesis*.) This can be a dauntingly tedious task unless carried out by computer. Even so, for many periodic functions the spectrum extends to infinity, and a complete graphical reconstruction is not possible. The following exercise illustrates the technique.

Exercise 2.7.1 Given that for a square wave $V = 1$ V and $f_0 = 1000$ Hz, use a computer to obtain the sum of the first seven harmonic terms of Eq. (2.7.1), and plot the resulting waveform over one complete cycle. (Ans. The result is shown in Fig. 2.7.3.)

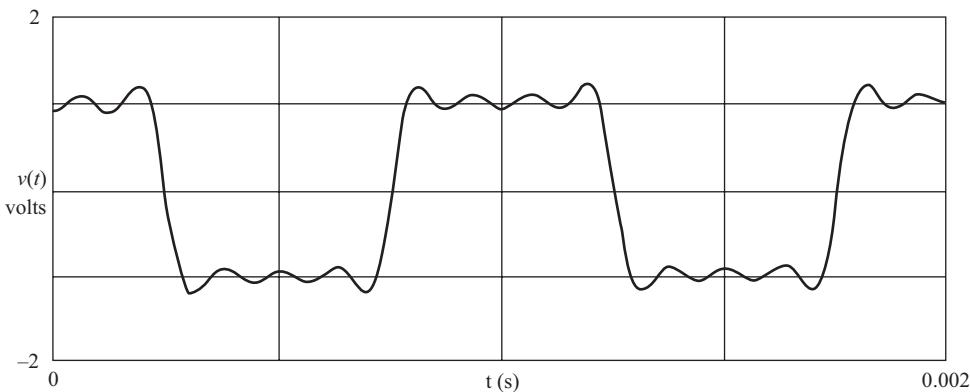


Figure 2.7.3 Square wave synthesis, using the first seven harmonic terms in the trigonometric series.

2.8 Sawtooth Waveform

Another waveform frequently encountered in practice is the sawtooth waveform shown in Fig. 2.8.1(a). The Fourier trigonometric series for this waveform is

$$v(t) = \frac{V}{2} - \frac{V}{\pi} \left(\sin \omega_0 t + \frac{1}{2} \sin 2\omega_0 t + \frac{1}{3} \sin 3\omega_0 t + \dots \right) \quad (2.8.1)$$

The dc component in this case is given by $V/2$ and the magnitude of the harmonics by $A_n = V/n\pi$. The harmonic terms are of the form $-\sin(n\omega_0 t)$, and so the phase lead relative to the cosine function is $(180 - 90) = 90^\circ$. This could also be shown as a 270° phase lag.

Note that when the dc component is zero, that is, the waveform is shifted down by amount $V/2$ to be symmetrical about the time axis, the waveform becomes an odd function, and only sine terms are present in the trigonometric Fourier series. This is confirmed by the harmonic terms in Eq. (2.8.1). The spectrum is shown in Fig. 2.8.1(b) and (c), where it will be seen that both odd and even harmonics are present. Note the need to distinguish carefully between even and odd functions, a mathematical property, and even and odd harmonies, a physical property.

2.9 Pulse Train

A periodic pulse train is shown in Fig. 2.9.1(a) for which the pulse width is τ . By choosing the zero time origin to make the waveform an even function only cosine terms are present. The Fourier series for this wave is

$$\begin{aligned} v(t) &= \frac{V\tau}{T_0} + \frac{2V}{\pi} \sum_{n=1}^{n=\infty} \frac{1}{n} \sin \frac{n\pi\tau}{T_0} \cdot \cos 2\pi n f_0 t \\ &= a_0 + \sum_{n=1}^{n=\infty} a_n \cdot \cos 2\pi n f_0 t \end{aligned} \quad (2.9.1)$$

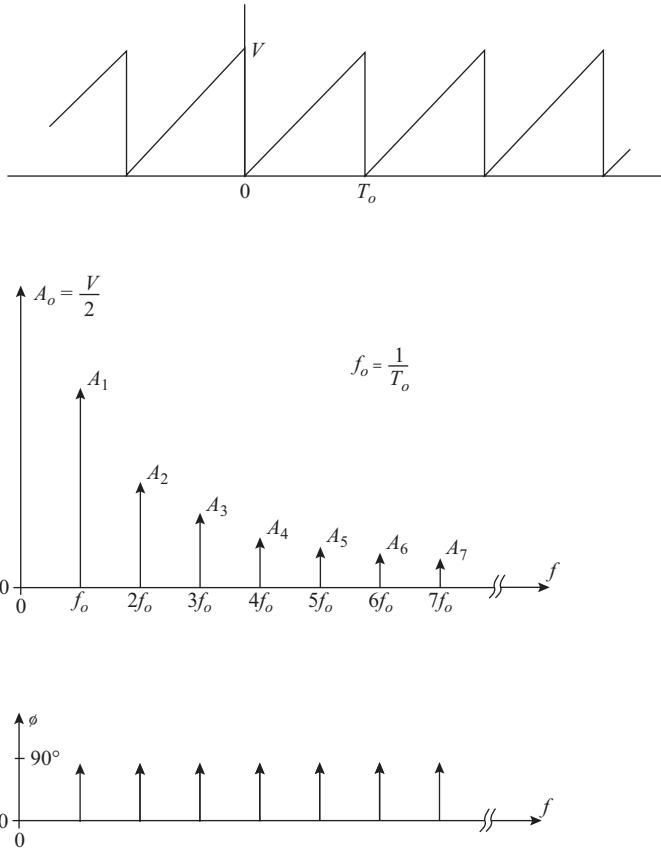


Figure 2.8.1 (a) Sawtooth voltage waveform, (b) its spectrum magnitude, and (c) the phase spectrum.

The dc component is seen to be given by

$$a_0 = \frac{V\tau}{T_0} \quad (2.9.2)$$

and the amplitude of the *n*th harmonic by

$$a_n = \frac{2V}{n\pi} \sin\left(\frac{n\pi\tau}{T_0}\right) \quad (2.9.3)$$

The spectrum magnitude is shown in Fig. 2.9.1(b), and the phase in Fig. 2.9.1(c).

In anticipation of later work, the form of the equation for *a_n* may be changed as follows. Let *x* = *nπτ*/*T₀*; then it is left as an exercise for the student to show that

$$a_n = \frac{2V\tau}{T_0} \cdot \frac{\sin x}{x} \quad (2.9.4)$$

The (*sin x*)/*x* function occurs frequently in spectrum studies. It is unity at *x* = 0 and zero for *x* = *kπ* for *k* = ±1, ±2, In the particular case of Eq. (2.9.4), *x* is a discrete variable, and *k* = *nτ*/*T₀* is not necessarily an integer.

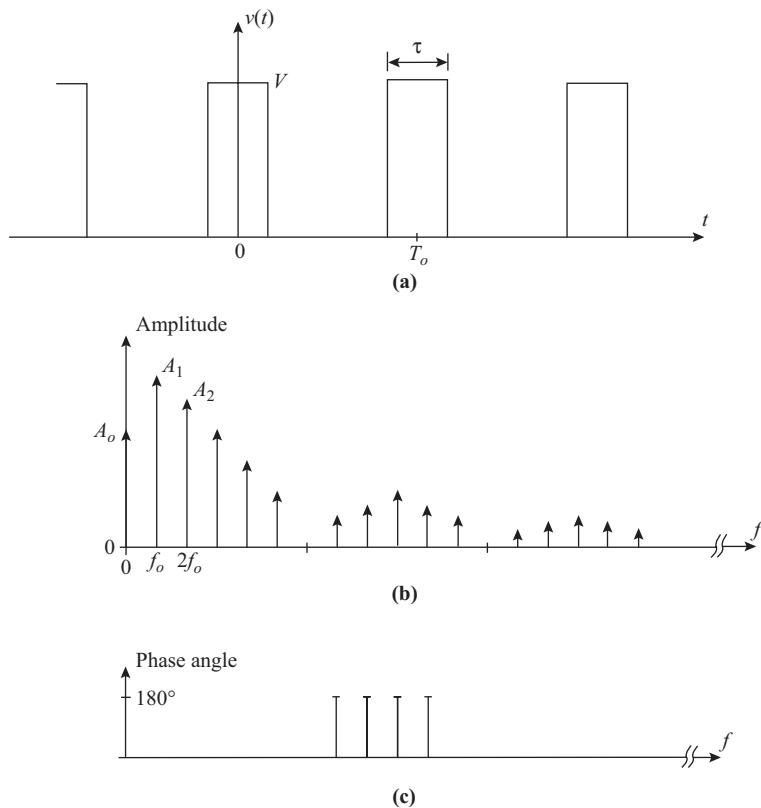


Figure 2.9.1 (a) Periodic pulse train, (b) its spectrum magnitude, and (c) the phase spectrum.

Exercise 2.9.1 Given that the duty cycle for the waveform of Fig. 2.9.1(a) is 0.1, determine the harmonic number for which the amplitude is first zero. (Ans. $n = 10$.)

Because of its significance in digital communications, the $(\sin x)/x$ function is given a special name. It is known as the *sampling function*, written as $\text{Sa}(x)$ where,

$$\text{Sa}(x) \equiv \frac{\sin x}{x} \quad (2.9.5)$$

Sometimes it is denoted by a related function known as the *sinc function*, written as $\text{sinc } k$, where

$$\text{sinc } k = \frac{\sin k\pi}{k\pi} \quad (2.9.6)$$

This is really just a difference in notation, but both functions are widely used in practice, and so it is best to be familiar with them. Both functions are available in graphical and tabular form. The sinc function will be used in this text, and Eq. (2.9.3) in terms of this is

$$a_n = \frac{2V\tau}{T_0} \cdot \text{sinc} \frac{n\pi}{T_0} \quad (2.9.7)$$

Exercise 2.9.2 Given that the duty cycle for the waveform of Fig. 2.9.1(a) is 0.1, determine the value of $\text{sinc } k$ for $n = 3$. (Ans. 0.858)

2.10 Some General Properties of Periodic Waveforms

When determining the spectra for the preceding waveforms, certain properties were deduced from the symmetry of the wave. Some of these properties and others are summarized below:

1. If one cycle of the waveform has equal areas above and below the time axis as sketched in Fig. 2.4.1(b) there will be no dc term.
2. If the waveform is symmetrical about the vertical axis as sketched in Fig. 2.7.3 the trigonometric expansion will contain only cosine terms. This symmetry requires that the waveform be an even function, or $v(-t) = v(t)$.
3. If the ac component of the waveform is skew-symmetric about the vertical axis as sketched in Fig. 2.4.1(b) the trigonometric expansion will contain only sine terms. This symmetry requires that the ac component be an odd function, or $v(-t) = -v(t)$.
4. If the waveform has finite discontinuities (such as the transitions from one level to another in a square wave, or the abrupt changes at the peaks of triangular waves), then the spectrum will contain an infinite number of harmonics. These decrease in amplitude at least as fast as $1/n$.

2.11 Exponential Fourier Series

The cosine of an angle can be written in terms of exponentials as

$$\cos \theta = \frac{e^{j\theta} + e^{-j\theta}}{2} \quad (2.11.1)$$

Thus an alternative way of writing the cosine wave of Eq. (2.4.3) is

$$\begin{aligned} v(t) &= A_n \cos (2\pi n f_0 t + \phi_n) \\ &= A_n \frac{e^{j(2\pi n f_0 t + \phi_n)} + e^{-j(2\pi n f_0 t + \phi_n)}}{2} \\ &= \frac{A_n}{2} e^{j\phi_n} \cdot e^{j2\pi n f_0 t} + \frac{A_n}{2} e^{-j\phi_n} \cdot e^{-j2\pi n f_0 t} \\ &= c_n e^{j2\pi n f_0 t} + c_{-n} e^{-j2\pi n f_0 t} \end{aligned} \quad (2.11.2)$$

where $c_n = \frac{A_n}{2} e^{j\phi_n}$ and $c_{-n} = \frac{A_n}{2} e^{-j\phi_n}$. It will be observed that c_{-n} is the complex conjugate of c_n .

The amplitudes of these two exponential terms are the same at $|c_n| = A_n/2$ and the phase angles differ by 180° . By showing the positive exponential as a phasor at frequency $+n f_0$ and the negative exponential at frequency $-n f_0$, the exponential form of the spectrum is created, as shown in Fig. 2.11.1. This is known as a *double-sided spectrum*, because it utilizes positive and negative frequencies. It should be understood, however, that these components must always come in pairs to create the corresponding trigonometric term at the real (positive) frequency $n f_0$.

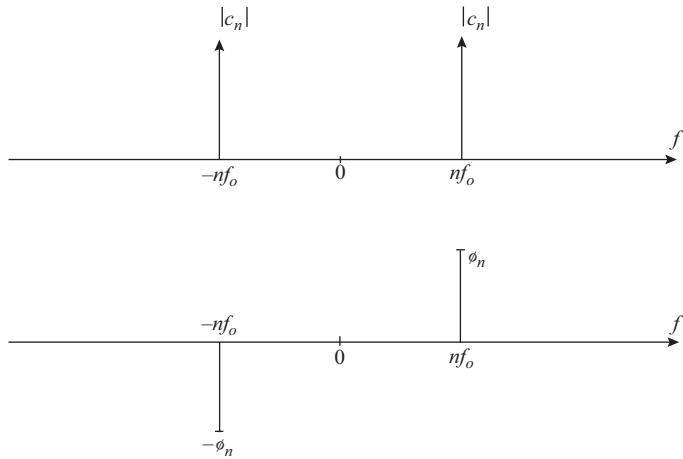


Figure 2.11.1 Double-sided spectrum for a cosine wave.

With one additional modification, the trigonometric Fourier series of Eq. (2.4.6) can be written in a very compact exponential form. The modification is to interpret the dc component as a cosine term of zero frequency and zero phase angle, which then allows the equation for $v(t)$ to be written as

$$v(t) = \sum_{n=-\infty}^{n=\infty} c_n e^{j2\pi n f_0 t} \quad (2.11.3)$$

The exponential Fourier series is widely used in more advanced texts for the ease with which it permits mathematical manipulations to be carried out. Such manipulations will not be required in this text, but, equally importantly, the exponential Fourier series forms the basis for most of the *fast Fourier transform* (fft) computer programs used to determine the Fourier coefficients. These programs are widely available, and to utilize them efficiently, a knowledge of the background to the method is needed.

It may be shown by straightforward manipulation of terms that $c_n = (a_n - jb_n)/2$ and that the integrals given in Section 2.5 can be combined to give

$$c_n = \frac{1}{T_0} \int_{T_0} v(t) e^{-j2\pi \cdot nf_0 \cdot t} dt \quad (2.11.4)$$

Once c_n is known, the amplitude A_n and phase angle ϕ_n of the corresponding cosine term in the trigonometric Fourier series can be found, which will generally be of more practical use. The equations are summarized next:

$$A_n = 2 \cdot |c_n| \quad (2.11.5)$$

$$\phi_n = \arg(c_n) \quad (2.11.6)$$

It must be remembered that the dc component represents a special case and is given by

$$A_0 = c_0 \quad (2.11.7)$$

As has already been mentioned, in many practical situations the functional form for $v(t)$ will not be known. In these situations the integral of Eq. (2.11.4) is evaluated by treating it as an area under a curve and obtaining an approximate value for the area, using an area rule such as the rectangular rule. Fast Fourier transform programs follow this approach, and this has a bearing on how the data must be entered into the program and how the results are to be interpreted, as will be shown shortly.

2.12 Approximate Formulas for the Fourier Coefficients

Consider first a periodic function $f(t)$ for which an approximate value of the area for one complete cycle must be found using the rectangular rule. This is shown in Fig. 2.12.1.

The function is sampled uniformly, the spacing between samples being T_s . The first sample is the value at $t = 0$, or $f(0)$. The second sample is $f(T_s)$, the third sample $f(2T_s)$, and so on. Denoting the number of the sample by k , where $k = 0, 1, 2, 3, \dots$, the value of the k th sample is simply $f(kT_s)$.

Applying the rectangular rule, the area under the curve for one cycle is

$$\begin{aligned} A &= \text{area}_1 + \text{area}_2 + \text{area}_3 + \dots \\ A &= f(0) \cdot T_s + f(T_s) \cdot T_s + f(2T_s) \cdot T_s + \dots \\ &= T_s \sum_{k=0}^{k=N-1} f(kT_s) \end{aligned} \quad (2.12.1)$$

where N is the total number of samples taken over one cycle. It will be noticed that the area for each rectangle in the approximation is taken to the right of the corresponding sample. The first sample is at $k = 0$, and the last sample is at $k = N - 1$, so the total number of samples is N ; but note that the N th sample belongs to the following cycle and is not included in the samples taken. In Fig. 2.12.1, $N = 8$, and hence the last sample number is $N - 1 = 7$. Applying this to Eq. (2.11.4),

$$\begin{aligned} c_n &= \frac{1}{T_0} \int_{T_0} v(t) e^{-j2\pi n f_0 t} dt \\ &\cong \frac{T_s}{T_0} \sum_{k=0}^{k=N-1} v(kT_s) e^{-j2\pi n f_0 kT_s} \\ &= \frac{1}{N} \sum_{k=0}^{k=N-1} v(kT_S) e^{-j\frac{2\pi nk}{N}} \end{aligned} \quad (2.12.2)$$

From Fig. 2.12.1, it will be seen that $T_s/T_0 = 1/N$, and this substitution has been made in Eq. (2.12.2).

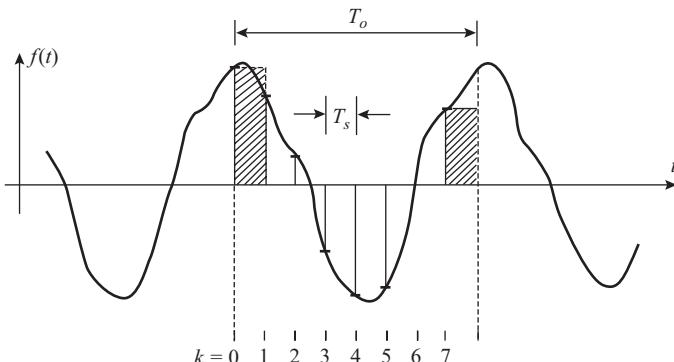


Figure 2.12.1 Rectangular rule used to find the area of a periodic function.

A question that arises naturally is how many samples of the waveform should be used. The sampling theorem, which is discussed more fully in Chapter 17, states in part that if the highest significant frequency in the spectrum is f_h then the sampling frequency f_s , should be at least $2f_h$. It follows from this that if the highest harmonic number is n_{\max} then the number of samples N should be at least $2n_{\max}$. This is illustrated in the following example.

EXAMPLE 2.12.1

Figure 2.12.2 shows a periodic waveform for which it is known that the highest harmonic is the third. Determine the trigonometric Fourier series for the wave.

SOLUTION Since $n_{\max} = 3$, making $N = 8$ should satisfy the conditions required by the sampling theorem. Since the periodic time is not specified, the sampling interval T_s may be taken as unity. The sample values, obtained from Fig. 2.12.2, are

Sample No.	0	1	2	3	4	5	6	7
Value	0	0.512	0.7	-0.088	0	0.088	-0.7	-0.512

$$\text{For } n = 0, \quad c_0 = \frac{1}{8} \sum_{k=0}^{k=7} v(k)e^0 = 0$$

$$\text{For } n = 1, \quad c_1 = \frac{1}{8} \sum_{k=0}^{k=7} v(k)e^{-j2\pi k/8} = -0.25j$$

$$\text{For } n = 2, \quad c_2 = \frac{1}{8} \sum_{k=0}^{k=7} v(k)e^{-j2\pi k/8} = -0.15j$$

$$\text{For } n = 3, \quad c_3 = \frac{1}{8} \sum_{k=0}^{k=7} v(k)e^{-j2\pi k/8} = 0.2j$$

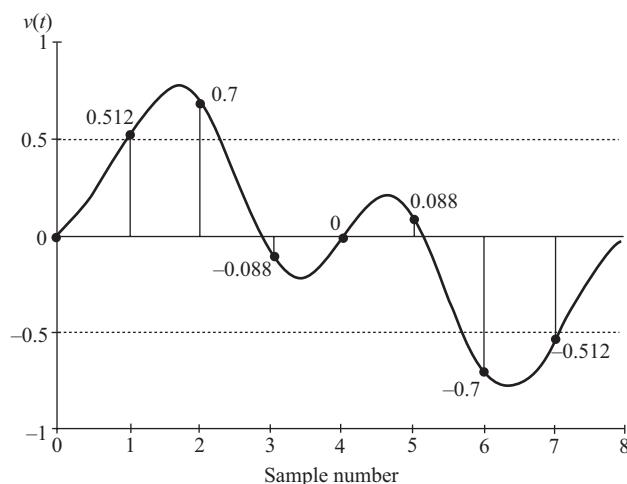


Figure 2.12.2 Periodic waveform for Example 2.12.1.

Hence, $A_0 = 0$, $A_1 = 2|c_1| = 0.5$, $A_2 = 2|c_2| = 0.3$, and $A_3 = 2|c_3| = 0.2$. The $-j$ multiplier signifies a phase lag of 90° , and the j multiplier a phase lead of 90° . Hence the series is

$$\begin{aligned} v(t) &= 0.5 \cos(2\pi ft - 90^\circ) + 0.3 \cos(2\pi 2f_0 t - 90^\circ) + 0.2 \cos(2\pi 3f_0 t + 90^\circ) \\ &= 0.5 \sin 2\pi f_0 t + 0.3 \sin 4\pi f_0 t - 0.2 \sin 6\pi f_0 t \end{aligned}$$

In this case, a relatively simple example was chosen to illustrate the method, but even so quite difficult summations are involved. In practice, a much larger number of samples is normally taken, 1024 being a typical number, and the computations become impractically large to carry out as shown. A computer routine known as the *fast Fourier transform* is widely available for such computations, and this is explained in Section 2.14.

2.13 Energy Signals and Fourier Transforms

Another type of signal encountered in communications engineering is that which lasts for a finite duration, such as a single pulse. By definition, such signals are nonperiodic but deterministic, and the power in such a signal, averaged over all time, must be zero. However, the average energy is finite.

EXAMPLE 2.13.1

Determine the average energy in a rectangular pulse of height 3 V and width 2 ms, developed across a $10\text{-}\Omega$ resistor.

SOLUTION The instantaneous power exists only during the pulse duration and is $p(t) = v(t)^2/R = 3^2/10 = 0.9$ W. Since this is constant, the average energy is $U = 0.9 \times 2 \times 10^{-3} = 1.8 \text{ mJ}$. Thus the average energy is finite. The average power, obtained by dividing the finite energy by the infinite time range, is zero.

Energy signals cannot be expanded into a Fourier series. This follows from the fact that a Fourier series consists of a summation of harmonics, each of which carries a fixed amount of average power, thus contradicting the statement of zero average power. Energy signals can, however, be represented in the frequency domain by means of a spectrum density function. The spectrum density function is related to the time waveform through what is known as a *Fourier transform*. Denoting the time waveform as $v(t)$ and the spectrum density function as $V(f)$, the relationship can be shown symbolically as

$$V(f) = F[v(t)] \quad (2.13.1)$$

The Fourier transform is an integral operation very similar to that shown in Eq. (2.11.4) for the harmonic coefficient c_n , and it is given next for completeness.

$$V(f) = \int_{-\infty}^{\infty} v(t) e^{-j2\pi ft} dt \quad (2.13.2)$$

The Fourier transforms for a wide range of pulse shapes are listed in many handbooks, and rather than solve the integral, use will be made of these published data to illustrate the physical significance of the spectrum density. Later, the use of the fast Fourier transform, introduced in Section 2.14, will be illustrated, as this provides a fast computer method of finding the spectrum density for arbitrary pulse shapes.

A rectangular pulse of height A and width τ is shown in Fig. 2.13.1(a). The spectrum density for this pulse is given by

$$V(f) = A\tau \operatorname{sinc} f\tau \quad (2.13.3)$$

This is seen to have the same functional form as Eq. (2.9.6) except that in this case the continuous variable $f\tau$ is used rather than the discrete variable k . The voltage spectrum density is plotted in Fig. 2.13.1(b), normalized to unity; that is, $A\tau = 1$. This therefore is a plot of the $\operatorname{sinc} f\tau$ function. The spectrum is double sided, requiring negative as well as positive frequencies, just like the Fourier coefficients of Section 2.11. Also, as for the exponential Fourier series, the negative half of the spectrum does not exist independently of the positive half; it is the complex conjugate of the positive half. The voltage spectrum density is a continuous curve, unlike the Fourier series spectrum, which is a line spectrum. Another fundamental difference is that the units for spectrum density are volts per hertz, so the voltage associated with some small bandwidth df centered about a frequency f_x is given by $V(f_x) df$. Note that the dimensions of $A\tau$ are volt-seconds or volts per hertz.

The voltage spectrum density is seen to consist of a main lobe and a number of sidelobes, which extend to infinity. Any practical transmission system will be unable to transmit the full theoretical range of frequencies, and the truncation in the frequency domain means that the pulse shape will be distorted after transmission. Examples of this will be met with in the chapter on digital transmission, but meantime it should be noted that the bulk of the spectrum, covered by the main lobe, is contained in a frequency range from zero to the first null at $f = 1/\tau$. As mentioned previously, negative frequencies are not to be considered separately from positive frequencies; that is, the bandwidth of the main lobe is not to be taken as $2/\tau$.

The inverse operation, that of finding a waveshape from a given spectrum density, is known as taking the *inverse Fourier transform*. Again, the mathematical operation will be stated without proof, but this will be illustrated later through the use of the inverse fast Fourier transform (ifft). The inverse Fourier transform equation is

$$v(t) = \int_{-\infty}^{\infty} V(f) e^{j2\pi ft} dt \quad (2.13.4)$$

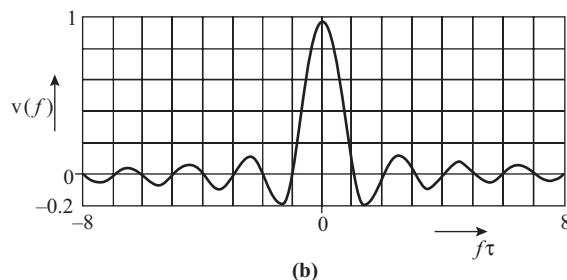
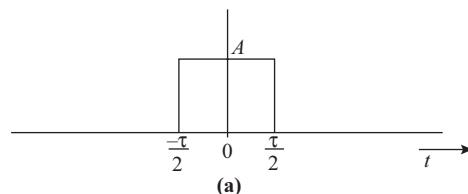


Figure 2.13.1 (a) Rectangular pulse. (b) Its spectral density.

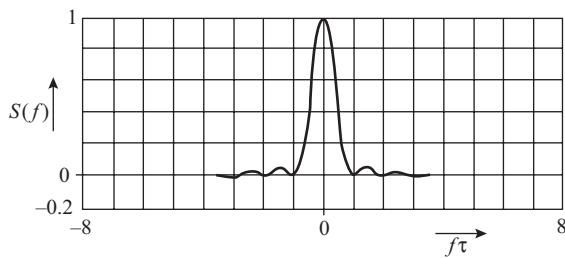


Figure 2.13.2 Energy spectral density curve for the pulse of Fig. 2.13.1.

The operation may be shown symbolically as

$$v(t) = F^{-1}[V(f)] \quad (2.13.5)$$

The Fourier transform and its inverse constitute the Fourier transform pair. The most significant property of the Fourier transform pair is *uniqueness*, which means that only one spectrum density is associated with a given waveform and, conversely, only one waveform is associated with a given spectrum density. The Fourier transform pair is often shown symbolically as

$$v(t) \leftrightarrow V(f) \quad (2.13.6)$$

It was shown that periodic waveforms have line spectra (measured in volts for a voltage waveform) rather than a voltage spectrum density function measured in volts per hertz. Although this aspect will not be pursued here, it may be stated that Fourier transforms can be generalized to cover periodic waveforms, also, through the introduction of a function known as the *delta function*. Details will be found in more advanced books on communication theory (see, for example, *Analog and Digital Communication Systems*, 3rd ed., by Martin S. Roden, Prentice Hall, 1991).

An important concept associated with energy signals is that of the energy spectrum density, which is the energy per unit bandwidth plotted as a function of frequency. Consider a voltage pulse developed across a resistor of $R \Omega$. Let $V(f)$ be the voltage spectrum density; then the energy spectrum density is $S(f) = |V(f)|^2/R$. Since the units for $V(f)$ are volts per hertz the units for $S(f)$ are (volts/hertz) $^2/\text{ohms}$. But volts $^2/\text{ohms} = \text{watts}$, and since hertz = s^{-1} , the units for $S(f)$ become watt \times second/hertz or joules/hertz. It is customary in analysis to assume a $1-\Omega$ resistor for R , and the plot of $|V(f)|^2$ then gives the energy spectrum density in joules per hertz as a function of frequency. The energy spectrum density curve for the pulse of Fig. 2.13.1 is plotted in Fig. 2.13.2. The total area under the energy spectrum density curve gives the total energy in the pulse for a $1-\Omega$ load resistor.

2.14 Fast Fourier Transform

Evaluation of the Fourier coefficients for a periodic function and of the Fourier transform for a pulse-type function require the evaluation of definite integrals. By thinking of the definite integral as an area under a curve, the rectangular rule can be used to approximate the integral by a summation, as already shown in Section 2.12.

To apply the rectangular rule, samples are taken at uniform intervals T_s in the time domain, a total of N samples being taken. Denoting the sample number by an integer k , where $0 \leq k \leq N - 1$, then the variable t in the integral is replaced by kT_s , and the limits on the integral are replaced by the summation limits 0 and

$N - 1$. In the rectangular rule approximation to the integral, the dt under the integral sign is replaced by T_s outside the summation sign.

As shown in Section 2.12, the exponential coefficients are given by

$$c_n = \frac{1}{N} \sum_{k=0}^{k=N-1} v(kT_s) e^{-j\frac{2\pi nk}{N}} \quad (2.14.1)$$

Applying a similar argument to the Fourier transform integral of Eq. (2.13.2) gives

$$V(n) = T_s \sum_{k=0}^{k=N-1} v(kT_s) e^{-j\frac{2\pi nk}{N}} \quad (2.14.2)$$

The fast Fourier transform algorithm provides a computationally efficient way of evaluating the summations. For the versions most commonly in use, the number of samples is always an integral power of 2. When N samples are taken in the time domain, the transform computation will produce N values in the frequency domain also. However, computer printouts will usually show only the positive frequency half of the spectrum, since the negative half is known to be the complex conjugate of the positive half. The printout for the frequency domain usually contains $(N/2 + 1)$ values, where the last value is known as the *fold-over value*. The reason for this will be explained shortly in connection with the inverse transform. For example, if $N = 32$ samples in the time domain, the printout will contain 17 spectrum values. With N a power of 2, $N = 2^m$, where m is an integer, and the printout will contain $2^m + 1$ components.

Figure 2.14.1(a) shows the relationships between sample points in the time domain, and Figure 2.14.1(b) the relationships in the frequency domain for a pulse waveform.

For periodic functions, the harmonics c_n are spaced by $n f_0$ in the frequency domain, where $f_0 = 1/T_0$ and T_0 is the periodic time. For aperiodic signals (for example, a pulselike function), T_0 represents the time base over which the pulse is sampled, and the spacing of the $V(n)$ values in the frequency domain are spaced by $n F_0$, where $F_0 = 1/T_0$. Although the equation form is the same, uppercase F is used to emphasize the fact that the values obtained for the aperiodic function are not harmonics, but point values on the continuous spectrum density curve.

When analyzing time functions, the sampling frequency is set at twice the highest frequency component in the spectrum of the waveform being analyzed, as required by the *Nyquist sampling theorem* (which is discussed more fully in Chapter 11). This would seem to be a catch, since the point of the analysis is to find the spectrum, and hence by definition the highest frequency is not known. However, in practice certain physical constraints, which are known, will limit the spectrum. For example, the signal will usually be filtered and the filter characteristics known. Assuming therefore that the highest significant frequency in the spectrum is known, and denoting this by f_h , the sampling frequency is given by $f_s \geq 2f_h$. The sampling interval is $T_s = 1/f_s$. The minimum number of samples would be T_0/T_s , where T_0 is the periodic time (or time base for a pulse), but the T_0/T_s has to be rounded up to the nearest integer power of 2. For example, if T_0/T_s were to equal 120, then $N = 128$ samples should be used.

In applying computer packages that provide the fft function, in order to be able to correctly interpret the results we must be aware that various versions of the summation formula exist. For example, the Mathcad formulation for the fft is of the form

$$g_n = \frac{1}{\sqrt{N}} \sum_{k=1}^{k=N} v(kT_s) e^{j\frac{2\pi nk}{N}} \quad (2.14.3)$$

To obtain c_n from the Mathcad g_n , the complex conjugate of the g_n values must be taken and the result multiplied by $1/\sqrt{N}$ so that the final multiplier is $1/N$, as required by Eq. (2.14.1). To obtain $V(n)$ from g_n ,

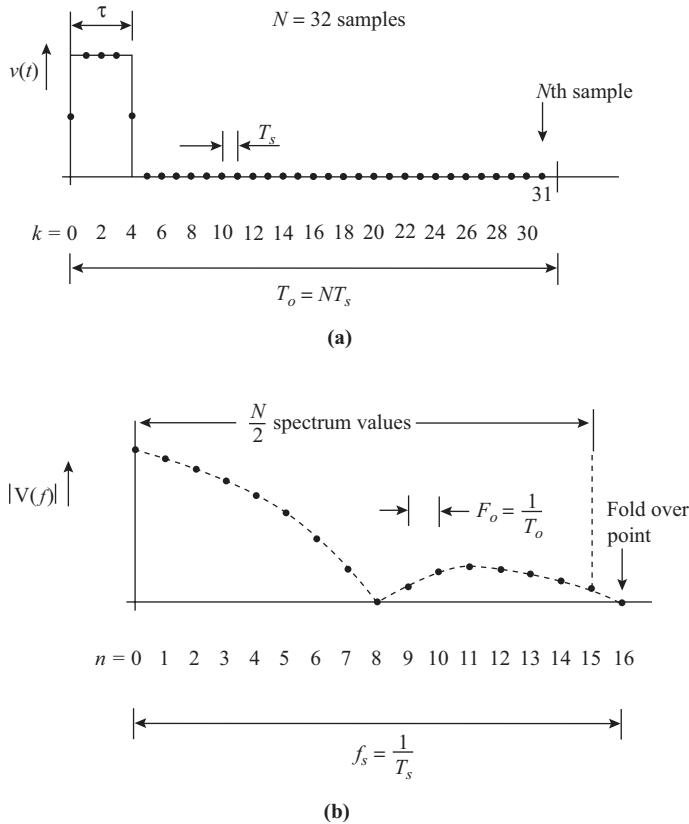


Figure 2.14.1 (a) Sampling a rectangular pulse. (b) Computed spectrum density samples.

again the complex conjugate must be used and the result multiplied by $T_s \cdot \sqrt{N}$ so that the final multiplying factor is T_s , as required by Eq. (2.14.2).

To check the results of any computer applications package, the best procedure is to apply it first to a known function and check the result. This is illustrated in the following example using the Mathcad package.

EXAMPLE 2.14.1

For the periodic waveform shown in Fig. 2.12.2, it is known that the highest harmonic is the third. Apply the Mathcad fft function to find the exponential Fourier series.

SOLUTION Since the highest harmonic is the third, the number of samples should be at least $2 \times 3 = 6$; but to meet the fft requirements, eight will be used. Set $N = 8$ and define an index k for eight samples:

$$N := 8 \quad k := 0 \dots 7$$

Denote the sample values by s . These are shown in Fig. 2.12.2 as

$$s_k :=$$

$$k$$

0	0
1	0.512
2	0.7
3	-0.088
4	0
5	0.088
6	-0.7
7	-0.512

Take the fft: $S := \text{fft}(s)$

Form the complex conjugate: $S := \bar{S}$

The required spectrum is: $c := \frac{1}{\sqrt{N}} \cdot \bar{S}$

These can be shown as: $c = \begin{bmatrix} 0 \\ 0.25i \\ 0.15i \\ -0.1i \\ 0 \end{bmatrix}$

These are the values for the positive half of the spectrum. Note that the last value, the fold-over value, is zero. The coefficients for the negative half are the complex conjugates of these.

2.15 Inverse Fast Fourier Transform

A similar approach is taken in evaluating the inverse Fourier transform, the approximate form of the Eq. (2.13.4) being

$$v(k) = F_0 \sum_{n=0}^{n=N-1} W(nF_0) e^{j \frac{2\pi n k}{N}} \quad (2.15.1)$$

where $F_0 = 1/T_0$ is the sampling interval in the frequency domain, and $W(nF_0)$ is the double-sided spectrum. Considering first the spectrum density function, the raw input data in this case are the spectrum density values $V(nF_0)$ for the positive half of the frequency range, and $W(nF_0)$ is formed from this by appending the complex conjugate values. In practice, the usual situation is that the user has only to enter the $V(nF_0)$ values and the computer routine will automatically generate $W(nF_0)$. The only restriction is that the number of sample values for $V(nF_0)$ must be $2^m + 1$, where m is a positive integer.

This is so for the Mathcad inverse fast Fourier transform (ifft), where the user inputs $2^m + 1$ samples for $V(nF_0)$ and the program automatically generates the $W(nF_0)$ vector of values. In doing so, the last value for $V(nF_0)$ is used as a *fold-over* value, which is common to $V(nF_0)$ and its complex conjugate. Also, the dc value in $V(nF_0)$ is common to both halves of the double-sided spectrum. As a result, the number of values returned for the time function is $2 \times (2^m + 1) - 2 = 2^m + 1$.

For a harmonic series, the c_n values are entered instead of $V(nF_0)$ and the multiplying factor F_0 in Eq. (2.15.1) is omitted, so the result is consistent with the summation given in Eq. (2.11.3). The result of the inverse transform in either case is $v(k)$, the time function evaluated at instants kT_s .

The Mathcad formulation for the ifft is of the form

$$d_k = \frac{1}{\sqrt{N}} \sum_{n=1}^{n=N} W(nF_0) e^{-j\frac{2\pi nk}{N}} \quad (2.15.2)$$

To convert the Mathcad d_k values to the desired $v(k)$ values, the input data must be the complex conjugate of the actual spectrum values. If the input data are spectrum density values $V(nF_0)$, the required input is $\overline{V(nF_0)}$, where the overbar signifies complex conjugate, and the resultant output has to be multiplied by $F_0 \cdot \sqrt{N}$. If the input data are harmonic values c_n , they must be converted to $\overline{c_n}$ and the resultant output multiplied by \sqrt{N} where N is the total number of samples in the time domain. As explained previously, the spectrum input has $2^m + 1$ values and the number of resulting time values is $2^m + 1$. The use of the Mathcad ifft is shown in the following example.

EXAMPLE 2.15.1

Use the computed values for the spectrum obtained in Example 2.14.1 and the ifft to obtain the sample values in the time domain.

MATHCAD SOLUTION The spectrum values, from Example 2.14.1, are

$$c : = \begin{bmatrix} 0 \\ -0.25i \\ -0.15i \\ 0.1i \\ 0 \end{bmatrix}$$

Form the complex conjugate: $c : = \bar{c}$

Take the ifft of this: $v : = \text{ifft}(c)$

The value of N is: $N : = \text{last}(v) + 1$

As a check, the value of N is: $N = 8$

The required sample values are: $v : = \sqrt{N} \cdot v$

These can be shown as:

$$v = \begin{bmatrix} 0 \\ 0.512 \\ 0.7 \\ -0.088 \\ 0 \\ 0.088 \\ -0.7 \\ -0.512 \end{bmatrix}$$

2.16 Filtering of Signals

One reason for wishing to know the spectrum of a signal is that filtering of the signal is determined by a multiplication of the spectrum by the frequency response of the filter. In Chapter 1 the transfer function of a filter, designated as $H(f)$, gives the ratio of the output signal to input signal at a given sinusoidal frequency f , as shown by Eq. (1.3.1). The transfer function $H(f)$ will be known or can be measured as a function of frequency. The input frequency spectrum $V_i(f)$ can be determined by Fourier methods, and the spectrum of the output signal is then

$$V_o(f) = H(f)V_i(f) \quad (2.16.1)$$

The inverse Fourier transform can be used to find the waveform from a given output spectrum. The time function (the waveform) is given by

$$v_o(t) = F^{-1}[V_o(f)] \quad (2.16.2)$$

This is illustrated in the following example using Mathcad.

EXAMPLE 2.16.1

The spectrum obtained in Example 2.14.1 is passed through a filter that has the $H(f)$ transfer characteristic shown. Determine the output waveform.

MATCAD SOLUTION The spectrum values, from Example 2.14.1, are

$$c := \begin{bmatrix} 0 \\ 0.25i \\ 0.15i \\ -0.1i \\ 0 \end{bmatrix}$$

In order to set up the filter function, define n

$$n := 0 \dots 4$$

The filter transfer function $H(f)$ in terms of n is

$$H_n := \frac{1}{\sqrt{1 + n^2}}$$

The filtered spectrum is

$$c_n := c_n \cdot H_n$$

From the complex conjugate:

$$c := \bar{c}$$

Take the ifft of this:

$$v := \text{ifft}(c)$$

The value of N is:

$$N := \text{last}(v) + 1$$

As a check, the value of N is: $N = 8$

The required sample values are: $v := \sqrt{N} \cdot v$

These can be shown as:

$$v = \begin{bmatrix} 0 \\ 0.339 \\ 0.417 \\ 0.071 \\ 0 \\ -0.071 \\ -0.417 \\ -0.339 \end{bmatrix}$$

2.17 Power Signals

Periodic voltage and current waves carry finite average power, and as such they represent signals that fall into a class known as *power signals*. For example, a sinusoidal current flowing through a resistor R develops an average power given by $P = I^2 \cdot R$, where $I = I_{\max}/\sqrt{2}$ is the root-mean-square current (rms).

Another example is the square wave, for which the rms current is equal to the maximum current of the square wave, denoted here by I_{sqwave} to avoid confusion with the peak value of the fundamental component of the spectrum. Hence the average power is $P = I_{\text{sqwave}}^2 \cdot R$.

In terms of the spectrum components, the peak of the fundamental component of the square wave is, from the results of Section 2.7, $I_{\max} = 4I_{\text{sqwave}}/\pi$, and the power is the sum of all spectrum components, which again from the results of Section 2.7 is

$$\begin{aligned} P &= \left(\frac{4I_{\text{sqwave}}}{\pi\sqrt{2}} \right)^2 \cdot R \left(1^2 + \frac{1}{3^2} + \frac{1}{5^2} + \dots \right) \\ &= 8 \left(\frac{I_{\text{sqwave}}}{\pi} \right)^2 \cdot R \left(\frac{\pi^2}{8} \right) \\ &= I_{\text{sqwave}}^2 \cdot R \end{aligned}$$

This demonstrates that the total average power is spread over all the spectral components, extending to infinity. This result holds in general for periodic waveforms.

Periodic waveforms are also known as *deterministic* waveforms because their values are known at all times. Another type of power waveform encountered in communications engineering is the *random* waveform, for which the values cannot be predicted. Noise waveforms are examples of random waveforms, as are all information signals such as audio and video signals. These signals carry finite power, but unlike periodic waves they do not have a harmonic, or *line spectra*. Fourier methods applied to these random power signals result in a *power spectral density* function. This is a curve that shows the energy distribution as a continuous function of frequency. Such a curve is sketched in Fig. 2.17.1.

The units for power spectral density are watts per hertz, which are equivalent to joules. The area under the curve has units of joules \times hertz, which is dimensionally equivalent to watts. The total area under the

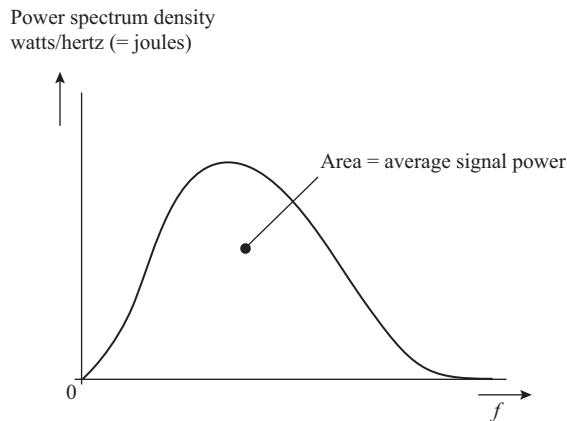


Figure 2.17.1 Power spectral density curve (positive half-frequencies only shown).

curve gives the average signal power. Note carefully how the power spectrum density curve under discussion here differs from the energy spectrum density curve for a pulse described in the previous section.

For binary digital waveforms (the topic of the next chapter), there is a connection between the energy spectrum density of the basic pulse making up the waveform and the power spectrum density of the waveform. Under certain conditions the power spectrum density $G(f)$ of the waveform is related to the energy spectrum density of the basic pulse as follows:

$$\begin{aligned} G(f) &= \frac{1}{\tau} |V(f)|^2 \\ &= \frac{1}{\tau} S(f) \end{aligned} \quad (2.17.1)$$

τ is the pulse period, and $S(f)$ is the energy spectrum density introduced in Section 2.13. It will be noted that $S(f)$ has units of joules per hertz, and $G(f)$ has units (1/s)(joules/hertz), which is equivalent to watts per hertz. The conditions required for Eq. (2.17.1) to apply are as follows:

1. The individual pulses must be independent of one another.
2. The binary 1's and 0's are generated with equal probability.
3. The waveform has zero mean.
4. The waveform is assumed to be infinitely long.

Examples of binary waveforms will be met with in Chapter 3, but to illustrate, the power spectrum density for a binary random waveform consisting of rectangular pulses of amplitudes $\pm A$ is shown Fig. 2.17.2. Combining Eqs. (2.13.3) and (2.17.1) gives $G(f)$ as

$$G(f) = \tau(A \operatorname{sinc} f\tau)^2 \quad (2.17.2)$$

Again, it must be emphasized that the power spectrum density curve does not give the power at any one frequency. What can be said is that the power in some small bandwidth Δf about some frequency f_x is given by $P(f_x) \cong G(f_x)\Delta f$. The concept of power spectral density is also fundamental to the study of noise waveforms, and this aspect is examined in Chapter 4.

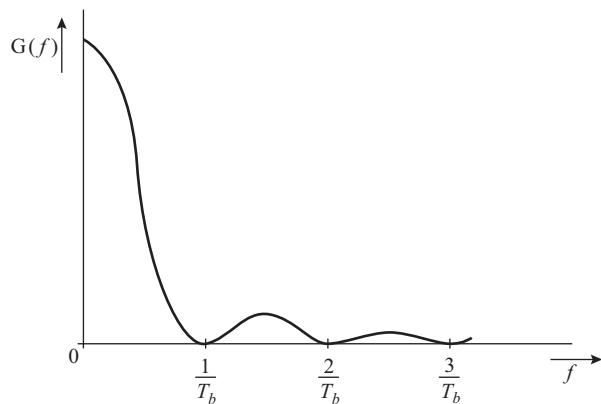


Figure 2.17.2 Power spectral density for a binary random waveform of rectangular pulses of amplitudes $\pm A$.

2.18 Bandwidth Requirements for Analog Information Signals

As Fourier methods show, signals can be described in terms of a frequency spectrum. From the transmission point of view, it is desirable to limit the bandwidth required for any signal, because this allows more channels to be accommodated in a given frequency band. Thus the bandwidth allocated for signals is often a compromise between minimizing the bandwidth and acceptable levels of distortion.

For speech signals, listening tests have shown that the energy content resides mainly in the low audio-frequency range; typically, about 80% of the energy lies in frequencies below 1 kHz. The intelligibility, or clarity, of the signal requires the higher frequencies, typically in the range from 1.5 to 2.5 kHz. For telephone service, a frequency range of 300 to 3400 Hz is the generally accepted standard, even though the spectrum for natural speech has a much greater frequency spread than this. Not all telephone administrations adhere to this standard, but it is widely used in determining system performance. Thus the audio bandwidth for speech signals is $3400 - 300 = 3100$ Hz. As will be seen later, when filtering requirements are taken into account, 4 kHz is usually allocated as the channel bandwidth for speech signals.

Signals associated with music, both instrumental and voice, require a much larger bandwidth than that for speech. Listening tests have shown that a frequency range of about 15 to 20,000 Hz is required for reasonable high-fidelity sound. For example, the specification sheets for certain pieces of "compatible" high-fidelity equipment specify the frequency ranges as CD player, 2 to 20,000 Hz; FM tuner section, 30 to 15,000 Hz; stereo tape deck, normal tape, 30 to 14,000 Hz, metal and chrome tapes, 30 to 15,000 Hz; and stereo turntable, 10 to 30,000 Hz.

Video signals, such as those produced by standard television systems, require a bandwidth of about 4 MHz, while a facsimile signal requires a bandwidth of only about 1000 Hz. These bandwidth requirements are discussed in detail in later chapters, but for the present they are given here for comparison purposes. This wide difference in bandwidth requirements reflects the difference in time taken for a transmission system to scan the picture information. With television the picture is scanned in about $1/30$ s, while a facsimile scanner may require about 10 min for a page of print. This illustrates the general connection between speed and bandwidth. The greater the rate at which information is generated, the greater the bandwidth needed, and of course vice versa. Thus the widespread use of facsimile transmission has come about (apart from its general appeal to the public) because it makes use of existing telephone networks that are often limited in bandwidth.

These bandwidth requirements are shown in Fig. 2.18.1 on a logarithmic frequency scale. The video signal extends down to zero frequency (which cannot be shown on the logarithmic scale; why?). The point

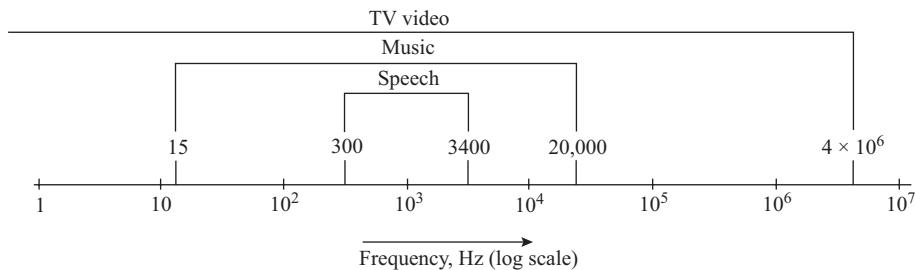


Figure 2.18.1 Bandwidth requirements for some baseband signals.

has been made elsewhere (source unknown) that the video bandwidth is about 1000 times the speech bandwidth, which gives some credence to the saying that “one picture is worth a thousand words”!

Digital signals also follow the general principle that the bandwidth requirements increase as the transmission rate increases. Digital signals are covered in the next chapter.

PROBLEMS

- 2.1. A sine wave is described by $5 \sin(300t + 27^\circ)$, where t is time in seconds. Determine the waveform
(a) amplitude, (b) rms value, (c) frequency, (d) periodic time, and (e) time lag or lead.
- 2.2. Express the waveform of Problem 2.1 as a cosine function.
- 2.3. By sketching the waveform of Problem 2.1, determine if it is an even or odd function, or neither.
- 2.4. A periodic waveform is described by $f(x) = |\sin x|$ for $-\pi < x < \pi$. By sketching this function, determine if it is an even or odd function, or neither.
- 2.5. Repeat Problem 2.4 for $f(x) = x$ for $-\pi < x < \pi$.
- 2.6. The trigonometric series for a periodic waveform consists of three sine terms consisting of a fundamental, a second harmonic, and a third harmonic. The fundamental has an amplitude of 1 V and provides the reference phase. The second harmonic has an amplitude of 0.3 V and a phase lag of 27° . The third harmonic has an amplitude of 0.5 V and a phase lead of 30° . Draw accurately to scale the resultant waveform.
- 2.7. The spectrum components of a waveform are fundamental, 1 V; second harmonic, 0.7 V; third harmonic, -0.35 V; all are sine waves with zero phase angle. The fourth harmonic is a 0.1-V cosine wave, also with zero phase angle. Construct accurately to scale one cycle of the resultant waveform. All voltages given are peak values.
- 2.8. A waveform consists of a 3-V dc component, a fundamental $2 \sin \omega t$, and a second harmonic $1.5 \sin 2\omega t$. Draw accurately to scale one cycle of the waveform.
- 2.9. A spectrum component of a waveform is given by $3 \sin(\omega t - 20^\circ)$. Express this as the sum of sine and cosine terms.
- 2.10. A spectrum component of a waveform is given by $5 \cos(\omega t + 70^\circ)$. Express this as the sum of sine and cosine terms.
- 2.11. The n th harmonic of a waveform can be written as $7 \cos(n\omega_0 t + 35^\circ)$. Determine the corresponding a_n and b_n amplitudes.

- 2.12.** The Fourier coefficients for the n th harmonic of a waveform are $a_n = 35$ V, $b_n = 45$ V. Express the harmonic in cosine form.
- 2.13.** For a waveform series $A_n = 1/n^2$, $a_0 = 0$ and $\phi_n = 0$. Given that all harmonics from $n = 1$ to 10 are present in the waveform, plot this using Eq. (2.4.6).
- 2.14.** The Fourier series for a square wave of 1-V amplitude is

$$v(t) = \frac{4}{\pi} \left(\sin \omega_0 t + \frac{\sin 3\omega_0 t}{3} + \frac{\sin 5\omega_0 t}{5} + \dots \right)$$

By comparing this with the series given in Eq. (2.7.1), sketch the square wave relative to the waveform shown in Fig. 2.7.1(a).

- 2.15.** The Fourier series for a half-wave rectified sinusoidal current wave of 1-A amplitude is

$$i(t) = \frac{1}{\pi} + \frac{1}{2} \sin \omega_0 t - \frac{2}{\pi} \left(\frac{\cos 2\omega_0 t}{1 \times 3} + \frac{\cos 4\omega_0 t}{3 \times 5} + \frac{\cos 6\omega_0 t}{5 \times 7} + \dots \right)$$

Draw accurately to scale the spectrum up to the eighth harmonic. What is the value of the dc component?

- 2.16.** Write out the next three terms for the wave in Problem 2.15. Using Mathcad or any other suitable computer program (or writing your own), construct accurately to scale two cycles of the waveform, using harmonics up to and including the eleventh.
- 2.17.** Calculate the rms value of the thirteenth harmonic of a square wave for which the peak-to-peak voltage is 10 V.
- 2.18.** Calculate the rms value of the tenth harmonic of a sawtooth wave for which the peak-to-peak voltage is 10 V.
- 2.19.** Interpret Eqs. (2.5.1), (2.5.2), and (2.5.3) for the waveform shown in Fig. 2.7.2(a) to determine which coefficients exist. Do not attempt to solve the integrals.
- 2.20.** Repeat Problem 2.19, but with the square wave set up as an odd function.
- 2.21.** Repeat Problem 2.19, for a 3-V peak, half-wave rectified sine wave. (Hints: Set up the waveform as an even function; use the expressions given in Problem 2.15 to find the dc component.)
- 2.22.** Plot the envelope of the amplitude function given by Eq. (2.9.3) for a periodic time of 1 ms, a pulse width of 0.1 ms, $V = 1V$ and $n = 1$ to 30 inclusive.
- 2.23.** Write down the defining equations for the *sine function* and for the *sampling function*. Evaluate (a) sine 0.3, (b) sine 0, (c) sine 1, (d) Sa(0.3), (e) Sa(0), (f) Sa(1).
- 2.24.** Given that sine $\lambda = 0.2$, determine the value of (a) λ and (b) Sa(λ).
- 2.25.** Plot the function sine x for x in steps of 0.1 over the range $-2.2 \leq x \leq 2.2$.
- 2.26.** One cycle of a periodic waveform is described by

$$f(x) = \cos x, \quad \text{for } 0 < x < \pi, \quad \text{and} \quad -\cos x, \quad \text{for } -\pi < x < 0$$

By applying rule 1 of Section 2.10, determine if the spectrum contains a dc component.

- 2.27.** For the waveform of Problem 2.26, apply rules 2 and 3 of Section 2.10 to determine if the trigonometric spectrum consists of sine or cosine terms.
- 2.28.** Apply rules 2 and 3 of Section 2.10 to the waveform of Fig. 2.8.1 to determine if it contains only even or odd harmonics, or both.

- 2.29. Apply rule 4 of Section 2.10 to determine the extent of the harmonic content of the waveforms described by Eq. (2.8.1). Assume harmonics less than 1% of V can be ignored.
- 2.30. Apply rules 1 through 3 of Section 2.10 to the waveform of Problem 2.4 and draw the necessary conclusions.
- 2.31. Apply rules 1 through 3 of Section 2.10 to the waveform of Problem 2.5 and draw the necessary conclusions.
- 2.32. Determine the exponential Fourier coefficients for the sine wave of Problem 2.1.
- 2.33. Determine the exponential Fourier coefficients for the waveform of Problem 2.6.
- 2.34. Write out in exponential form the equations for the waveforms in Problems 2.8, 2.9, and 2.10.
- 2.35. Write out in exponential form the series given by Eqs. (2.7.1), (2.7.2), (2.8.1), and (2.9.1).
- 2.36. The amplitude of a harmonic term in an exponential Fourier series is $c_n = 3$ V. Determine (a) the amplitude and (b) the rms value of the corresponding sinusoidal function.
- 2.37. The amplitude of a term in an exponential Fourier series is $c_n = 3 + j4$ V. Write out the equation for the corresponding trigonometric function.
- 2.38. The first two terms in an exponential Fourier series are $c_0 = 1$ V and $c_1 = 2 - j5$ V. Write out the corresponding trigonometric terms.
- 2.39. If the fundamental frequency in the waveform of Problem 2.6 is 1000 Hz, at what rate should it be sampled in order to compute the spectrum from the samples?
- 2.40. Assuming that harmonics less than 10% of the ac peak amplitude V can be ignored, determine the number of samples to be taken for the square wave shown in Fig. 2.7.2(a) in order that the waveform may be analyzed through an fft routine. The periodic time of the waveform is 3 ms.
- 2.41. Repeat Problem 2.40 for the sawtooth waveform of Fig. 2.8.1(a).
- 2.42. The highest-frequency component in a speech signal spectrum is 4 kHz. At what rate should this signal be sampled in order to compute the spectrum from the samples?
- 2.43. By using an fft computer routine, determine the spectrum for the waveform of Fig. 2.7.2(a). Compare the results with those obtained in Problem 2.19.
- 2.44. Repeat Problem 2.43 with the square wave set up as an odd function.
- 2.45. By using an fft computer routine, determine the spectrum for a 60-Hz, 3-V peak, half-wave rectified sine wave. Compare the results with those obtained in Problem 2.21.
- 2.46. A cosine wave has a periodic time of 1 s and a maximum value of 1 V. When this is passed through an amplifier, horizontal peak clipping is observed. Using the center of a positive peak as the time zero reference, the clipping extends to 0.1 s and starts again at 0.45 s, repeating in this fashion for every cycle. Sketch one cycle of the clipped waveform, and use a computer fft routine to find its spectrum.
- 2.47. Use Eqs. (2.5.1), (2.5.2), and (2.5.3) to find the theoretical spectrum for the waveform in Problem 2.46. Compare the results in volts with those obtained in Problem 2.46 for the first 6 terms.
- 2.48. Explain what is meant by an *energy signal*. The voltage spectral density curve for a pulse can be approximated by a linear rise from 0 to 5 V/Hz over the frequency range from 0 to 5000 Hz, followed immediately by an exponential fall-off expressed as $5 \exp(1 - f/5000)$ V/Hz for $f > 5000$ Hz. Sketch the voltage spectral density curve and the energy spectral density curve, assuming the voltage is developed across a load of 1Ω .
- 2.49. Determine the energy in the pulse described in Problem 2.48.
- 2.50. Using Eq. (2.13.3), plot the voltage spectral density for $A = 1$ V, $\tau = 1$ ms over the frequency range $-2.2/\tau \leq f \leq 2.2/\tau$ in steps of $0.1/\tau$. State clearly the units on the graph axes.

- 2.51.** A rectangular pulse has an amplitude of 5 V and a width of 3 ms. Determine the voltage spectral density at frequencies of (a) 30 Hz, (b) 100 Hz, and (c) 3000 Hz, stating clearly the units used.
- 2.52.** Determine the spectrum for a 3-V, 0.5-s rectangular pulse such as shown in Fig. 2.13.1(a), using a computer fft routine. Compare the results with the theoretical spectrum given by Eq. (2.13.3).
- 2.53.** Explain what is meant by a *power signal*. Give one example each of a deterministic and a nondeterministic power signal.
- 2.54.** The power spectral density curve for a nondeterministic signal can be approximated by a linear rise from 0 to 3 J over the frequency range from 0 to 300 Hz, remaining constant at 3 J up to 3000 Hz, followed by a linear drop-off reaching 0 J again at 8000 Hz. Sketch the power spectral density curve and determine the average signal power.
- 2.55.** A waveform consists of an infinitely long, random sequence of pulses, the pulses being independent of one another and the waveform having zero dc level. The basic pulse has a voltage spectrum density the magnitude of which is rectangular in shape, being 2 mV/Hz from 100 to 3000 Hz and zero outside these limits. Given that the pulse width in the time domain is 6 ms, determine the average power in the waveform.
- 2.56.** Can the pulse shape in the time domain be determined from the information given in Problem 2.55?
- 2.57.** Using Eq. (2.17.2), plot the power spectrum density curve for a binary waveform that meets the conditions stated in the text, over the frequency range 0.1 Hz to 10 kHz. The basic rectangular pulse has an amplitude of 5 volts and a duration of 3 milliseconds.
- 2.58.** Briefly describe the bandwidth requirements for speech and video signals. How many speech channels, approximately, could be fitted into the bandwidth required for a TV video signal?
- 2.59.** Plot the complex sine wave $v(t) = \sum_{n=1}^m a_n \cos(2\pi n f_o t)$ using MATLAB, for $m = 5$ and $m = 10$.
- Let $a_n = \frac{1}{n^2}$ and $f_o = 100\text{Hz}$.
- 2.60.** The MATLAB function *fft(.)* can be used to obtain the Fourier transform of a discrete sequence. Apply *fft(.)* on the discrete sequence $v(n)$ obtained from Problem 2.1. (Hint: Let $t=[0:0.01:5]$)
- 2.61.** Consider a sinusoidal signal, $e(t) = 10\sin(2000\pi t)$. Plot the entire waveform and one full wave of $e(t)$.
- 2.62.** Assume that an arbitrary waveform sequence is available as one row vector, $v(n)$. Obtain and plot 10 *replicas* of $v(n)$. (Hint: Use the column operator $(:)$ in MATLAB. $V_x = V * \text{ones}(1, 10); Vx = Vx(:); \text{stem}(Vx);$)
- 2.63.** Generate one wave of the arbitrary waveform shown in Figure P2.63, using MATLAB.

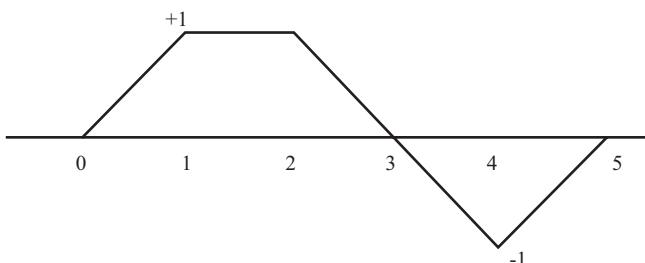


Figure P2.63

- 2.64. Replicate the waveform given above 5 times and plot it.
- 2.65. Using MATLAB, clip the lower half cycle of a sine wave and plot it.
- 2.66. One cycle of a periodic waveform is described by $v(x) = \sin(x)$, for $0 < x < \pi$ and $= -\sin(x)$, for $\pi < x < 2\pi$. Plot $v(x)$ using MATLAB. Check whether the waveform has a dc value.
- 2.67. Plot, using MATLAB, a Gaussian pulse train consisting of 10 pulses. (Hint: Use *gauspuls(.)* function along with the solution to Problem 2.4.)
- 2.68. Obtain the dc value of the waveform depicted in Problem 2.7. Verify the result using MATLAB. (Hint: Use *mean(.)* function.)



Digital Line Waveforms

3.1 Introduction

Communications systems can be broadly divided into *analog* and *digital* systems. In an analog system, the electrical waveform that carries the information is a replica of the information signal. A good example is the current or voltage waveform generated by a microphone. The electrical waveform follows the variations in sound pressure impinging on the microphone and hence can be said to be an analog of the sound pressure wave. The spectra of analog-type waveforms are the subject of Chapter 2.

In digital systems, by contrast, the electrical waveforms are *coded representations* of the original information. Pressing the letter A on a computer keyboard will generate a signal, usually a series of binary pulses, that is a code for the letter A. Alternatively, the code could be a single discrete voltage level, but binary pulses have the advantage of being more easily distinguishable in the presence of noise and interference. Thus, for a string of letters such as a teletext transmission, a data sequence is generated that is a coded representation of the information, rather than an analog of the information signal. If the original information is an analog signal, such as the microphone output mentioned previously, this must be converted to a series of discrete values that can then be transmitted digitally. The process of converting the original information into a data sequence is referred to as *source coding*.

Whatever the original source of information, text, speech, or the like, the most commonly used coding scheme is *binary digital* in which the message appears as a binary sequence such as ...00110010.... For transmission purposes, this has to be converted to a continuous electrical waveform; the conversion process is referred to as *modulation*. These waveforms are *analog* representations of the binary data, and hence the frequency analysis methods of the previous chapter are equally applicable here. However, many of the properties specifically apply to the digital aspects of the transmission, and these will be emphasized in this chapter.

3.2 Symbols, Bunits, Bits, and Bauds

The word *symbol* refers generally to a mark or token that represents something else. The letters of the alphabet are symbols representing spoken sounds, and numerals are symbols representing quantity. A binary alphabet

has only two symbols, **0** and **1**, known as *binit*s (derived from binary digits). A closely related term is the *bit*, which is a unit of information (derived from binary *unit*). The information, in bits, conveyed by the transmission of a symbol is formally defined as

$$\begin{aligned} I &= \log_2 \frac{1}{P_{\text{sym}}} \\ &= -\log_2 P_{\text{sym}} \end{aligned} \quad (3.2.1)$$

where P_{sym} is the probability of selecting the symbol from the alphabet of symbols making up the source of information. To put this on a more practical footing, suppose that a keyboard is capable of generating 128 characters and a key is selected at random. Since each symbol has a probability of $\frac{1}{128}$ of being selected, the information contained in the symbol is

$$\begin{aligned} I &= -\log_2 \frac{1}{128} \\ &= 7 \text{ bits per character} \end{aligned} \quad (3.2.2)$$

It will be seen that information, as defined here, has nothing to do with the semantic content of the message. Suppose now that the characters are coded in binary so each character requires 7 binit. All possible binit combinations will be used, since $2^7 = 128$. For a random character selection, therefore, the probability of a binary **1** occurring is equal to that of a **0** occurring, which in turn is equal to 1/2. Thus the information content of an equiprobable binit is

$$I = -\log_2 \frac{1}{2} = 1 \text{ bit per binit} \quad (3.2.3)$$

This could also have been deduced from the fact that one character conveys 7 bits of information and is encoded into 7 binit, and thus the information rate is 1 bit per binit. For this reason, the term *bit* is more commonly used as the name for the binary symbol and, unless otherwise stated, this practice will be followed here.

As shown, a source of M characters can be represented by a binary output of m bits per character where $M = 2^m$. The converse is also true; the output from a binary source can be combined in groups of bits, each group being identified by a separate symbol. Thus four symbols would be required to represent bits grouped in two's, one for each of **00**, **01**, **10**, and **11**. This is referred to as *quaternary encoding* since four separate symbols are required. In general, M symbols may be used, and the coding is referred to as M -ary encoding, where as before

$$M = 2^m \quad (3.2.4)$$

Once the symbols are modulated into a waveform, each symbol occupies a given time termed the *symbol period*, denoted here by T_{sym} . The transmission rate is measured in *bauds* (named after Emile Baudot, a French telegraph engineer); 1 baud is defined as one symbol per second. In terms of the symbol period, the symbol rate is

$$R_{\text{sym}} = \frac{1}{T_{\text{sym}}} \quad (3.2.5)$$

In the special but common case of binary transmission, it is more usual to refer to the bit period T_b and the bit rate in bits per second (bps) or multiples of this. The bit rate is given by

$$R_b = \frac{1}{T_b} \quad (3.2.6)$$

In the process of converting m bits into an M symbol, the relationship between symbol duration and bit duration is maintained as

$$T_{\text{sym}} = mT_b \quad (3.2.7)$$

Thus it follows that the transmission rates are related by

$$R_{\text{sym}} = \frac{R_b}{m} \quad (3.2.8)$$

Because rectangular pulses are commonly used as the analog representation of the symbols, it is important to realize that the pulse duration (or width) is not necessarily the same as the symbol period. If the pulse width is τ , then what is necessary is that $\tau \leq T_{\text{sym}}$ and that the pulses be equispaced on the time axis by amount T_{sym} .

3.3 Functional Notation for Pulses

The shape of the pulse representing a symbol is a function of time and may be written very generally as $p(t)$. This is a dimensionless function, and, for example, if a voltage pulse of amplitude A is used, it would be written as $Ap(t)$. The rectangular pulse is encountered so frequently in communications theory and practice that it is given its own symbol:

$$p(t) \equiv \text{rect}\left(\frac{t}{\tau}\right) \quad (3.3.1)$$

This represents a rectangular pulse of duration τ centered at $t = 0$ and normalized amplitude as shown in Fig. 3.3.1(a) and is a shorthand way of writing

$$\begin{aligned} p(t) &= 0, & \text{for } t < -\frac{\tau}{2} \\ &= 1, & \text{for } -\frac{\tau}{2} \leq t \leq \frac{\tau}{2} \\ &= 0, & \text{for } \frac{\tau}{2} < t \end{aligned} \quad (3.3.2)$$

If the pulse is displaced along the time axis by amount T , as shown in Fig. 3.3.1(b), it is written as

$$p(t - T) \equiv \text{rect}\left(\frac{t - T}{\tau}\right) \quad (3.3.3)$$

It will be seen that, in this case, $t = T$ gives the value of $p(0) = 1$. Two particular types of rectangular pulses are met with in practice. In what is referred to as a *non-return-to-zero* (NRZ) pulse, the pulse width is equal to the symbol period. In a *return-to-zero* (RZ) pulse, the pulse width is less than the symbol period and is

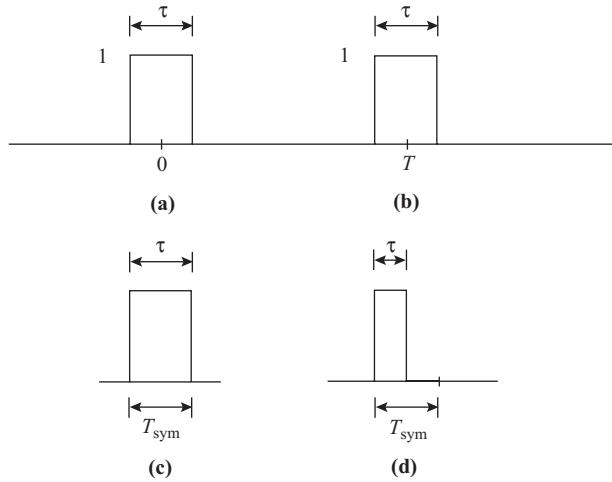


Figure 3.3.1 (a) Rectangular pulses centered at (a) $t = 0$ and (b) $t = T$. (c) NRZ pulse, and (d) RZ pulse.

often made equal to one-half the symbol period. The NRZ pulse is shown in Fig. 3.3.1(c) and the RZ pulse in Fig. 3.3.1(d). Waveforms made up of these basic rectangular pulses are discussed in the following section.

3.4 Line Codes and Waveforms

The *line code* refers to the way in which symbols are encoded by means of pulses. The simplest way of encoding would be to have each symbol represented by a NRZ rectangular pulse. This is referred to as *level* encoding because the symbol is encoded into the pulse level. The *line waveform* then consists of a series of rectangular pulses of duration T_{sym} , there being M levels for a code with M symbols. Such a line waveform using rectangular pulses can be expressed in mathematical notation as

$$v(t) = \sum_{k=-\infty}^{\infty} a_k \text{rect}\left(\frac{t - kT_{\text{sym}}}{T_{\text{sym}}}\right) \quad (3.4.1)$$

The amplitude a_k indicates the amplitude of the k th symbol, which is centered at time kT_{sym} . The fact that k ranges over $\pm\infty$ should not be of concern; this simply indicates an indefinitely long sequence of symbols. The same technique is readily accepted when describing a sinewave by $\sin \omega t$, where it is implicitly understood that t ranges over $\pm\infty$. Figure 3.4.1(b) shows a *unipolar* binary waveform of amplitude $2A$ for the binary sequence ...101101001.... The symbol period is equal to the bit period in this case, as shown, and $a_k = 0$ or $2A$. The reason for using $2A$ for the pulse amplitude is that it simplifies comparison with binary *polar* waveforms, which swing between levels $+A$ and $-A$, the difference between levels being $2A$ in each case.

For the rectangular pulse, $p(0) = 1$, and it follows that if the waveform given by Eq. (3.4.1) is *sampled* at time $t = KT_{\text{sym}}$ the sample value will be

$$v(KT_{\text{sym}}) = a_K \quad (3.4.2)$$

This is because at $t = KT_{\text{sym}}$ and $k = K$ the pulse becomes $p(KT_{\text{sym}} - KT_{\text{sym}}) = p(0) = 1$, and at any other value of k not equal to K the pulse is zero, by definition. Recall that the definition of the rectangular pulse

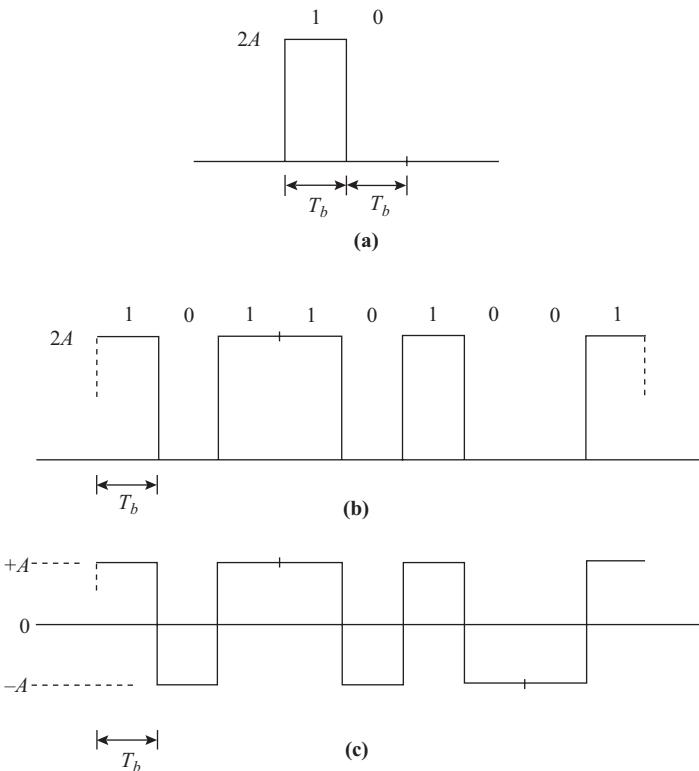


Figure 3.4.1 Unipolar NRZ-L code for (a) binary **1 0**; (b) part of a random binary stream; (c) the ac waveform.

requires the pulse width to be equal to or less than the symbol period. Thus any one pulse does not overlap with any other. Sampling is a key factor in digital communications. Unfortunately, in practice, it is virtually impossible to maintain the rectangular pulse shapes, and the resulting distortion during transmission, plus the unavoidable noise that is added in, considerably complicates the sampling and detection functions. Nevertheless, the rectangular pulses are basic to digital communications, and in this chapter some of the more important properties of their frequency spectra will be examined. The study of the effects of pulse distortion and noise is deferred until Chapter 12.

Unipolar NRZ-L Line Code

A unipolar NRZ-L waveform for a binary sequence is shown in Fig. 3.4.1(b). As mentioned previously, unipolar means that pulses of one polarity only are used. These may be positive or negative, but not both. Although, in principle, for binary waveforms the pulses could change between any two fixed levels, in practice, one of the levels is set equal to zero. In Fig. 3.4.1, the binary **1** is represented by a rectangular pulse of amplitude $2A$ and duration T_b and the binary **0** by a zero level of duration T_b . The main attraction of the unipolar code is its simplicity, but it has a number of drawbacks that make it unsuitable for general use on the normal telecommunications networks.

For an infinitely long random binary stream, there will be as many **1's** as **0's** on average, assuming these are equiprobable binites, and it follows therefore that the average or dc level for the waveform is A .

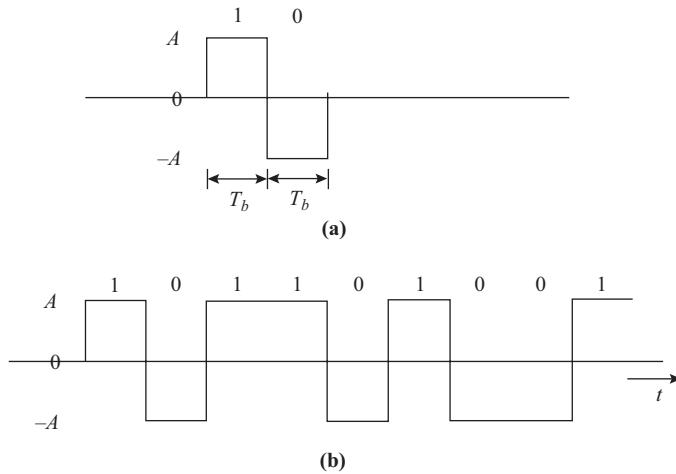


Figure 3.4.2 Polar NRZ-L code for (a) binary **1 0**; (b) part of a random binary stream.

Hence the dc power dissipated in a $1\text{-}\Omega$ load is A^2 . Since most telephone lines are ac coupled, this power is dissipated at the input termination and does not contribute to the received signal. Also, for a $1\text{-}\Omega$ load, the average power for a binary **1** is $4A^2$ and zero for a binary **0**. Hence the total average power in the waveform, obtained by averaging over a binary **10** pair is $2A^2$. This is the total power, and hence the dc power is one-half of this.

The unipolar NRZ-L waveform can be thought of as an ac waveform superimposed on the dc level, and the power spectrum for the ac waveform alone can be found using the method of Section 2.13. The ac waveform is as shown in Fig. 3.4.2(b), where the basic pulse shape is seen to be rectangular of height A . Applying the results of Section 2.13 (the mathematical details are omitted here), the magnitude of the voltage spectrum density is

$$|V(f)| = AT_b |\operatorname{sinc} f T_b| \quad (3.4.3)$$

Applying Eq. (2.17.1), the power spectrum density is

$$\begin{aligned} G(f) &= \frac{1}{T_b} |V(f)|^2 \\ &= T_b (A \operatorname{sinc} f T_b)^2 \end{aligned} \quad (3.4.4)$$

This is plotted in Fig. 3.4.3, with the peak value normalized to unity. This spectrum is actually two-sided, requiring negative as well as positive frequencies, but the negative half is symmetrical with respect to the positive half and is not shown. Recall too that the spectrum shown here applies only to the ac component of the waveform, it being understood that a dc component of power exists. It will be seen that the spectrum peaks at the low-frequency end, indicating that the low-frequency content of the ac waveform is high (this is quite separate from the dc component). Because of the dc power, which is wasted as heat, and the considerable low-frequency content, the unipolar waveform is not suitable for general-purpose communications links. It is used for very short distance communications, for example, between pieces of digital equipment or circuits in close proximity to one another.

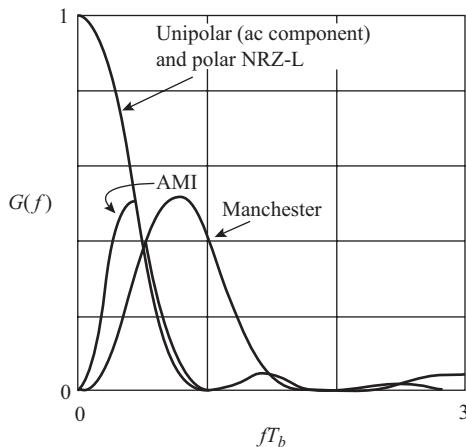


Figure 3.4.3 Power spectral density curves for NRZ waveforms: (a) unipolar (ac component only), (b) polar, (c) Manchester, (d) alternate mark inversion (AMI). The curves have been normalized to unity.

Polar NRZ-Line Code

The *polar* waveform has both positive and negative pulses as shown in Fig. 3.4.2. We might be tempted to call this a bipolar waveform, and, in fact, European practice is to do so. However, in North America, the term bipolar is reserved for a particular form of alternate polarity pulses to be described shortly.

For an infinitely long random binary stream, there will be as many **1**'s as **0**'s on average, assuming these are equiprobable binites, and it follows therefore that the average or dc level for the waveform is zero, compared to A for the unipolar waveform. Also, the total signal power in this case is A^2 in a $1\text{-}\Omega$ load compared to $2A^2$ for the unipolar waveform. Thus the polar NRZ waveform requires one-half the power of a unipolar NRZ waveform for the same swing between levels.

Since both the polar waveform and the ac component of the unipolar waveform swing between $\pm A$, the results for the power spectrum density for the ac component of the unipolar waveform [Eq. (3.4.4)] apply directly to the polar waveform, as shown in Fig. 3.4.3.

DC Wander

As already seen, ac coupling requires that the output waveform should have a zero mean level. When, therefore, the input is a long string of like-polarity pulses, the output level has to adjust to maintain zero mean. This is illustrated in Fig. 3.4.4(a). The input is a string of negative-going pulses, which in fact may be thought of as one long negative pulse. AC coupling will pass the initial rising transient, but thereafter the output signal level will decay. At the end of the string, the output overshoots zero and becomes positive and then decays to zero. The overshoot occurs to bring the average value to zero, the signal energy in the overshoot being obtained from storage elements (L and C) in the transmission system.

It will be seen that, in effect, the pulse levels decay and may in fact decay below the threshold level of the level detector in the receiver. This effect is termed *dc wander*, because the dc level of the signal shifts to maintain zero mean at the output. Figure 3.4.4(b) shows how the dc wander affects pulses that follow a long string of like symbols. DC restorer circuits can be and are used in receivers in some systems, but a better solution is to modify the waveform to eliminate the low-frequency content. This is discussed shortly.

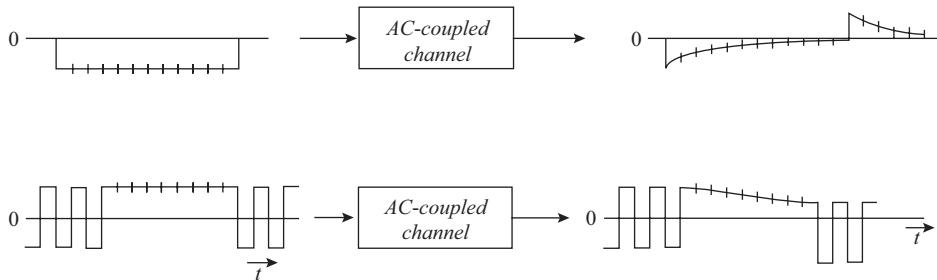


Figure 3.4.4 DC wander. (a) A long string of negative input pulses and the corresponding output waveform. (b) Mixed input, showing shift of dc level at output.

Long strings of like symbols also create a problem in recovering the timing information from the waveform. As mentioned previously, the output waveform at the receiver is reconstructed from samples of the transmitted waveform taken every bit period. The transitions in the waveform are used to synchronize the sampling clock with the bit rate, and therefore, if the waveform has few transitions over a given period, such as occur with long strings of **1**'s or **0**'s, the synchronization may be lost. Again, modifications to the basic waveform to introduce more transitions is the favored solution, and this can be achieved by using RZ pulses rather than NRZ pulses. Alternative coding techniques can also be used, as described next.

Manchester Line Code

The Manchester code, shown in Fig. 3.4.5, is a NRZ-L code in which the bunits are encoded as transitions between levels. A binary **1** is coded as a transition from $+A$ to $-A$, and a binary **0** by a transition from $-A$ to $+A$. The transitions take place at the midpoint on a pulse so that the dc level is entirely eliminated. Also, since transitions always occur, timing information derived from these is always available. (Bit-timing recovery is required in the receiver and is described in Chapter 12.)

The disadvantage of the Manchester code (also known as a split-phase code) is that it requires twice the bandwidth of the polar NRZ-L code. This can be seen in a simple way. The highest frequency for the Manchester code occurs when a very long string of like pulses is being transmitted. Under these conditions, the waveform appears like a square wave of periodic time T_b . This is shown in Fig. 3.4.5(c). With the polar waveform, the highest frequency occurs when pulses of alternate polarity are being transmitted, and the periodic time of the resulting square wave is $2T_b$, as shown in Fig. 3.4.5(d). Hence the highest frequency expected in the Manchester code is twice that expected in the polar code.

The power spectrum density for the Manchester code is found as follows. The magnitude of the voltage spectrum density for the basic pulse shape of Fig. 3.4.5 (obtained by taking the Fourier transform, the details of which are omitted here) is

$$|V(f)| = T_b A \left| \operatorname{sinc} \frac{f T_b}{2} \sin \frac{\pi f T_b}{2} \right| \quad (3.4.5)$$

From the results of Section 2.13, the power spectrum density is

$$\begin{aligned} G(f) &= \frac{1}{T_b} |V(f)|^2 \\ &= T_b \left(A \operatorname{sinc} \frac{f T_b}{2} \sin \frac{\pi f T_b}{2} \right)^2 \end{aligned} \quad (3.4.6)$$

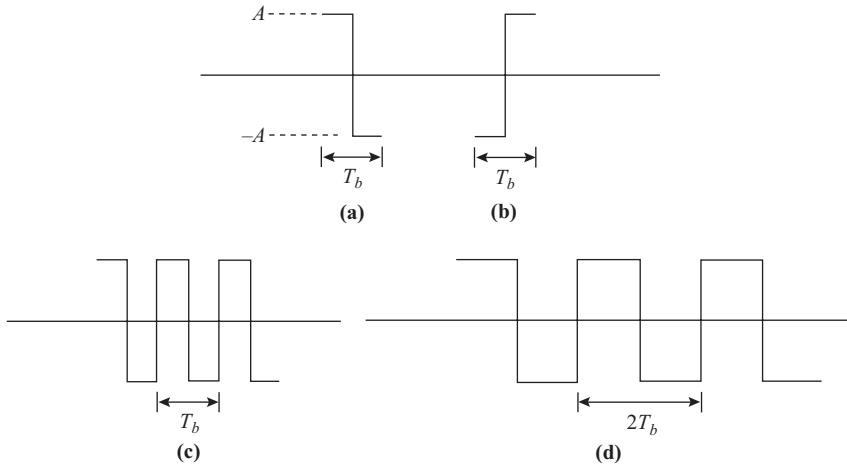


Figure 3.4.5 Basic pulse shapes used in the Manchester line code: (a) binary 1 and (b) binary 0. (c) The sequence ...1111... in the Manchester code. (d) The sequence ...1010... in the NRZ-L polar code.

This is plotted in Fig. 3.4.3, normalized to unity magnitude, where it will be seen that the power spectrum density is zero at zero frequency, and the first null occurs at $f = 2/T_b$, compared to $f = 1/T_b$ for the polar waveform.

Alternate Mark Inversion (AMI) Line Code

In the AMI code, binary 0's are coded as zero voltage, and binary 1's are coded alternately $+A$ and $-A$ (Fig. 3.4.6), which gives rise to the name *alternate mark inversion* (AMI) code. The code is also known as a *bipolar code*. As already mentioned, the word bipolar is used in a restricted sense, because, strictly speaking, the polar waveform is also bipolar. Alternating the binary 1 symbol between positive and negative voltages of equal magnitude means that the dc level remains at zero even where long strings of binary 1's occur; that is, the dc wander problem is eliminated. The code is relatively easy to implement and is widely used for digital communications. However, it is less efficient, in a coding sense, than the other forms of coding described because it requires three voltage levels to encode the two binary symbols, rather than just two levels. Because it uses three levels to encode two symbols, it is also known as a *pseudoternary code*.

Because the coding symbols are not independent of one another, the power spectrum density is more difficult to derive; but the results of spectrum analysis yield the following expression for the power spectrum density:

$$G(f) = T_b(A \operatorname{sinc} fT_b \sin \pi fT_b)^2 \quad (3.4.7)$$

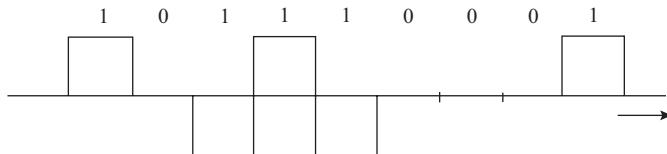


Figure 3.4.6 Alternate mark inversion (AMI) sequence.

This is seen to be similar to that for the Manchester code, but with the nulls occurring at $1/T_b$ rather than $2/T_b$. The power spectrum density for the AMI code is plotted in Fig. 3.4.3, where it is seen that the spectrum width is half that of the Manchester code. Like the Manchester code, the AMI code does not lose timing information where long strings of 1's occur. However, timing information may be lost where long strings of 0's occur. To combat this, a special class of codes based on the AMI code has been devised, known as high-density bipolar codes.

High-density Bipolar (HDB n) Line Codes

High-density bipolar codes are so called because they contain a higher density of marks (binary 1's) than the normal bipolar (AMI) code. More specifically, a HDB n code is one in which sequences of more than n zeros are encoded as special sequences that can be identified according to certain rules. To illustrate these, S will be used to indicate a Space (encoding for binary 0), P a positive pulse (encoding for a binary 1), and N a negative pulse (also encoding for a binary 1). The most popular is the HDB3 code, in which a **0000** sequence is encoded by one of the special sequences SSSX, or YSSX. Here, X stands for a mark that violates the AMI code and Y for a mark that does not violate the AMI code. Thus the SSSX sequence will be either an SSSN or SSSP, depending on the previous mark. The YSSX will be either a PSSP or NSSN again, depending on the previous mark.

The SSSX sequence is used for the first occurrence of **0000** and where an odd number of 1's occurs between successive **0000** sequences in the original message. The YSSX sequence is used where an even number of 1's occurs between successive **0000** sequences in the original message. This is illustrated in the following tables. In Table 3.4.1, the initial sequence **101** is AMI encoded as PSN. The first **0000** sequence to be encountered is encoded by SSSN, since this violates the AMI code. An odd number of 1's (one in this case) occurs between this and the next **0000** sequence, and therefore this second sequence is encoded as SSSP. It follows that where an odd number of 1's occurs between substitution sequences the encoding will alternate between SSSP and SSSN, and hence the dc component is eliminated (or at least attenuated).

The data in Table 3.4.1, along with the corresponding AMI code, are shown in Fig. 3.4.7.

TABLE 3.4.1

Original message	1	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0
HDB3	P	S	N	S	S	S	N	S	P	S	S	S	P	N	P	S

In Table 3.4.2, the original data sequence is similar to that in Table 3.4.1 except that the eighth and ninth bits are **1, 1** instead of **0, 1**. This places an even number of 1's between the successive **0000** sequences, and so the second **0000** sequence is encoded as PSSP.

The data in Table 3.4.2, along with the corresponding AMI code, are shown in Fig. 3.4.8.

The equation for the power spectrum for the HDB3 code is difficult to derive, because the encoded bits are not independent, and on average there will be more marks (P's and N's) than spaces (S's). However, the curve is similar to that for the normal AMI code, occupying about the same bandwidth. The total average

TABLE 3.4.2

Original message	1	0	1	0	0	0	0	1	1	0	0	0	0	1	1	0
HDB3	P	S	N	S	S	S	N	P	N	P	S	S	P	N	P	S

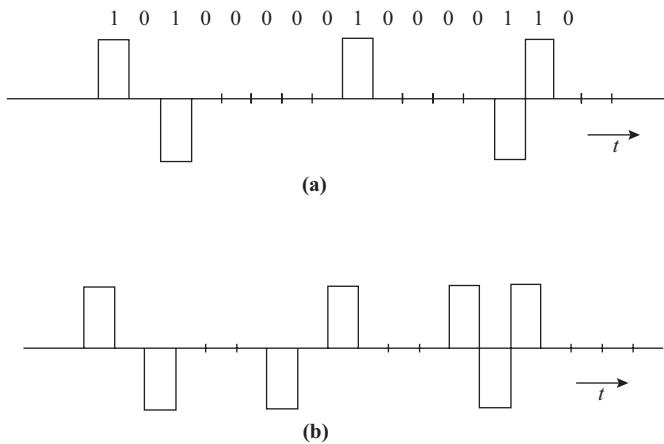


Figure 3.4.7 (a) AMI code and (b) the HDB3 code for the data of Table 3.4.1.

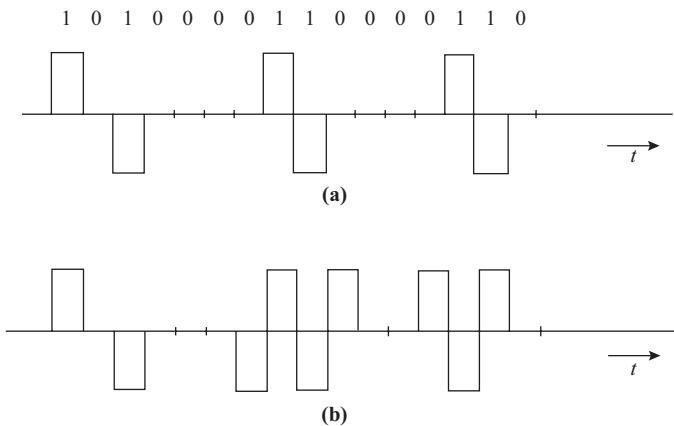


Figure 3.4.8 (a) AMI code and (b) the HDB3 code for the data of Table 3.4.2.

power in the HDB3 waveform is somewhat greater (about 10%) than that for the AMI waveform because of the increase in the number of mark pulses it contains. This results in a double-peaked spectrum, the peaks being slightly higher than the single peak for the AMI case.

Differential Encoding

It is possible for a binary signal to be unknowingly inverted during transmission; that is, **1**'s are converted to **0**'s and **0**'s to **1**'s. With digitized speech signals this may not matter, but it is a serious matter with data signals. To overcome this problem, the original data signal is often transmitted by comparing any given bit with the previous bit. If it is the same, a **0** is transmitted and, if different, a **1** is transmitted. This is known as *differential encoding*. A reference bit, which can be **0** or **1**, must be transmitted at the beginning of the message in order for the first message bit to be encoded differentially.

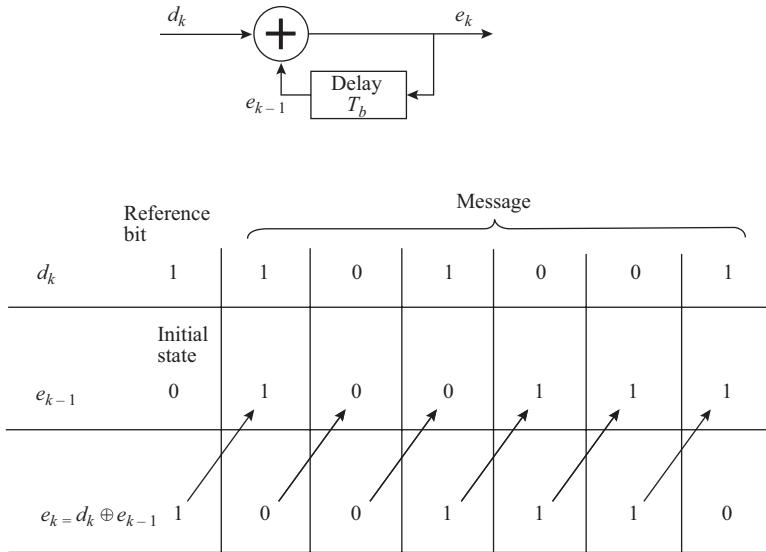


Figure 3.4.9 Differential encoding of a binary message.

Figure 3.4.9 shows how differential encoding may be achieved. A modulo 2 adder is shown, the arithmetic rules for which are

$$\begin{aligned} 0 \oplus 0 &= 0 \\ 0 \oplus 1 &= 1 \\ 1 \oplus 0 &= 1 \\ 1 \oplus 1 &= 0 \end{aligned}$$

The input binary sequence is denoted by d_k , where the subscript k denotes the k th bit. The output sequence is denoted by e_k , and the feedback sequence, which is the output sequence delayed by a 1-bit period, is denoted by e_{k-1} .

The initial output state of the modulo 2 adder is **0**, and with the reference bit a **1**, the output is also a **1**, as shown in Fig. 3.4.9. This establishes the feedback sequence. The first message bit is a **1**, and when this is modulo 2 added to the feedback bit **1**, the output becomes a **0**. This output bit is also the delayed feedback bit, which is modulo 2 added to the next message bit. The next message bit is a **0**, and modulo 2 addition gives an output of **0**. The operation continues in this way as shown by the sample sequences in Fig. 3.4.9.

Decoding the differentially encoded signal is achieved as shown in Fig. 3.4.10. Considering first the situation, where no phase inversion occurs in transmission, it is assumed that the received logic signal is e_k . The operation of the decoder is similar to that of the encoder, the decoding sequence being illustrated in the first table in Fig. 3.4.10. It is seen that the output sequence matches the input sequence of the encoder shown in Fig. 3.4.9; that is, decoding is achieved.

Consider what happens now if an inversion takes place in the transmission. The signal e_k is converted to its complement \bar{e}_k , but so also is e_{k-1} converted to \bar{e}_{k-1} , and hence the output of the modulo 2 adder is the same as that obtained without inversion. This is shown in the second table in Fig. 3.4.10.

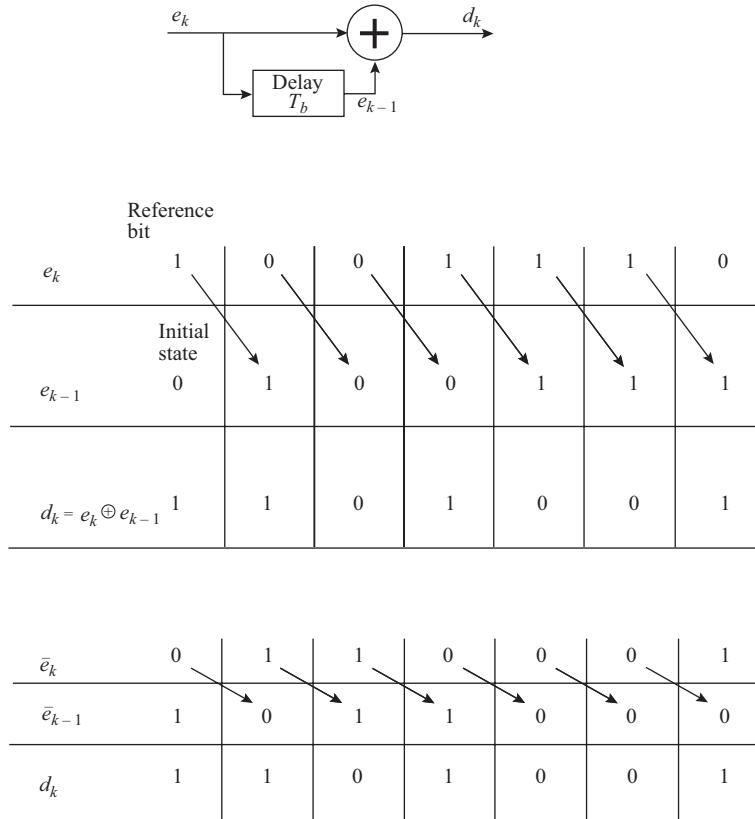


Figure 3.4.10 Decoding of a differentially encoded binary message.

These statements can be proved more formally from the rules of binary arithmetic. The encoder output is

$$e_k = d_k \oplus e_{k-1} \quad (3.4.8)$$

At the output of the digital transmission system, the received signal will be either e_k or its complement \bar{e}_k , if inversion takes place. For the first possibility, the output is

$$\begin{aligned} e_k \oplus e_{k-1} &= d_k \oplus e_{k-1} \oplus e_{k-1} \\ &= d_k \end{aligned} \quad (3.4.9)$$

This last statement follows because $e_{k-1} \oplus e_{k-1} = 0$. For the second possibility, the output is

$$\begin{aligned} \bar{e}_k \oplus \bar{e}_{k-1} &= d_k \oplus \bar{e}_{k-1} \oplus \bar{e}_{k-1} \\ &= d_k \end{aligned} \quad (3.4.10)$$

This result makes use of the fact that $\overline{d_k \oplus e_{k-1}} = d_k \oplus \bar{e}_{k-1}$. Thus, in either case, the correct output is obtained.

3.5 M-ary Encoding

With M -ary encoding the symbol rate is less than the bit rate, as shown by Eq. (3.2.8), repeated here for reference:

$$R_{\text{sym}} = \frac{R_b}{m} \quad (3.5.1)$$

m is the number of bits contained in a symbol and $M = 2^m$. As with binary transmission, polar transmission for M -ary waveforms is more efficient in terms of power when the levels are centered on zero. For a separation of $2A$ between levels (as used for the binary waveforms), the M -ary waveform requires that

$$\begin{aligned} a_k &= 0, \pm 2A, \pm 4A, \dots, \pm(M - 1)A, && \text{for } M \text{ odd} \\ &= \pm A, \pm 3A, \pm 5A, \dots, \pm(M - 1)A && \text{for } M \text{ even} \end{aligned} \quad (3.5.2)$$

Figure 3.5.1 shows an example where the binary stream is partitioned into groups of three, thus requiring an eight-level code.

The advantage of M -ary encoding is that it allows information at a fixed bit rate to be transmitted over a smaller bandwidth system compared to binary transmission. To illustrate this, consider NRZ-L waveforms. The highest frequency is encountered when the signal alternates between maximum and minimum levels; that is, a square wave is generated of period $2T_b$ for the binary transmission and period $2T_{\text{sym}}$ for the M -ary transmission. Hence the frequency of the square wave, which is the reciprocal of the periodic time, is $f_b = R_b/2$ for binary transmission and $f_M = R_{\text{sym}}/2$ for M -ary. But since $R_M = R_b/m$, it follows that

$$f_M = \frac{f_b}{m} \quad (3.5.3)$$

The M -ary system requires $1/m$ of the bandwidth required for the binary system, or, alternatively, for a given bandwidth the M -ary system can transmit information at m times the bit rate of the binary system. M -ary

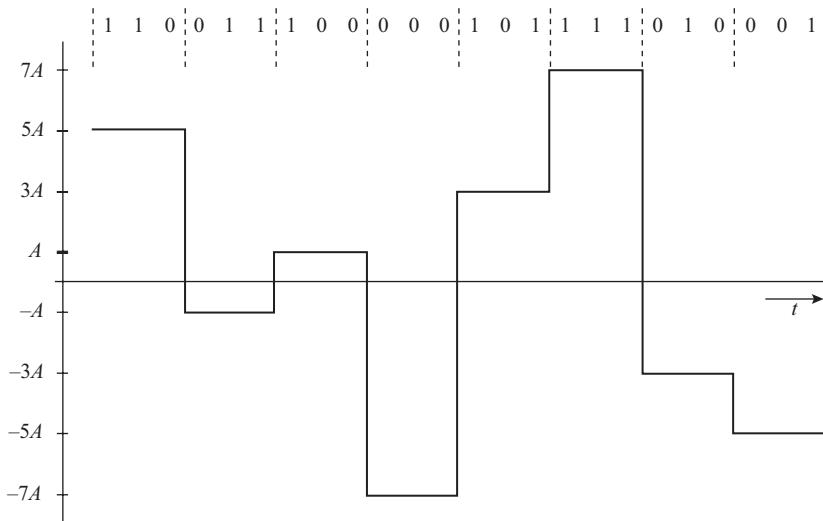


Figure 3.5.1 M -ary coding with $M = 8$ levels.

coding suffers from the disadvantage that it requires more power in order to maintain the same peak-to-peak excursion between adjacent levels, compared to binary. Also, the circuitry required to implement it is more complex. However, M -ary encoding is used extensively in digital radio systems, and some of these systems will be examined in later chapters.

3.6 Intersymbol Interference

When a waveform is transmitted over a communications channel, the frequency response of the channel will, in general, introduce *linear distortion*. Linear distortion means that the shape of the waveform undergoes a change, but no new frequency components are generated in the spectrum. However, the linear distortion can take the form of “ringing,” in which a pulse, for example, may have a “tail” added, as illustrated in Fig. 3.6.1.

The tail is a result of the natural buildup and decay of energy in the inductive and capacitive elements in the transmission path, rather in the same way that a resonant circuit behaves. The problem this creates with digital waveform transmission, for example, with a polar binary waveform, is that many of the tails can combine at some later time to produce a polarity inversion in the waveform. If the inversion coincides with the time that the waveform is sampled to determine if a **0** or **1** is present, a bit error will occur. The interference caused by the tails is referred to as *intersymbol interference* (ISI). By keeping the waveform spectrum much narrower than the frequency bandwidth of the transmission channel, the linear distortion, and hence the ISI, can be held to negligible levels, but in general this is not an efficient solution to the problem. As shown in Fig. 3.4.3, the main lobe of the spectrum for a polar NRZ-L waveform occurs at $f = 1/T_b$. Thus decreasing the waveform spectrum means increasing the bit period, which in turn means that fewer bits per second can be transmitted.

As an empirical guide, to avoid ISI with rectangular pulses the product of the -3 dB bandwidth and the pulse time-width should not be less than 0.5. Denoting the product by BT, then $BT \geq 0.5$.

In practice, it is not necessary to preserve the shape of the waveform if it can be sampled at suitable instants and a new waveform generated from the samples. Suitable instants would be at the center of the pulses making up the waveform, where the pulses are near their maximum amplitudes. Correct synchronization is required to ensure that the pulses are sampled at the optimum times, and this topic is covered in Chapter 12. It is still necessary, however, to prevent or minimize the intersymbol interference, and this is achieved through pulse shaping, as described next.

3.7 Pulse Shaping

The shape of a pulse at the output or receiving end of a transmission channel is determined by the spectrum of the input pulse, the frequency response of the transmitter, the frequency response of the channel, and the frequency response of the receiver. This is shown diagrammatically in Fig. 3.7.1.

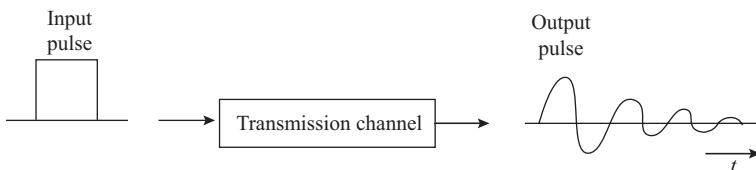
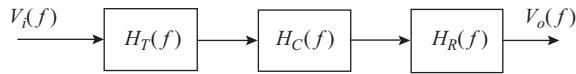


Figure 3.6.1 Linear distortion of a pulse that produces “ringing.”



$$V_o(f) = V_i(f)H_T(f)H_C(f)H_R(f)$$

$$V_o(t) = \mathcal{F}^{-1} \{V_o(f)\}$$

Figure 3.7.1 Factors affecting the spectrum of an output pulse.

Denoting the input pulse spectrum by $V_i(f)$, the transmit filter by $H_T(f)$, the channel frequency response by $H_C(f)$, and the receive filter by $H_R(f)$, the spectrum of the output pulse is given by

$$V_o(f) = V_i(f)H_T(f)H_C(f)H_R(f) \quad (3.7.1)$$

In general there will be a transmission delay, and for present purposes this is included in the channel response $H_C(f)$. To correctly shape the output pulse, its spectrum is shaped by adjustment of $H_T(f)$ or $H_R(f)$, or both. One major advantage of digital systems is that shaping the spectrum in this way has no effect on the information content of the signal, in contrast to analog systems, in which the message spectrum, and hence the analog waveshape, is directly affected by the frequency response of the transmission system. The designer also has control over the input pulse, although this is usually chosen to be rectangular. The frequency response of the channel is seldom within the control of the designer.

Figure 3.7.2 shows in a general way how pulse shaping can be used to avoid ISI. Three rectangular pulses representing **010** are shown at the input to the system. The output pulses have tails that overlap, but the zeros in the tails are arranged to occur at the sampling instants. Pulses that are shaped to eliminate ISI are known as *Nyquist pulses*. One particular shape that is widely used is the *raised-cosine response*, named

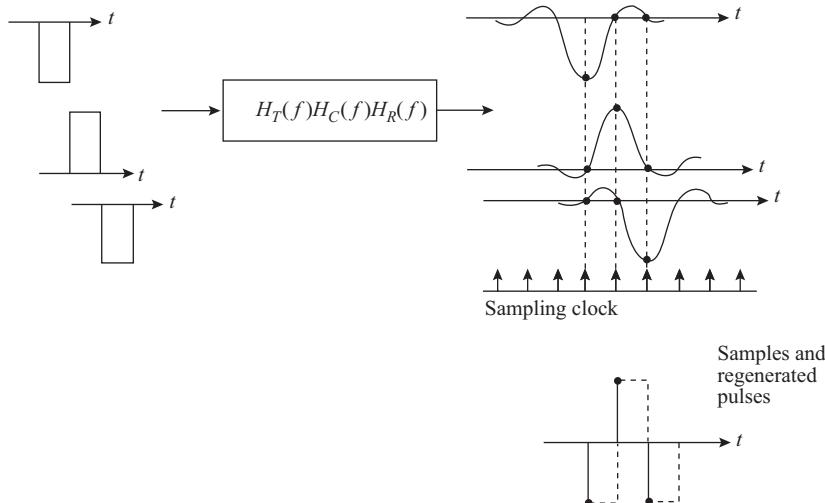


Figure 3.7.2 Pulse shaping to avoid ISI.

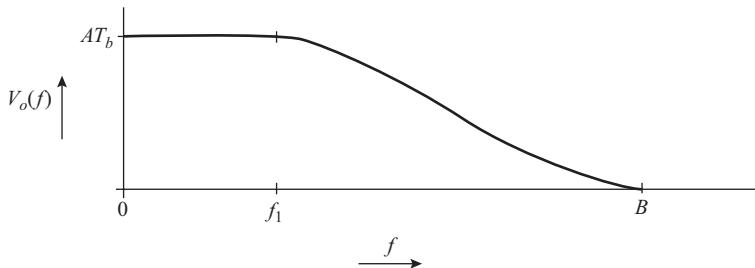


Figure 3.7.3 Raised-cosine spectrum.

after the shape of its spectrum. The spectrum is sketched in Fig. 3.7.3, where A is the peak value of the pulse in the time domain and T_{sym} is the symbol period, not the pulse width. As will be shown shortly, the pulse is spread out in time, and therefore a precise pulse width cannot be defined. The frequency spectrum of the raised cosine pulse can be described mathematically by

$$\begin{aligned} V_o(f) &= AT_{\text{sym}}, \quad \text{for } |f| < f_1 \\ &= AT_{\text{sym}} \cos^2 \frac{\pi(|f| - f_1)}{2(B - f_1)} \quad \text{for } f_1 < |f| < B \\ &= 0 \quad \text{for } |f| > B \end{aligned} \quad (3.7.2)$$

The frequencies f_1 and B are determined by a design parameter known as the *roll-off factor*, denoted here by the symbol ρ and by the symbol period. The design equations are

$$B = \frac{1 + \rho}{2T_{\text{sym}}} \quad (3.7.3)$$

$$f_1 = \frac{1 - \rho}{2T_{\text{sym}}} \quad (3.7.4)$$

The roll-off factor is a specified parameter that lies between the limits $0 \leq \rho \leq 1$. Strictly, the raised cosine response is a theoretical model, but it is one that can be closely approximated in practice for moderate to high values of ρ . It is left as an exercise for the student to show that when $\rho = 0$ the raised cosine spectrum becomes rectangular in shape, and this is usually referred to as the ideal low-pass response. Recalling that the symbol rate is given by $R_{\text{sym}} = 1/T_{\text{sym}}$ the equations relating to the raised cosine pulse can be rewritten in terms of the symbol rate if desired.

The time waveform corresponding to the raised-cosine spectrum is obtained by taking the inverse Fourier transform of $V_o(f)$. The mathematical details will be omitted here, but the result is

$$v(t) = A \operatorname{sinc} \frac{t}{T_{\text{sym}}} \cdot \frac{\cos \rho \pi t / T_{\text{sym}}}{1 - (2\rho t / T_{\text{sym}})^2} \quad (3.7.5)$$

It will be seen that the sinc function, previously encountered in the frequency domain, now enters into the time domain also. The shape of the pulse is shown in Fig. 3.7.4 for a roll-off factor of 0.25. The pulse

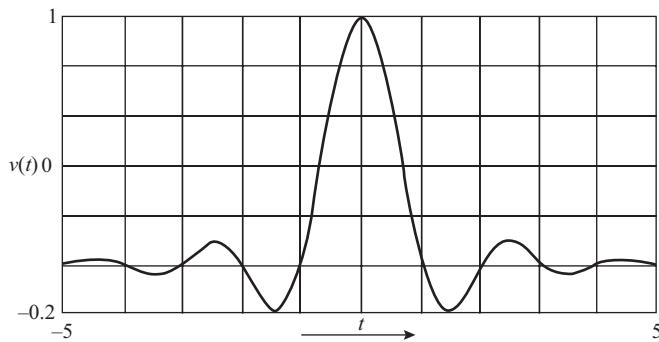


Figure 3.7.4 Raised-cosine pulse with $\rho = 0.25$.

has its maximum at $t = 0$, and it has periodic zeros at integer multiples of the symbol period, $t = kT_{\text{sym}}$. This can be seen by rewriting the pulse equation as

$$v(kT_{\text{sym}}) = A \operatorname{sinc} k \cdot \frac{\cos \rho \pi k}{1 - (2\rho k)^2} \quad (3.7.6)$$

The sinc function is equal to zero for integer values of k and is maximum at unity for $k = 0$. The cosine term is also unity for $k = 0$. Therefore, sampling the pulse at $k = 0$ yields $v(0) = A$, while sampled at any other integer value, k yields zero output. This means that a waveform consisting of a sequence of such pulses will have no ISI. Denoting the basic pulse shape by

$$p(t) = \operatorname{sinc} \frac{t}{T_{\text{sym}}} \cdot \frac{\cos \rho \pi t / T_{\text{sym}}}{1 - (2\rho t / T_{\text{sym}})^2} \quad (3.7.7)$$

allows the waveform to be written in the form given by Eq. (3.4.1):

$$v(t) = \sum_{k=-\infty}^{\infty} a_k p(t - kT_{\text{sym}}) \quad (3.7.8)$$

For example, the resultant binary waveform for the sequence **010** is shown in Fig. 3.7.5. Although the pulses interfere with one another, by sampling at integer multiples of the bit period, ISI is avoided. In practice, some mistiming is likely to occur (referred to as timing jitter), which results in sampling at points where the ISI is not zero. However, the timing jitter is kept to a minimum by careful design. With the raised cosine pulse, the denominator term $[1 - (2\rho k)^2]$ increases rapidly for large values of k , so the tails of the pulse decrease rapidly, which helps to minimize the ISI.

The raised-cosine response when $\rho = 0$ has special significance. For this condition it is seen from Eqs. (3.7.3) and (3.7.4) that $B = f_1 = 1/2T_{\text{sym}}$. The spectrum shown in Fig. 3.7.6(a) becomes rectangular, and it is the narrowest bandwidth spectrum that still avoids ISI. It is referred to as the *ideal* low-pass spectrum because the rectangular shape cannot be realized in practice. It does, however, provide a reference against which system performance can be assessed.

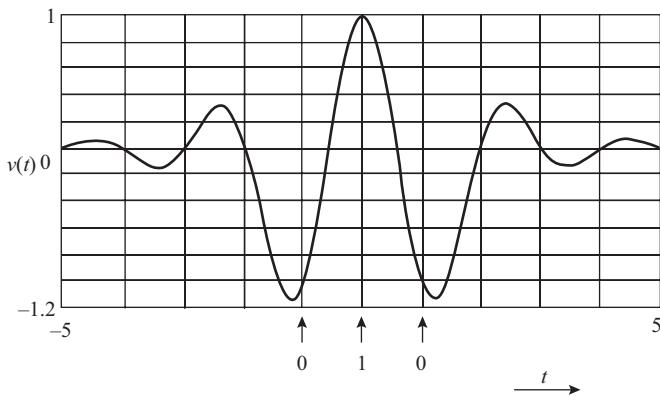


Figure 3.7.5 Raised-cosine pulses with $\rho = 0.25$ for the binary sequence **010**.

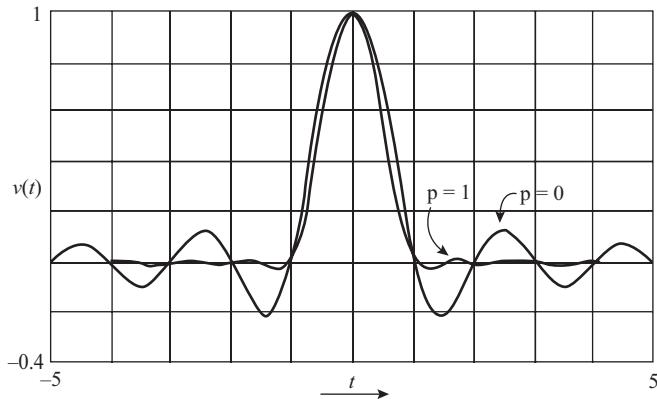


Figure 3.7.6 Raised-cosine response for $\rho = 0$ and $\rho = 1$.

The corresponding pulse in the time domain is obtained from Eq. (3.7.5) as

$$\begin{aligned} v(t) &= A \operatorname{sinc} \frac{t}{T_{\text{sym}}} \\ &= A \operatorname{sinc} R_{\text{sym}} t \end{aligned} \quad (3.7.9)$$

This is shown in Fig. 3.7.6. The pulse samples taken at intervals $kT_{\text{sym}} = k/R_{\text{sym}}$ are given by $A \sin k$, which is equal to A for $k = 0$ and is zero for all other values of k (recalling that k is integer). The problem with the ideal response, apart from the fact that it cannot be realized in practice, is that the pulse tails do not decrease nearly as rapidly as when $\rho > 0$, and so the ISI that might occur as a result of timing jitter could be severe.

For the ideal response

$$\begin{aligned} B &= \frac{1}{2T_{\text{sym}}} \\ &= \frac{R_{\text{sym}}}{2} \end{aligned} \quad (3.7.10)$$

This shows that the signaling rate in symbols per second is equal to twice the ideal bandwidth. An important parameter in digital communications is the ratio of bit rate to the spectrum bandwidth. Denoting this ratio by α , then for the raised-cosine response

$$\begin{aligned}\alpha &= \frac{R_b}{B} \\ &= \frac{R_b}{(1 + \rho)R_{\text{sym}}/2} \\ &= \frac{2m}{1 + \rho}\end{aligned}\tag{3.7.11}$$

The units for α are bps/Hz, which in fact is a dimensionless quantity, since hertz are measured in cycles per second, and both cycles and bits are dimensionless. However, using bps/Hz helps to keep track of the meaning of the parameter.

For the ideal bandwidth system ($\rho = 0$), $\alpha = 2m$ and thus increasing m makes more efficient use of available bandwidth in terms of bps/Hz. However, more complex circuitry is needed to implement M -ary coding. Also, if it is desired to maintain the same separation between levels as in the binary system (a condition required to maintain the same performance against noise), then higher power levels are required for the M -ary system.

It is seen that the pulses at the receiver do not have well-defined shapes (such as rectangular) and, in addition, noise is an inescapable part of any communication system, which will further distort the wave-shape. The result is that the waveform at the receiver will bear very little resemblance to the well-defined line waveforms described in Section 3.4. The function of the pulse regenerator in the receiver is to regenerate a “clean” waveform from the distorted and noisy signal, from which the data can be recovered with as few errors as possible. This topic is taken up again in Chapter 12.

PROBLEMS

- 3.1.** Briefly discuss the difference between analog and digital signals. A thermostat is set so that it generates a voltage of 5 V at its low setting and a voltage of 10 V at its high setting. Are these signal levels analog or digital or is the situation indeterminate?
- 3.2.** The thermostat in Problem 3.1 generates a continuous voltage proportional to temperature between its low and high limits. Is this an analog or digital signal?
- 3.3.** Define and explain the terms *binit*, *bit*, and *baud*. Which of these are symbols?
- 3.4.** A discrete source contains 32 symbols, each of which has equal probability of being selected for transmission. Calculate the information in each symbol. What is the information content in the transmission of two successive symbols?
- 3.5.** In a certain binary transmission system, the probability of a binary **1** being transmitted is 0.6. What is the probability of a binary **0** being transmitted? What is the information content in each binit?
- 3.6.** The transmission rate in a binary system is 3000 binit per second. What is the rate in bauds? Given that the rate is also equal to 3000 bps, what condition is implied by this statement?
- 3.7.** For the source alphabet of Problem 3.4, determine the number of binit required to encode each symbol into a binary code. Given that the binary code is transmitted at the rate of 1000 bps, what is the source-symbol transmission rate?

- 3.8. The binary code in Problem 3.7 is recoded into a quaternary code. Determine the code-symbol transmission rate.
- 3.9. A rectangular pulse has an amplitude of 5 V and a width of 3 ms. Express this in the notation of Section 3.3 for (a) a pulse centered on the zero time reference, (b) a pulse delayed by 7 ms from the zero time reference, and (c) a pulse advanced by 2 ms from the zero time reference. Sketch all three pulses.
- 3.10. A rectangular pulse of width τ and amplitude A starts at $t = 0$. Write the expression for the pulse in the notation of Section 3.3 and sketch the pulse.
- 3.11. A pulse shape can be described by $p(t) = \sin \pi t$ for $0 \leq t \leq 1$ s and zero for all other values of t . (a) Sketch the pulse. Write the expressions for similar pulses and sketch these for (b) a time delay of 0.5 s, and (c) a time advance of 0.5 s.
- 3.12. Explain what is meant by (a) return-to-zero (RZ) and (b) not-return-to-zero (NRZ) pulses. Discuss briefly the reasons why both types are used in practice. Information bits of period T_b are encoded in rectangular pulses of both types. Denoting the pulse width by τ , express the bits in the notation of Section 3.3 when the bit period is centered about $t = 0$ and the pulse width for the RZ pulse is one-half the bit period.
- 3.13. For a unipolar binary signal, the a_k terms of Eq. (3.4.1) for $k = -3, -2, -1, 0, 1$ are 0, 1, 1, 0, 1. Write out the corresponding terms of Eq. (3.4.1) for a bit period of 1 μ s, showing the time range for each term.
- 3.14. A finite binary message **11100** is transmitted as a unipolar NRZ-L waveform using a rectangular pulse of height 5 V for binary 1's. Calculate the energy dissipated in a 1- Ω load resistor when the waveform is developed across this, the pulse width being 2 ms.
- 3.15. Explain why $\text{sinc } x = 1$ for $x = 0$, and $\text{sinc } x = 0$ for $x = \pm 1, \pm 2, \dots$
- 3.16. Given that $\text{sinc } x = 0.2$, determine the value of x .
- 3.17. Plot the function $\text{sinc } x$ for x in steps of 0.1 over the range $-3 \leq x \leq 3$.
- 3.18. For a unipolar NRZ-L waveform using rectangular pulses of height $2A$ volts, the dc component of power is A^2 for a load resistance of 1 Ω . State the conditions that apply to the waveform for this relationship to hold. The power density spectrum for such an NRZ-L waveform is given by Eq. (3.4.4). Derive the expression for power spectrum density at zero frequency. What is the distinction between this and the dc component of power?
- 3.19. A unipolar binary waveform has the spectrum density function given by Eq. (3.4.4). Given that the basic pulse for the waveform has a magnitude of 5 V and a width of 2 μ s, calculate the double-sided spectrum density at a frequency of 425 kHz, stating clearly the units for this. What would be the one-sided spectrum density at this frequency?
- 3.20. Assuming that the curve given by Eq. (3.4.3) can be approximated as flat at its peak value for a frequency range of $\pm 5\% / T_b$ about zero, calculate the power (for a 1 Ω load) in this range centered at zero frequency. The pulse amplitude is 1 V and width is 5 μ s.
- 3.21. What is the main advantage of the polar NRZ-L waveform compared to the unipolar version? A digital signal consisting of an infinitely long stream of random and equiprobable binary symbols is encoded as a polar NRZ-L waveform. The pulses are rectangular of amplitude 1 V and width 3 ms. Assuming the waveform is developed across a 1- Ω resistor, calculate (a) the dc power, (b) the average signal power, and (c) the power spectrum density at zero frequency.
- 3.22. A finite binary sequence **10101110** is encoded as a polar NRZ-L using rectangular pulses of height 3 V and width 5 μ s. Sketch the waveform and determine the waveform energy, assuming a 1- Ω load resistance.

- 3.23. Given that the average signal power in an infinitely long, random, polar NRZ-L waveform is A^2 , where A is the pulse height, and using Eq. (3.4.4), deduce the value of the area under the $(\text{sinc } fT_b)^2$ curve from $f = -\infty$ to $f = \infty$.
- 3.24. Explain what is meant by dc wander, and why this is to be avoided in a digital transmission system.
- 3.25. State the main advantages and main disadvantages of the Manchester code. A finite binary sequence **111000** is to be transmitted using the Manchester code. Sketch the waveform, and calculate the energy in the waveform assuming a 1Ω load resistance and rectangular pulses of 1 V.
- 3.26. Calculate the power spectrum density for an infinitely long random binary sequence encoded as a Manchester waveform at the following frequencies: (a) zero, and (b) $0.5/T_b$, $1/T_b$, and $2/T_b$. The pulse height is 5 V and the bit period is 1 ms.
- 3.27. Explain what is meant by an alternate mark inversion (AMI) code and why this is also referred to as a pseudoternary code. Sketch the AMI waveform for the binary sequence **11010011**.
- 3.28. A digital message consists of an infinitely long random sequence of equiprobable bunits, which is encoded as an AMI waveform. What is the probability of a mark being encoded as a negative pulse? What is the average signal power in such a waveform? What is the power spectrum density at zero frequency? A pulse height of A volts and a load resistor of 1Ω may be assumed.
- 3.29. A binary sequence **111000111** is to be transmitted on a digital link. Compare the average energies for the following waveforms: (a) polar NRZ-L, (b) Manchester, and (c) AMI. The same pulse width and peak value are used in all cases.
- 3.30. Give the reason for the use of high-density bipolar codes, and show that these are a development of the AMI code. A binary sequence **100000011** is encoded as an HDB3 code. Sketch the resulting waveform.
- 3.31. The binary sequence **10100000001110010** is to be encoded in the HDB3 code. Sketch the resulting waveform.
- 3.32. The binary sequence **1000001111000011001** is to be encoded in the HDB3 code. Sketch the resulting waveform.
- 3.33. Draw and compare the waveform for the binary sequence **10100010000110000** encoded in (a) AMI and (b) HDB3 line codes.
- 3.34. The binary sequence shown in Problem 3.33 is differentially encoded. Given that a reference **1** bit is inserted ahead of the sequence, write out the differentially encoded sequence.
- 3.35. A binary stream is being generated at the rate of 64 kbps and is to be encoded into a quaternary (four-level) code in real time. Calculate the symbol rate for the quaternary code.
- 3.36. A polar M -ary code has eight levels, the spacing between levels being 1 V. Write down the permissible voltage levels for the code.
- 3.37. A polar M -ary code has five levels, the spacing between levels being 1 V. Write down the permissible voltage levels for the code.
- 3.38. A binary stream **1110110001010001** is partitioned in groups of two and encoded in quaternary code. Sketch the resulting waveform.
- 3.39. Explain what is meant by intersymbol interference. A rectangular pulse of width 1 ms is transmitted through a channel that can be modeled as an RC low-pass filter with a time constant of 1 μ s. At the receiver the pulses are sampled at their midpoint. Is ISI likely to be significant?
- 3.40. A rectangular pulse of width 1 μ s is transmitted through a channel that can be modeled as an RC low-pass filter with a time constant of 1 μ s. At the receiver the pulses are sampled at their midpoint. Is ISI likely to be significant?

- 3.41. Discuss briefly the factors that affect the shape of the output pulse in a digital transmission system.
- 3.42. Explain what is meant by the *raised-cosine response*. A transmission system has rectangular input pulses of width 5 ms, and the output response is shaped to be a raised-cosine curve with a roll-off factor of 0.5. Determine the cutoff bandwidth of the raised-cosine response. Determine also the frequency at which the raised-cosine section of the curve starts.
- 3.43. A transmission system has rectangular input pulses, and the output response is shaped to be a raised-cosine curve with a roll-off factor of 0.5. The transmission rate is 3000 bps. Determine the cutoff bandwidth of the raised-cosine response. Determine also the frequency at which the raised-cosine section of the curve starts.
- 3.44. A transmission system has a raised-cosine output for a rectangular input pulse of width 2 ms and a pulse amplitude of 1 V. Plot the raised-cosine response for roll-off factors of (a) 1, (b) 0.5, and (c) 0.
- 3.45. For the raised-cosine responses specified in Problem 3.44, plot the corresponding pulses in the time domain.
- 3.46. Show that for the raised-cosine response $T_b = 1/(B + f_1)$ and $\rho = (B - f_1)/(B + f_1)$.
- 3.47. A rectangular pulse is transmitted through a channel that has a raised-cosine output, the frequency parameters for which are $f_1 = 350$ kHz and $B = 650$ kHz. The peak value of the output pulse in the time domain is 1 V. Plot the output pulse as a function of time up to the third null.
- 3.48. The binary sequence **111** is coded as a rectangular pulse sequence $+A, -A, +A$ and transmitted through a channel having a raised-cosine response. Plot the output waveform as a function of t/T_b given that the roll-off factor is 0.6.
- 3.49. Develop a MATLAB program to generate a random binary sequence.
- 3.50. Generate and plot the $sinc(x)$ function using MATLAB.
- 3.51. Generate and plot a train of ten $sinc(x)$ pulses.
- 3.52. Using *dubits*, encode and plot the waveform for the binary sequence “000110111000111101”.
- 3.53. Plot the binary sequence “000110110111” when applied to a QPSK modulator.
- 3.54. Plot the binary sequence “01010101001000” when applied to an AMI encoder.
- 3.55. Plot the binary sequence “00110001111110” when applied to a 8-ary QAM.
- 3.56. Explore the following MATLAB functions to generate waveforms: (a) *rectpuls* (b) *saw-tooth* (c) *tripuls* and (d) *pulstran*.
- 3.57. Generate qam using the *modulate(.)* command in MATLAB.



Noise

4.1 Introduction

Noise, as commonly understood, is a disturbance one “hears,” but in telecommunications the word noise is also used as a label for the electrical disturbances that give rise to audible noise in a system. These electrical disturbances also appear as interference in video systems, for example, the white flecks seen on a television picture when the received signal is weak, referred to as a “noisy picture.”

Noise can arise in a variety of ways. One obvious example is when a faulty connection exists in a piece of equipment, which, if it is a radio receiver, results in an intermittent or “crackling” type of noise at the output. Such sources of noise can, in principle anyway, be eliminated. Noise also occurs when electrical connections that carry current are made and broken, as, for example, at the brushes of certain types of motors. Again in principle, this type of noise can be suppressed at the source.

Natural phenomena that give rise to noise include electric storms, solar flares, and certain belts of radiation that exist in space. Noise arising from these sources may be more difficult to suppress, and often the only solution is to reposition the receiving antenna to minimize the received noise, while ensuring that reception of the desired signal is not seriously impaired.

Noise is mainly of concern in receiving systems, where it sets a lower limit on the size of signal that can be usefully received. Even when precautions are taken to eliminate noise from faulty connections or that arising from external sources, it is found that certain fundamental sources of noise are present within electronic equipment that limit the receiver sensitivity. One might think that any signal, however small, could simply be amplified up to any desired level. Unfortunately, adding amplifiers to a receiving system also adds noise, and the signal-to-noise ratio, which is the significant quantity, may be degraded by the addition of the amplifiers. Thus the study of the fundamental sources of noise within equipment is essential if the effects of the noise are to be minimized.

4.2 Thermal Noise

It is known that the free electrons within an electrical conductor possess kinetic energy as a result of heat exchange between the conductor and its surroundings. The kinetic energy means that the electrons are in

motion, and this motion in turn is randomized through collisions with imperfections in the structure of the conductor. This process occurs in all real conductors and is what gives rise to the conductors' resistance. As a result, the electron density throughout the conductor varies randomly, giving rise to a randomly varying voltage across the ends of the conductor (Fig. 4.2.1). Such a voltage may sometimes be observed in the flickerings of a very sensitive voltmeter. Since the noise arises from thermal causes, it is referred to as *thermal noise* (and also as *Johnson noise*, after its discoverer).

The average or mean noise voltage across the conductor is zero, but the root-mean-square value is finite and can be measured. (It will be recalled that a similar situation occurs for sinusoidal voltage, which has a mean value of zero and a finite rms value.) It is found that the mean-square value of the noise voltage is proportional to the resistance of the conductor, to its absolute temperature, and to the frequency bandwidth of the device measuring (or responding to) the noise. The rms voltage is of course the square root of the mean-square value.

Consider a conductor that has resistance R , across which a true rms measuring voltmeter is connected, and let the voltmeter have an ideal band-pass frequency response of bandwidth B_n as shown in Fig. 4.2.2. The subscript n signifies noise bandwidth, which for the moment may be assumed to be the same as the bandwidth

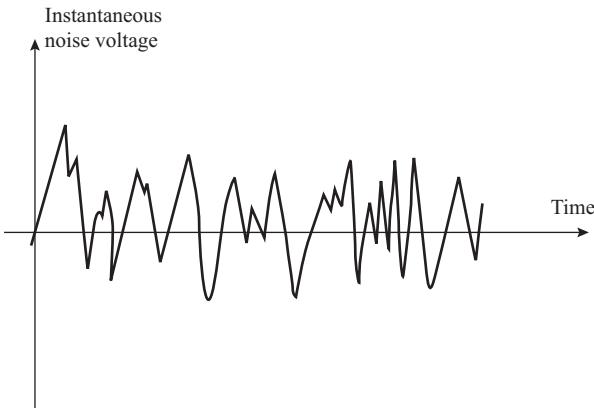


Figure 4.2.1 Thermal noise voltage.

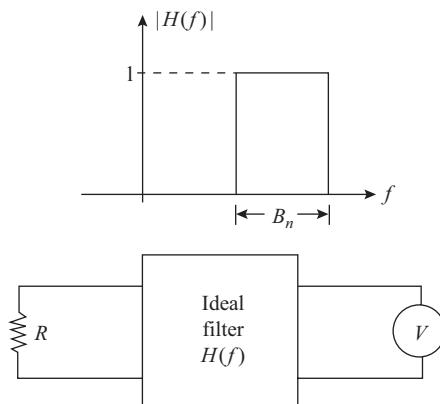


Figure 4.2.2 Measurement of thermal noise.

of the ideal filter. The relationship between noise bandwidth and actual frequency response will be developed more fully later. The mean-square voltage measured on the meter is found to be

$$E_n^2 = 4RkTB_n \quad (4.2.1)$$

where E_n = root-mean-square noise voltage, volts

R = resistance of the conductor, ohms

T = conductor temperature, kelvins

B_n = noise bandwidth, hertz

k = Boltzmann's constant

= 1.38×10^{-23} J/K

The equation is given in terms of mean-square voltage rather than root mean square, since this shows the proportionality between the noise power (proportional to E_n^2) and temperature (proportional to kinetic energy).

The rms noise voltage is given by

$$E_n = \sqrt{4RkTB_n} \quad (4.2.2)$$

The presence of the mean-square voltage at the terminals of the resistance R suggests that it may be considered as a generator of electrical noise power. Attractive as the idea may be, thermal noise is not unfortunately a free source of energy. To abstract the noise power, the resistance R would have to be connected to a resistive load, and in thermal equilibrium the load would supply as much energy to R as it receives.

The fact that the noise power cannot be utilized as a free source of energy does not prevent the power being calculated. In analogy with any electrical source, the *available average power* is defined as the maximum average power the source can deliver. For a generator of emf E volts (rms) and internal resistance R , the available power is $E^2/4R$. Applying this to Eq. (4.2.1) gives for the available thermal noise power:

$$P_n = kTB_n \quad (4.2.3)$$

EXAMPLE 4.2.1

Calculate the thermal noise power available from any resistor at room temperature (290 K) for a bandwidth of 1 MHz. Calculate also the corresponding noise voltage, given that $R = 50 \Omega$.

SOLUTION For a 1-MHz bandwidth, the noise power is

$$\begin{aligned} P_n &= 1.38 \times 10^{-23} \times 290 \times 10^6 \\ &= 4 \times 10^{-15} \text{ W} \\ E_n^2 &= 4 \times 50 \times 1.38 \times 10^{-23} \times 290 \\ &= 810^{-13} \\ \therefore E_n &= 0.895 \mu\text{V} \end{aligned}$$

The noise power calculated in Example 4.2.1 may seem to be very small, but it may be of the same order of magnitude as the signal power present. For example, a receiving antenna may typically have an induced signal emf of 1 μV , which is of the same order as the noise voltage.

The thermal noise properties of a resistor R may be represented by the equivalent voltage generator of Fig. 4.2.3(a). This is one of the most useful representations of thermal noise and is widely used in determining

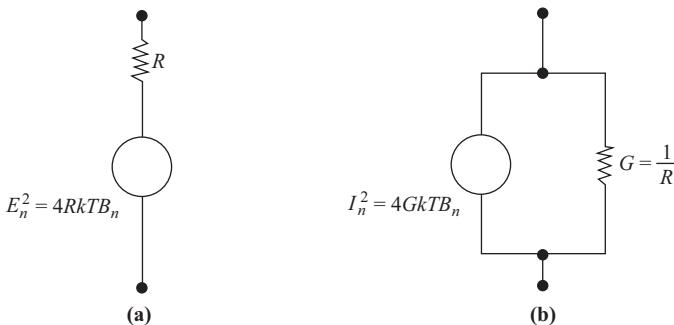


Figure 4.2.3 Equivalent sources for thermal noise: (a) voltage source and (b) current source.

the noise performance of equipment. It is best to work initially in terms of E_n^2 rather than E_n , for reasons that will become apparent shortly.

Norton's theorem may be used to find the equivalent current generator and this is shown in Fig. 4.2.3(b). Here, using conductance $G (= 1/R)$, the rms noise current I_n is given by

$$I_n^2 = 4GkTB_n \quad (4.2.4)$$

It will be recalled that the bandwidth is that of the external circuit, not shown in the source representations, and this must be examined in more detail. Suppose the resistance is left open circuited; then the bandwidth ideally would be infinite, and Eq. (4.2.3) suggests that the open-circuit noise voltage would also be infinite! Two factors prevent this from happening. The first relates to the derivation of the noise energy, which is based on classical thermodynamics and ignores quantum mechanical effects. The quantum mechanical derivation shows that the energy drops off with increasing frequency, and this therefore sets a fundamental limit to the noise power available. However, quantum mechanical effects only become important at frequencies well into the infrared region. The second and more significant practical factor from the circuit point of view is that *all* real circuits contain reactance (for example, self-inductance and self-capacitance), which sets a finite limit on bandwidth. In the case of the open-circuited resistor, the self-capacitance sets a limit on bandwidth, a situation that is covered in more detail later.

Resistors in Series

Let R_{ser} represent the total resistance of the series chain, where $R_{\text{ser}} = R_1 + R_2 + R_3 + \dots$; then the noise voltage of the equivalent series resistance is

$$\begin{aligned} E_n^2 &= 4R_{\text{ser}} kTB_n \\ &= 4(R_1 + R_2 + R_3 + \dots)kTB_n \\ &= E_{n1}^2 + E_{n2}^2 + E_{n3}^2 + \dots \end{aligned} \quad (4.2.5)$$

This shows that the total noise voltage *squared* is obtained by summing the mean-square values. Hence the noise voltage of the series chain is given by

$$E_n = \sqrt{E_{n1}^2 + E_{n2}^2 + E_{n3}^2 + \dots} \quad (4.2.6)$$

Note that simply adding the individual noise voltages would have given the wrong result.

Resistors in Parallel

With resistors in parallel it is best to work in terms of conductance. Thus let G_{par} represent the parallel combination where $G_{\text{par}} = G_1 + G_2 + G_3 + \dots$; then

$$\begin{aligned} I_n^2 &= 4G_{\text{par}} kTB_n \\ &= 4(G_1 + G_2 + G_3 + \dots)kTB_n \\ &= I_{n1}^2 + I_{n2}^2 + I_{n3}^2 + \dots \end{aligned} \quad (4.2.7)$$

Again, it is to be noted that the mean-square values are added to obtain the total mean-square noise current. Usually, it is more convenient to work in terms of noise voltage rather than current. This is most easily done by first determining the equivalent parallel resistance from $1/R_{\text{par}} = 1/R_1 + 1/R_2 + 1/R_3 + \dots$ and using

$$E_n^2 = 4R_{\text{par}} kTB_n \quad (4.2.8)$$

EXAMPLE 4.2.2

Two resistors of 20 and 50 kΩ are at room temperature (290 K). For a bandwidth of 100 kHz, calculate the thermal noise voltage generated by (a) each resistor, (b) the two resistors in series, and (c) the two resistors in parallel.

SOLUTION (a) For the 20-kΩ resistor

$$\begin{aligned} E_n^2 &= 4 \times (20 \times 10^3) \times (4 \times 10^{-21}) \times (100 \times 10^3) \\ &= 32 \times 10^{-12} \text{ V}^2 \\ \therefore E_n &= 5.66 \mu\text{V} \end{aligned}$$

The voltage for the 50-kΩ resistor may be found by simple proportion:

$$\begin{aligned} E_n &= 5.66 \times \sqrt{\frac{50}{20}} \\ &= 8.95 \mu\text{V} \end{aligned}$$

(b) For the series combination, $R_{\text{ser}} = 20 + 50 = 70$ kΩ. Hence

$$\begin{aligned} E_n &= 5.66 \times \sqrt{\frac{70}{20}} \\ &= 10.59 \mu\text{V} \end{aligned}$$

(c) For the parallel combination, $R_{\text{par}} = \frac{20 \times 50}{20 + 50} = 14.29$ kΩ.

$$\therefore E_n = 5.66 \times \sqrt{\frac{14.29}{20}} = 4.78 \mu\text{V}$$

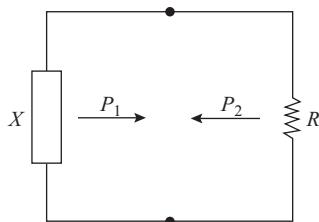


Figure 4.2.4 Power exchange between a reactance and a resistance is \$P_1 = P_2 = 0\$.

Reactance

Reactances do not generate thermal noise. This follows from the fact that reactance cannot dissipate power. Consider an inductive or capacitive reactance connected in parallel with a resistor R (Fig. 4.2.4). In thermal equilibrium, equal amounts of power must be exchanged; that is, if the resistor supplies thermal noise power P_2 to the reactance, the reactance must supply thermal noise power $P_1 = P_2$ to the resistor. But since the reactance cannot dissipate power, the power P_2 must be zero, and hence P_1 must also be zero.

The effect of reactance on the noise bandwidth must, however, be taken into account, as shown in the next section.

Spectral Densities

Thermal noise falls into the category of power signals as described in Section 2.17, and hence it has a spectral density. As pointed out previously, the bandwidth B_n is a property of the external measuring or receiving system and is assumed flat so that, from Eq. (4.2.3), the available power spectral density, in watts per hertz, or joules, is

$$\begin{aligned} G_a(f) &= \frac{P_n}{B_n} \\ &= kT \end{aligned} \quad (4.2.9)$$

The spectral density for the mean-square voltage is also a useful function. This has units of volts² per hertz and is given by

$$\begin{aligned} G_v(f) &= \frac{E_n^2}{B_n} \\ &= 4RkT \end{aligned} \quad (4.2.10)$$

The spectral densities are flat, that is, independent of frequency, as shown in Fig. 4.2.5, and as a result thermal noise is sometimes referred to as *white noise*, in analogy to white light, which has a flat spectrum. When white noise is passed through a network, the spectral density will be altered by the shape of the network frequency response. The total noise power at the output is found by summing the noise contributions over the complete frequency range, taking into account the shape of the frequency response.

Consider a power spectral response as shown in Fig. 4.2.6. At frequency f_1 , the available noise power for an infinitesimally small bandwidth δf about f_1 is $\delta P_{n1} = S_p(f_1)\delta f$. This is so because the bandwidth δf may

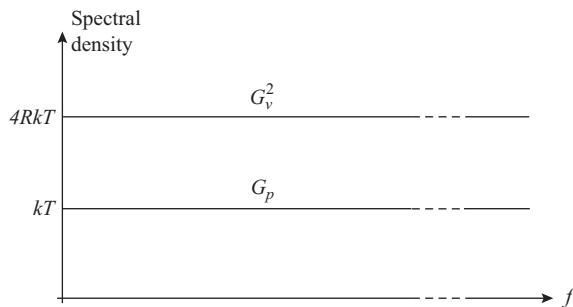


Figure 4.2.5 Thermal noise spectral densities.

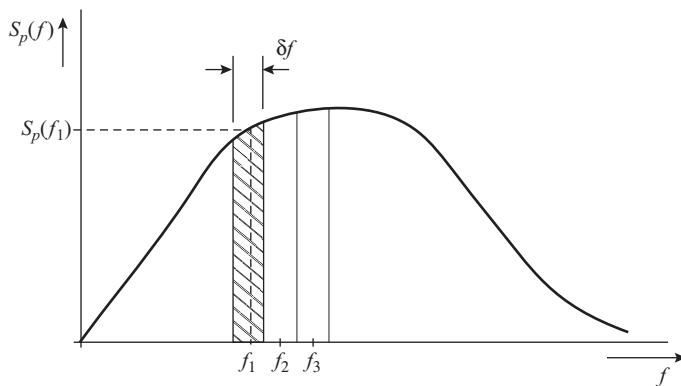


Figure 4.2.6 Nonuniform noise spectral density.

be assumed flat about f_1 , and the available power is given as the product of spectral density (watts/hertz) \times bandwidth (hertz). The available noise power is therefore seen to be equal to the area of the shaded strip about f_1 . Similar arguments can be applied at frequencies f_2, f_3, \dots , and the total power, given by the sum of all these contributions, is equal to the sum of all these small areas, which is the total area under the curve. More formally, this is equal to the integral of the spectral density function over the frequency range $f = 0$ to $f = \infty$.

A similar argument can be applied to mean-square voltage. The spectral density curve in this case has units of V^2/Hz , and multiplying this by bandwidth δf Hz results in units of V^2 , so the area under the curve gives the total mean-square voltage.

Equivalent Noise Bandwidth

Suppose that a resistor R is connected to the input of an LC filter, as shown in Fig. 4.2.7(a). This represents an input generator of mean-square voltage spectral density $4RkT$ feeding a network consisting of R and the LC filter. Let the transfer function of the network including R be $H(f)$, as shown in Fig. 4.2.7(b). The spectral density for the mean-square output voltage is therefore $4RkT|H(f)|^2$. This follows since $H(f)$ is the ratio of output to input voltage, and here mean-square values are being considered.

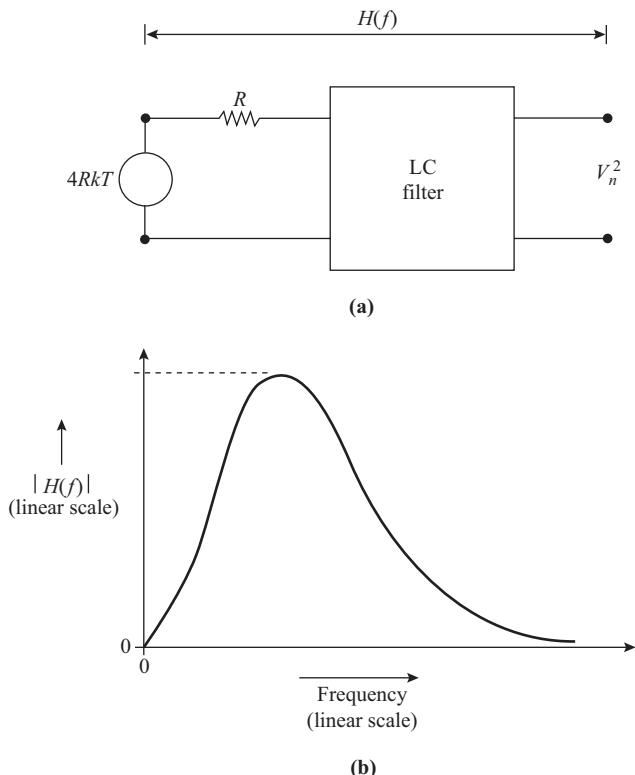


Figure 4.2.7 (a) Filtered noise and (b) the transfer function of the filter including R .

From what was shown previously, the total mean-square output voltage is given by the area under the output spectral density curve

$$\begin{aligned} V_n^2 &= \int_0^\infty 4RkT |H(f)|^2 df \\ &= 4RkT \times (\text{area under } |H(f)|^2 \text{ curve}) \end{aligned} \quad (4.2.11)$$

Now the total mean-square voltage at the output can be stated as $V_n^2 = 4RkTB_n$, and equating this with Eq. (4.2.11) gives, for the equivalent noise bandwidth of the network,

$$\begin{aligned} B_n &= \int_0^\infty |H(f)|^2 df \\ &= (\text{area under } |H(f)|^2 \text{ curve}) \end{aligned} \quad (4.2.12)$$

As a simple example consider the circuit of Fig. 4.2.8, which consists of a resistor in parallel with a capacitor. The capacitor may in fact be the self-capacitance of the resistor, or an external capacitor, for example, the input capacitance of the voltmeter used to measure the noise voltage across R .

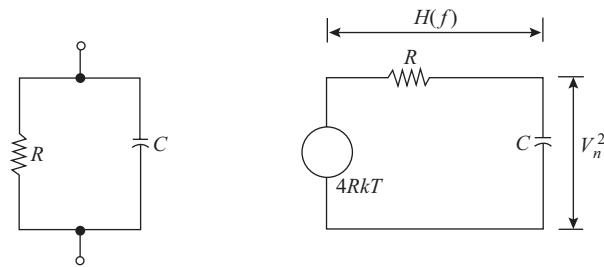


Figure 4.2.8 *RC* network and its transfer function used in determining noise bandwidth.

The transfer function of the *RC* network is

$$|H(f)| = \frac{1}{\sqrt{1 + (\omega CR)^2}} \quad (4.2.13)$$

The equivalent noise bandwidth of the *RC* network is found using Eq. (4.2.12) as

$$\begin{aligned} B_n &= \int_0^\infty |H(f)|^2 df \\ &= \frac{1}{4RC} \end{aligned} \quad (4.2.14)$$

(Details of the integration are left as an exercise for the reader.) The mean-square output voltage is given by

$$\begin{aligned} V_n^2 &= 4RkT \times \frac{1}{4RC} \\ &= \frac{kT}{C} \end{aligned} \quad (4.2.15)$$

This is a surprising result. It shows that the mean-square output voltage is independent of R , even though it originates from R , and it is inversely proportional to C , even though C does not generate noise.

A second example is that of the tuned circuit shown in Fig. 4.2.9. Here the capacitor is assumed lossless, and the inductor has a series resistance r that generates thermal noise.

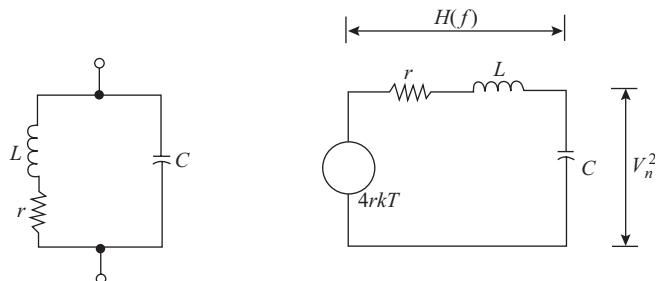


Figure 4.2.9 Tuned circuit and its transfer function used in determining noise bandwidth.

The transfer function in this case is

$$|H(f)| = \left| \frac{X_c}{Z_s} \right| \quad (4.2.16)$$

where $Z_s = r(1 + jyQ)$ is the impedance of the series tuned circuit as given by Eq. (1.3.10) and $X_c = 1/j\omega C$ is the reactance of C . As before, the equivalent noise bandwidth is found by solving Eq. (4.2.12).

Consider first the situation where the circuit is resonant at f_0 , and the noise is restricted to a small bandwidth $\Delta f \ll f_0$ about the resonant frequency. The transfer function is then approximated by $|H(f)| \approx 1/\omega_0 Cr = Q$, and the area under the $|H(f)|^2$ curve over a small constant bandwidth δf is $Q^2 \delta f$. Hence the mean-square noise voltage is

$$\begin{aligned} V_n^2 &= 4rkTB_n \\ &= 4rQ^2kT\delta f \\ &= 4R_D kT\Delta f \end{aligned} \quad (4.2.17)$$

Here, use is made of the relationship $Q^2 r = R_D$ developed in Section 1.4. This is an important result, because the bandwidth is often limited in practice to some small percentage about f_0 . An example will illustrate this.

EXAMPLE 4.2.3

The parallel tuned circuit at the input of a radio receiver is tuned to resonate at 120 MHz by a capacitance of 25 pF. The Q -factor of the circuit is 30. The channel bandwidth of the receiver is limited to 10 kHz by the audio sections. Calculate the effective noise voltage appearing at the input at room temperature.

SOLUTION

$$\begin{aligned} R_D &= \frac{Q}{\omega_o C} \\ &= \frac{30}{2 \times \pi \times 120 \times 10^6 \times 25 \times 10^{-12}} \\ &= 1.59 \text{ k}\Omega \\ \therefore V_n &= \sqrt{4 \times 1.59 \times 10^3 \times 4 \times 10^{-21} \times 10^4} \\ &= 0.5 \mu\text{V} \end{aligned}$$

Where the complete frequency range 0 to ∞ has to be taken into account, the integral becomes much more difficult to solve, and only the result will be given here. This is

$$B_n = \frac{1}{4R_DC} \quad (4.2.18)$$

where R_D is the dynamic resistance of the tuned circuit.

The noise bandwidth can be expressed as a function of the -3 -dB bandwidth of the circuit. From Eq. (1.3.17), $B_{3 \text{ dB}} = f_0/Q$ and from Eq. (1.4.4) $R_D = Q/\omega_o C$. Combining these expressions along with that for the noise bandwidth gives

$$B_n = \frac{\pi}{2} B_{3 \text{ dB}} \quad (4.2.19)$$

By postulating that the noise originates from a resistor R_D and is limited by the bandwidth B_n , the mean-square voltage at the output can be expressed as

$$\begin{aligned} V_n^2 &= 4R_D kT \times \frac{1}{4R_D C} \\ &= \frac{kT}{C} \end{aligned} \quad (4.2.20)$$

In the foregoing, to simplify the analysis it was assumed that the Q -factor remained constant, independent of frequency. This certainly would not be true for the range zero to infinity, but the end result still gives a good indication of the noise expected in practice.

For most radio receivers the noise is generated at the front end (antenna input) of the receiver, while the output noise bandwidth is determined by the audio sections of the receiver. The equivalent noise bandwidth is equal to the area under the normalized power-gain/frequency curve for the low-frequency sections. By normalized is meant that the curve is scaled such that the maximum value is equal to unity. Usually this information is available in the form of a frequency response curve showing output in decibels relative to maximum and with frequency plotted on a logarithmic scale, as sketched in Fig. 4.2.10(a). Before determining the area under the curve, the decibel axis must be converted to a linear power-ratio scale and the frequency axis to a linear frequency scale, as shown in Fig. 4.2.10(b). The equivalent noise bandwidth is then equal to the area under this curve for a single-sideband receiver. Where the receiver is of the double-sideband type, then the noise bandwidth appears on both sides of the carrier and is effectively doubled. This is shown in Fig. 4.2.10(c).

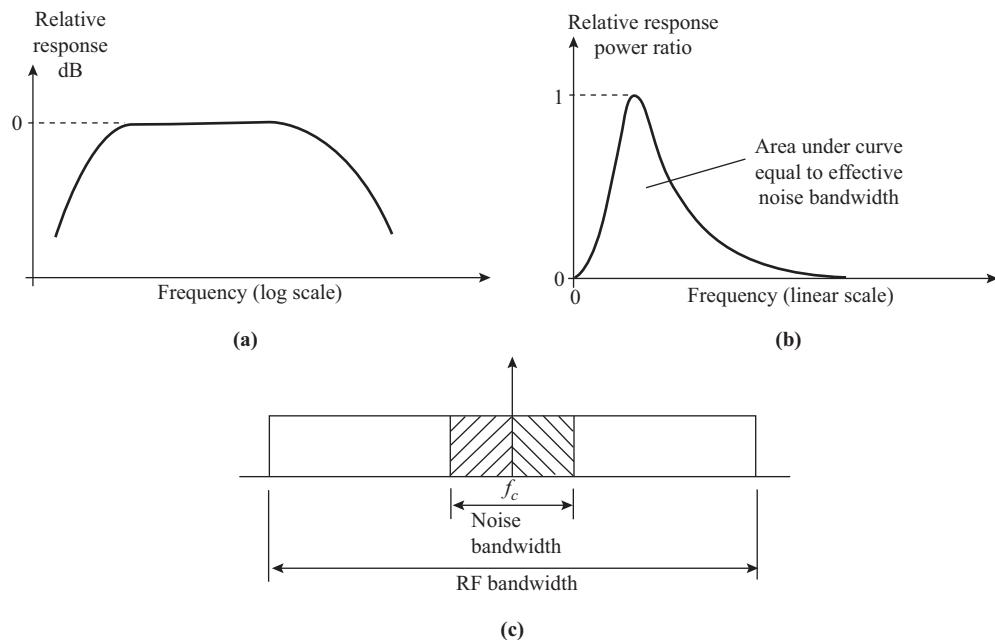


Figure 4.2.10 (a) Amplifier frequency response curve. (b) Curve of (a) using linear scales. (c) Noise bandwidth of a double-sideband receiver.

4.3 Shot Noise

Shot noise is a random fluctuation that accompanies any direct current crossing a potential barrier. The effect occurs because the carriers (holes and electrons in semiconductors) do not cross the barrier simultaneously, but rather with a random distribution in the timing for each carrier, which gives rise to a random component of current superimposed on the steady current. In the case of bipolar junction transistors, the bias current crossing the forward biased emitter–base junction carries shot noise. With vacuum tubes the electrons emitted from the cathode have to overcome a potential barrier that exists between cathode and vacuum. The name *shot noise* was first coined in connection with tubes, where the analogy was made between the electrons striking the plate and lead shot from a gun striking a target.

Although it is always present, shot noise is not normally observed during measurement of direct current because it is small compared to the dc value; however, it does contribute significantly to the noise in amplifier circuits. The idea of shot noise is illustrated in Fig. 4.3.1.

Shot noise is similar to thermal noise in that its spectrum is flat (except in the high microwave frequency range). The mean-square noise component is proportional to the dc flowing, and for most devices the mean-square, shot-noise current is given by

$$I_n^2 = 2I_{dc}q_eB_n \text{ amperes}^2 \quad (4.3.1)$$

where I_{dc} is the direct current in amperes, q_e the magnitude of electron charge ($= 1.6 \times 10^{-19}\text{C}$), and B_n is the equivalent noise bandwidth in hertz.

EXAMPLE 4.3.1

Calculate the shot noise component of current present on a direct current of 1 mA flowing across a semiconductor junction, given that the effective noise bandwidth is 1 MHz.

SOLUTION

$$\begin{aligned} I_n^2 &= 2 \times 10^{-3} \times 1.6 \times 10^{-19} \times 10^6 \\ &= 3.2 \times 10^{-16} \text{ A}^2 \\ \therefore I_n &= 18 \text{ nA} \end{aligned}$$

4.4 Partition Noise

Partition noise occurs wherever current has to divide between two or more electrodes and results from the random fluctuations in the division. It would be expected therefore that a diode would be less noisy than a transistor (other factors being equal) if the third electrode draws current (such as base or gate current). It is for this reason that the input stage of microwave receivers is often a diode circuit, although, more recently, gallium arsenide field-effect transistors, which draw zero gate current, have been developed for low-noise microwave amplification. The spectrum for partition noise is flat.

4.5 Low Frequency or Flicker Noise

Below frequencies of a few kilohertz, a component of noise appears, the spectral density of which increases as the frequency decreases. This is known as *flicker noise* (and sometimes as $1/f$ noise). In vacuum tubes it arises from slow changes in the oxide structure of oxide-coated cathodes and from the migration of impurity ions

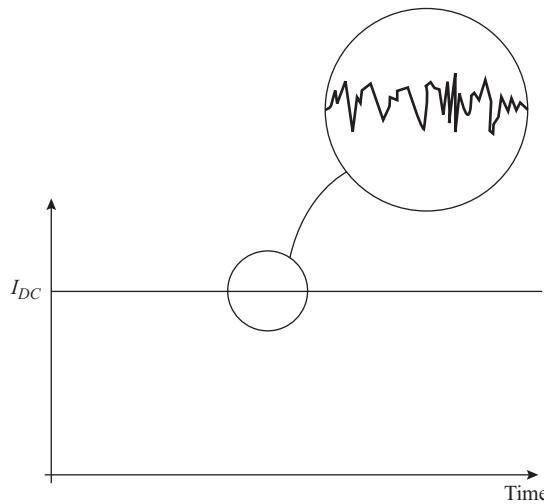


Figure 4.3.1 Shot noise.

through the oxide. In semiconductors, flicker noise arises from fluctuations in the carrier densities (holes and electrons), which in turn give rise to fluctuations in the conductivity of the material. It follows therefore that a noise voltage will be developed whenever direct current flows through the semiconductor, and the mean-square voltage will be proportional to the square of the direct current. Interestingly enough, although flicker noise is a low-frequency effect, it plays an important part in limiting the sensitivity of microwave diode mixers used for Doppler radar systems. This is because, although the input frequencies to the mixer are in the microwave range, the Doppler frequency output is in the low (audio-frequency) range, where flicker noise is significant.

4.6 Burst Noise

Another type of low-frequency noise observed in bipolar transistors is known as *burst noise*, the name arising because the noise appears as a series of bursts at two or more levels (rather like noisy pulses). When present in an audio system, the noise produces popping sounds, and for this reason is also known as “popcorn” noise. The source of burst noise is not clearly understood at present, but the spectral density is known to increase as the frequency decreases.

4.7 Avalanche Noise

The reverse-bias characteristics of a diode exhibit a region where the reverse current, normally very small, increases extremely rapidly with a slight increase in the magnitude of the reverse-bias voltage. This is known as the *avalanche region* and comes about because the holes and electrons in the diode’s depletion region gain sufficient energy from the reverse-bias field to ionize atoms by collision. The ionizing process means that additional holes and electrons are produced, which in turn contribute to the ionization process, and thus the descriptive term *avalanche*.

The collisions that result in the avalanching occur at random, with the result that large noise spikes are present in the avalanche current. In diodes such as zener diodes, which are used as voltage reference sources, the avalanche noise is a nuisance to be avoided. However, avalanche noise is put to good use in noise measurements, as described in Section 4.19. The spectral density of avalanche noise is flat.

4.8 Bipolar Transistor Noise

Bipolar transistors exhibit all the sources of noise discussed previously, that is, thermal, shot, partition, flicker, and burst noise. The thermal noise is generated by the bulk or extrinsic resistances of the electrodes, but the only significant component is that generated by the extrinsic base resistance. It should be emphasised at this point that the small-signal equivalent resistances for the base-emitter and the base-collector junctions do not generate thermal noise, but they do enter into the noise calculations made using the small-signal equivalent circuit for the transistor.

The bias currents in the transistor show shot noise and partition noise, and, in addition, the flicker and burst noise components are usually associated with the base current.

4.9 Field-effect Transistor Noise

In field-effect transistors (both JFETs and MOSFETs), the main source of noise is the thermal noise generated by the physical resistance of the drain-source channel. Flicker noise also originates in this channel. Additionally, there will be shot noise associated with the gate leakage current. This will develop a noise component of voltage across the signal-source impedance and is only significant where this impedance is very high (in the megohm range).

4.10 Equivalent Input Noise Generators and Comparison of BJTs and FETs

An amplifier may be represented by the block schematic of Fig. 4.10.1(a), in which a noisy amplifier is shown and where the source and load resistances generate thermal noise. The circuit may be redrawn as shown in

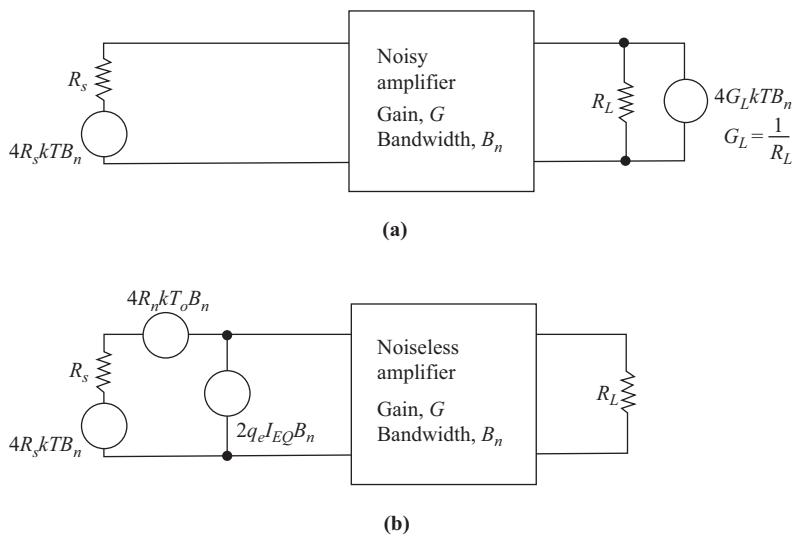


Figure 4.10.1 (a) Noisy amplifier and (b) the equivalent input noise generators.

Fig. 4.10.1(b) in which the amplifier itself is considered to be noiseless, the amplifier noise being represented by *fictitious noise generators* $V_{na} = \sqrt{4R_n kT_o B_n}$ and $I_{na} = \sqrt{2q_e I_{EQ} B_n}$ at the input. Here, B_n is the equivalent noise bandwidth of the amplifier in hertz, T_o is room temperature in kelvins, k is Boltzmann's constant $= 1.38 \times 10^{-23}$ J/K, and $q_e = 1.6 \times 10^{-19}$ C is the magnitude of the electron charge. These terms have all been defined previously. What is new here is the *fictitious resistance* R_n ohms, known as the *equivalent input noise resistance* of the amplifier, and I_{EQ} amperes, the *equivalent input shot noise current*. Both these parameters have to be calculated or specified for a transistor under given operating conditions.

The noise generated by the load resistance R_L is generally very small compared to the other sources and is assumed to be negligible, so this is dropped from the equivalent circuit. The thermal noise generated by the signal-source resistance R_s is generally significant and must be taken into account as shown in Fig. 4.10.1(b).

The total noise voltage at the input to the amplifier is found as follows. Referring to the equivalent circuit of Fig. 4.10.2(a), the noise sources are

$$V_{ns}^2 = 4R_s kT_o B_n \quad (4.10.1)$$

$$V_{na}^2 = 4R_n kT_o B_n \quad (4.10.2)$$

$$I_{na}^2 = 2q_e I_{EQ} B_n \quad (4.10.3)$$

As a first step in simplifying this, the emf sources can be combined as shown in Fig. 4.10.2(b). Next, the Thevenin equivalent circuit can be obtained by combining the current source as an equivalent emf as shown in Fig. 4.10.2(c). Thus the equivalent noise voltage at the input to the amplifier is

$$V_n = \sqrt{V_{ns}^2 + V_{na}^2 + (I_{na} R_s)^2} \quad (4.10.4)$$

Throughout, it is assumed that all the noise sources are uncorrelated. Correlation actually exists between V_{na} and I_{na} , but this is only significant at high frequencies, where the analysis must take correlation into account.

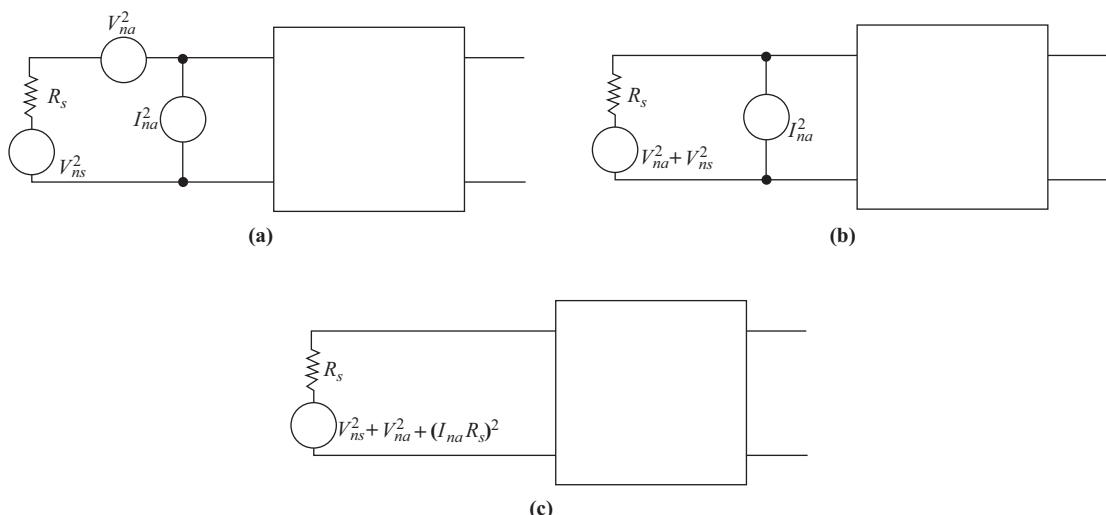


Figure 4.10.2 (a) Equivalent input noise generators; (b) the voltage sources combined; (c) all sources combined.

A detailed comparison of the performance of BJT and FET amplifiers is too involved to be included here, but the following general remarks may be made. R_n is generally smaller and I_{EQ} larger for BJTs compared to FETs. For input signal sources with low resistances, where the noise voltage $I_{na}R_s$ is small enough to be neglected, the BJT will produce lower noise because of its smaller value of R_n . Where, however, R_s is large such that the $I_{na}R_s$ voltage is significant, the FET will produce lower noise than the BJT because of its lower I_{EQ} . There will be an intermediate range for R_s where, in fact, the thermal noise generated by R_s itself dominates, and the type of transistor may have little bearing on the overall noise performance of the amplifier.

4.11 Signal-to-Noise Ratio

In a communications link it is the signal-to-noise ratio, rather than the absolute value of noise, that is important. Signal-to-noise ratio is defined as a power ratio, and since at a given point in a circuit power it is proportional to the square of the voltage, then

$$\begin{aligned}\frac{S}{N} &= \frac{P_s}{P_n} \\ &= \frac{V_s^2}{V_n^2}\end{aligned}\quad (4.11.1)$$

EXAMPLE 4.11.1

The equivalent noise resistance for an amplifier is $300\ \Omega$, and the equivalent shot noise current is $5\ \mu A$. The amplifier is fed from a $150\text{-}\Omega$, $10\text{-}\mu V$ rms sinusoidal signal source. Calculate the individual noise voltages at the input and the input signal-to-noise ratio in decibels. The noise bandwidth is 10 MHz .

SOLUTION Assume room temperature so that $kT = 4 \times 10^{-21}\ J$ and $q_e = 1.6 \times 10^{-19}\ C$. The shot noise current is $I_{na} = \sqrt{2q_e I_{EQ} B_n} = 4\ nA$. The noise voltage developed by this across the source resistance is $I_{na}R_s = 0.6\ \mu V$.

Note that the shot noise current *does not* develop a voltage across R_n . The noise voltage generated by R_n is $V_{na} = \sqrt{4R_n kT_o B_n} = 6.93\ \mu V$. The thermal noise voltage from the source is

$$V_{ns} = \sqrt{4R_s kT_o B_n} = 4.9\ \mu V$$

The total noise voltage at the input to the amplifier is

$$V_n = \sqrt{4.9^2 + 6.93^2 + .6^2} = 8.51\ \mu V$$

The signal-to-noise ratio in decibels is

$$\frac{S}{N} = 20 \log \frac{V_s}{V_n} = 1.4\ dB$$

4.12 S/N Ratio of a Tandem Connection

In an analog telephone system it is usually necessary to insert amplifiers to make up for the loss in the telephone cables, the amplifiers being known as repeaters. As shown in Fig. 4.12.1, if the power loss of a line section is L , then the amplifier power gain G is chosen so that $LG = 1$. A long line will be divided into

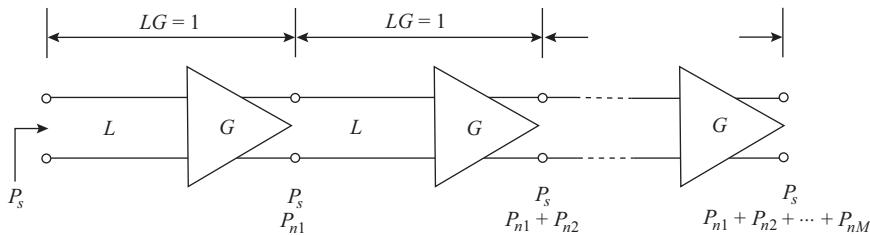


Figure 4.12.1 Tandem connection of repeaters.

sections that are near enough identical, and each repeater adds its own noise, so the noise accumulates with the signal as it travels along the system.

Consider the situation where the input signal power to the first section of the line is P_s , and at this point the input noise may be assumed negligible. After traveling along the first section of line, the signal is attenuated by a factor L . At the output of the first repeater the signal power is again P_s since the gain G exactly compensates for the loss L . The noise at the output of the first repeater is shown as P_{nl} and consists of the noise added by the line section and amplifier, or what is termed the *first link* in the system.

As the signal progresses along the links, the power output at each repeater remains at P_s because $LG = 1$ for each link. However, the noise powers are additive, and the total noise at the output of the M th link is $P_n = P_{nl} + P_{n2} + \dots + P_{nM}$. If the links are identical such that each link contributes P_n , the total noise power becomes $P_{nM} = MP_n$. The output signal-to-noise ratio in this case is

$$\begin{aligned} \left(\frac{N}{S}\right)_o \text{ dB} &= 10 \log \frac{P_s}{MP_n} \\ &= \left(\frac{S}{N}\right)_1 \text{ dB} - (M) \text{ dB} \end{aligned} \quad (4.12.1)$$

where (S/N) is the signal-to-noise ratio of any one link, and (M) dB is the number of links expressed as a power ratio in decibels (that is, in decilogs).

EXAMPLE 4.12.1

Calculate the output signal-to-noise ratio in decibels for three identical links, given that the signal-to-noise ratio for any one link is 60 dB.

SOLUTION $(S/N)_o = 60 - 10 \log 3 = 55.23 \text{ dB}$

If the S/N ratio of any one link is much worse than the others, that link will determine the overall S/N ratio. Suppose for example that the S/N ratio of the first link is much lower than the others; then the $(N/S)_1$ ratio will be much greater than the other noise-to-signal ratios. Hence

$$\begin{aligned} \left(\frac{N}{S}\right)_o &= \frac{P_{n1}}{P_s} + \frac{P_{n2}}{P_s} + \dots \\ &= \left(\frac{N}{S}\right)_1 + \left(\frac{N}{S}\right)_2 + \dots \\ &\approx \left(\frac{N}{S}\right)_1 \end{aligned} \quad (4.12.2)$$

EXAMPLE 4.12.2

Calculate the output signal-to-noise ratio in decibels for three links, the first two of which have S/N ratios of 60 dB and the third an S/N of 40 dB.

SOLUTION The noise-to-power ratio of the first two links is -60 dB, or a power ratio of 10^{-6} , while that of the third link is -40 dB, or a power ratio of 10^{-4} . The overall noise-to-signal ratio is

$$\begin{aligned}\left(\frac{S}{N}\right)_o &= 10^{-6} + 10^{-6} + 10^{-4} \\ &\approx 10^{-4}\end{aligned}$$

Thus the output signal-to-noise is approximately **40 dB**.

This example shows that the S/N ratio is approximately equal to that of the worst link, and the old saying that “a chain is no stronger than its weakest link” applies here also!

4.13 Noise Factor

Consider a signal source at room temperature $T_o = 290$ K providing an input to an amplifier. As explained in Section 4.2, the available noise power from such a source would be $P_{ni} = kT_oB_n$. Let the available signal power from the source be denoted by P_{si} ; then the available signal-to-noise ratio from the source is

$$\left(\frac{S}{N}\right)_{in} = \frac{P_{si}}{kT_oB_n} \quad (4.13.1)$$

With the source connected to an amplifier, this represents the available input signal-to-noise ratio and hence the use of the subscript *in*. If now the amplifier has an available power gain denoted by G , the available output signal power would be $P_{so} = GP_{si}$, and if the amplifier was entirely noiseless, the available output noise power would be $P_{no} = GkT_oB_n$, as shown in Fig. 4.13.1(a). Hence the available output signal-to-noise ratio would be the same as that at the input since the factor G would cancel for both signal and noise.

However, it is known that all real amplifiers contribute noise, and the available output signal-to-noise ratio will be less than that at the input. The noise factor F is defined as

$$F = \frac{\text{available S/N power ratio at the input}}{\text{available S/N power ratio at the output}} \quad (4.13.2)$$

In terms of the symbols, this can be written as

$$\begin{aligned}F &= \frac{P_{si}}{kT_oB_n} \times \frac{P_{no}}{GP_{si}} \\ &= \frac{P_{no}}{GkT_oB_n}\end{aligned} \quad (4.13.3)$$

It follows from this that the available output noise power is given by

$$P_{no} = FGkT_oB_n \quad (4.13.4)$$

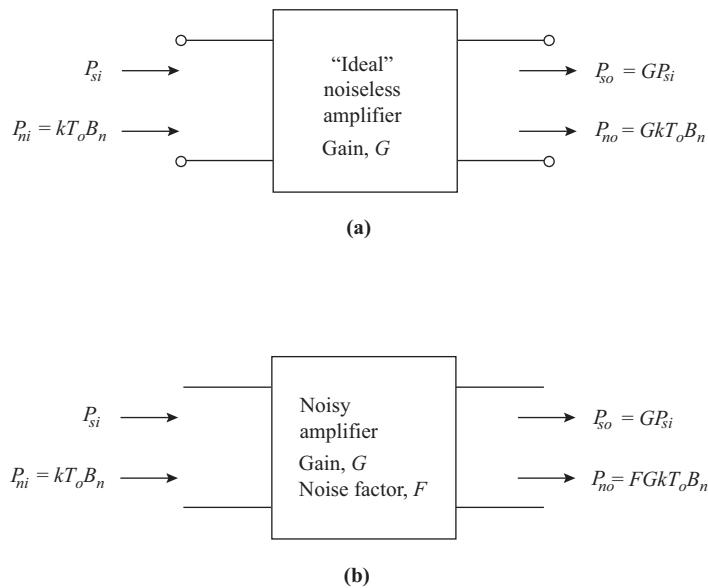


Figure 4.13.1 Noise factor F .

This is shown in Fig. 4.13.1(b). F can be interpreted as the factor by which the amplifier increases the output noise, for, if the amplifier were noiseless, the output noise would be GkT_oB_n .

A few comments are in order here regarding the definitions. Available power gain G is used because it can be defined unambiguously; that is, it does not depend on the load impedance. It may be thought that this definition requires the input to be matched for maximum power transfer, but this is not so. The available output power depends on the actual input power delivered to the amplifier and hence takes into account any input mismatch that may be present. It must also be noted that noise factor is defined for the source at room temperature $T = 290$ K.

Noise factor is a measured parameter and will usually be specified for a given amplifier or network (the definition given applies for any linear network). It is usually specified in decibels, when it is referred to as the *noise figure*. Thus

$$\text{noise figure} = (F) \text{ dB} = 10 \log F \quad (4.13.5)$$

EXAMPLE 4.13.1

The noise figure of an amplifier is 7 dB. Calculate the output signal-to-noise ratio when the input signal-to-noise ratio is 35 dB.

SOLUTION From the definition of noise factor it follows that

$$\begin{aligned} (\text{S/N})_o &= (\text{S/N})_{in} - (F) \text{ dB} \\ &= 35 - 7 \\ &= \mathbf{28 \text{ dB}} \end{aligned}$$

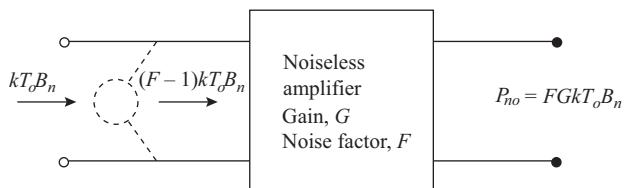


Figure 4.14.1 Equivalent input noise power source for an amplifier.

4.14 Amplifier Input Noise in Terms of F

Amplifier noise is generated in many components throughout the amplifier, but it proves convenient to imagine it to originate from some equivalent power source at the input of the amplifier. (This is somewhat similar to the equivalent input generator approach described in Section 4.10.) From Eq. (4.13.4), the total available input noise is

$$\begin{aligned} P_{ni} &= \frac{P_{no}}{G} \\ &= FkT_oB_n \end{aligned} \quad (4.14.1)$$

This is illustrated in Fig. 4.14.1.

The source contributes an available power kT_oB_n and hence the amplifier must contribute an amount P_{na} , where

$$\begin{aligned} P_{na} &= FkT_oB_n - kT_oB_n \\ &= (F - 1)kT_oB_n \end{aligned} \quad (4.14.2)$$

EXAMPLE 4.14.1

An amplifier has a noise figure of 13 dB. Calculate the equivalent amplifier input noise for a bandwidth of 1 MHz.

SOLUTION 13 dB is a power ratio of approximately 20:1. Hence

$$P_{na} = (20 - 1)4 \times 10^{-21} 10^6 = 1.44 \text{ pW}$$

It will be noted in the example that the noise figure must be converted to a power ratio F to be used in the calculation.

4.15 Noise Factor of Amplifiers in Cascade

Consider first two amplifiers in cascade as shown in Fig. 4.15.1. The problem is to determine the overall noise factor F in terms of the individual noise factors and available power gains.

The available noise power at the output of amplifier 1 is $P_{no1} = F_1 G_1 kT_o B_n$ and this is available to amplifier 2. Amplifier 2 has noise $(F_2 - 1)kT_o B_n$ of its own at its input, and hence the total available noise power at the input of amplifier 2 is

$$P_{ni2} = F_1 G_1 kT_o B_n + (F_2 - 1)kT_o B_n \quad (4.15.1)$$

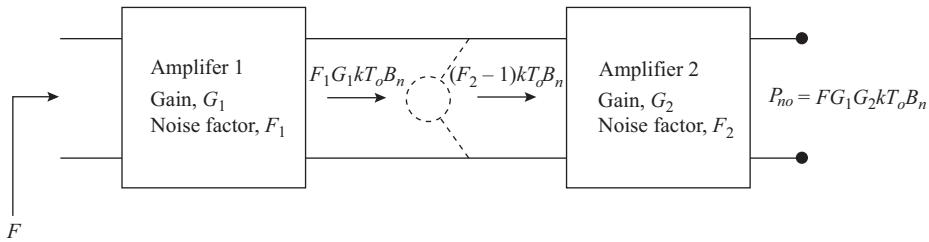


Figure 4.15.1 Noise factor of two amplifiers in cascade.

Now since the noise of amplifier 2 is represented by its equivalent input source, the amplifier itself can be regarded as being noiseless and of available power gain G_2 , so the available noise output of amplifier 2 is

$$\begin{aligned} P_{no2} &= G_2 P_{ni2} \\ &= G_2 (F_1 G_1 k T_o B_n + (F_2 - 1) k T_o B_n) \end{aligned} \quad (4.15.2)$$

The overall available power gain of the two amplifiers in cascade is $G = G_1 G_2$, and let the overall noise factor be F ; then the output noise power can also be expressed as [see Eq. 4.13.4]

$$P_{no} = FGkT_oB_n \quad (4.15.3)$$

Equating the two expressions for output noise and simplifying yields

$$F = F_1 + \frac{F_2 - 1}{G_1} \quad (4.15.4)$$

This equation shows the importance of having a high-gain, low-noise amplifier as the first stage of a cascaded system. By making G_1 large, the noise contribution of the second stage can be made negligible, and F_1 must also be small so that the noise contribution of the first amplifier is low.

The argument is easily extended for additional amplifiers to give

$$F = F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1 G_2} + \dots \quad (4.15.5)$$

This is known as *Friis's formula*.

There are two particular situations where a low-noise, front-end amplifier is employed to reduce noise. One of these is in satellite receiving systems and this is discussed more fully in Chapter 19. The other is in radio receivers used to pick up weak signals, such as short-wave receivers. In most receivers, a stage known as the *mixer stage* is employed to change the frequency of the incoming signal, and it is known that mixer stages have notoriously high noise factors. By inserting an RF amplifier ahead of the mixer, the effect of the mixer noise can be reduced to negligible levels. This is illustrated in the following example.

EXAMPLE 4.15.1

A mixer stage has a noise figure of 20 dB, and this is preceded by an amplifier that has a noise figure of 9 dB and an available power gain of 15 dB. Calculate the overall noise figure referred to the input.

SOLUTION It is first necessary to convert all decibel values to the equivalent power ratios:

$$\begin{aligned}F_2 &= 20 \text{ dB} = 100 : 1 \text{ power ratio} \\F_1 &= 9 \text{ dB} = 7.94 : 1 \text{ power ratio} \\G_1 &= 15 \text{ dB} = 31.62 : 1 \text{ power ratio}\end{aligned}$$

$$\begin{aligned}F &= F_1 + \frac{F_2 - 1}{G_1} \\&= 7.94 + \frac{100 - 1}{31.62} \\&= 11.07\end{aligned}$$

This is the overall noise factor. The overall noise figure is

$$\begin{aligned}(F) \text{ dB} &= 10 \log 11.07 \\&= \mathbf{10.44 \text{ dB}}\end{aligned}$$

4.16 Noise Factor and Equivalent Input Noise Generators

The noise factor is a function of source resistance as well as amplifier input noise. Referring once again to Fig. 4.10.2, the total mean-square input noise voltage is V_n^2 , while the noise from the source alone is V_{ns}^2 . In terms of these quantities, the noise factor is

$$F = \frac{V_n^2}{V_{ns}^2} \quad (4.16.1)$$

Substituting from Eqs. (4.10.1) through (4.10.3) and simplifying gives

$$\begin{aligned}F &= 1 + \frac{R_n}{R_s} + \frac{q_e I_{EQ} \cdot R_s}{2kT_o} \\&= 1 + \frac{R_n}{R_s} + \frac{I_{EQ} R_s}{2V_T} \quad (4.16.2)\end{aligned}$$

Here, $V_T = kT_o/q_e = 26 \text{ mV}$ is a constant.

The second term is inversely proportional to R_s , and the third term is proportional to R_s , which means that there must be an optimum value for R_s that minimizes F . This can be found by differentiating Eq. (4.16.2) and equating to zero. After simplifying, this results in

$$\begin{aligned}R_{s \text{ opt}} &= \frac{V_{na}}{I_{na}} \\&= \sqrt{\frac{2R_n V_T}{I_{EQ}}} \quad (4.16.3)\end{aligned}$$

Thus, knowing the input generator parameters allows the optimum value of source resistance to be determined. An input transformer coupling circuit may be necessary in order to transform the actual source resistance to the optimum value.

4.17 Noise Factor of a Lossy Network

When a signal source is matched through a lossy network, such as a connecting cable, the available signal power at the output of the network is reduced by the insertion loss of the network. The output noise remains unchanged at kT_oB_n (assuming source and network to be at room temperature), since available noise power is independent of source resistance. In effect, the network attenuates the source noise, but at the same time adds noise of its own. The S/N ratio is therefore reduced by the amount that the output power is attenuated.

Denoting the power insertion loss ratio as L , the output S/N ratio will be $1/L$ times the input S/N ratio, and, from the definition of noise factor given by Eq. (4.13.2),

$$\begin{aligned} F &= \frac{\text{available } S/N \text{ power ratio at the input}}{\text{available } S/N \text{ power ratio at the output}} \\ &= L \end{aligned} \quad (4.17.1)$$

In Section 1.3 the insertion loss IL was defined in terms of currents. In terms of power, the power insertion loss is $L = (\text{IL})^2$. Alternatively, specifying the insertion loss in decibels, which apply equally to current and power ratios, also specifies the noise figure in decibels.

EXAMPLE 4.17.1

Calculate the noise factor of an attenuator pad that has an insertion loss of 6 dB.

SOLUTION The insertion loss is 6 dB, and therefore the noise figure is 6 dB. This is equivalent to a noise factor of 4.

The available power gain of a lossy network is $1/L$, and therefore when a lossy network, such as a connecting cable, is placed ahead of an amplifier, Friis's formula gives for the overall noise factor

$$\begin{aligned} F &= F_{nw} + \frac{F_a - 1}{G_{nw}} \\ &= L + (F_a - 1) \cdot L \end{aligned} \quad (4.17.2)$$

The subscript nw refers to the lossy network and a to the amplifier. It will be seen therefore that the loss L adversely affects the overall noise factor in two ways: by its direct contribution and by increasing the effect of the amplifier noise.

Alternatively, if the amplifier is placed ahead of the network, the overall noise factor is

$$\begin{aligned} F &= F_a + \frac{F_{nw} - 1}{G_a} \\ &= F_a + \frac{L - 1}{G_a} \end{aligned} \quad (4.17.3)$$

In this case, provided the amplifier has high gain, the overall noise factor of the system is essentially that of the amplifier alone. This situation is met with in satellite receiving systems (see Problem 4.42 and Chapter 19).

4.18 Noise Temperature

The concept of noise temperature is based on the available noise power equation given in Section 4.2, which is repeated here for convenience:

$$P_n = kT_aB_n \quad (4.18.1)$$

Here, the subscript a has been included to indicate that the noise temperature is associated only with the available noise power. In general, T_a will not be the same as the physical temperature of the noise source. As an example, an antenna pointed at deep space will pick up a small amount of cosmic noise. The equivalent noise temperature of the antenna that represents this noise power may be a few tens of kelvins, well below the physical ambient temperature of the antenna. If the antenna is pointed directly at the sun, the received noise power increases enormously, and the corresponding equivalent noise temperature is well above the ambient temperature.

When the concept is applied to an amplifier, it relates to the equivalent noise of the amplifier referred to the input. If the amplifier noise referred to the input is denoted by P_{na} , the equivalent noise temperature of the amplifier referred to the input is

$$T_e = \frac{P_{na}}{kB_n} \quad (4.18.2)$$

In Section 4.14, it was shown that the equivalent input power for an amplifier is given in terms of its noise factor by $P_{na} = (F - 1)kT_oB_n$. Substituting this in Eq. (4.18.2) gives for the equivalent input noise temperature of the amplifier

$$T_e = (F - 1)T_o \quad (4.18.3)$$

This shows the proportionality between T_e and F , and knowing one automatically entails knowing the other. In practice, it will be found that noise temperature is the better measure for low-noise devices, such as the low-noise amplifiers used in satellite receiving systems, while noise factor is a better measure for the main receiving system.

Friis's formula can be expressed in terms of equivalent noise temperatures. Denoting by T_e the overall noise of the cascaded system referred to the input, and by T_{e1}, T_{e2} , and so on, the noise temperatures of the individual stages, then Friis's formula is easily rearranged to give

$$T_e = T_{e1} + \frac{T_{e2}}{G_1} + \frac{T_{e3}}{G_1 G_2} + \dots \quad (4.18.4)$$

EXAMPLE 4.18.1

A receiver has a noise figure of 12 dB, and it is fed by a low-noise amplifier that has a gain of 50 dB and a noise temperature of 90 K. Calculate the noise temperature of the receiver and the overall noise temperature of the receiving system.

SOLUTION 12 dB represents a power ratio of 15.85:1. Hence

$$T_{em} = (15.85 - 1) \times 290 \cong \mathbf{4306 \text{ K}}$$

The 50-dB gain represents a power ratio of 10^5 :1. Hence

$$T_e = 90 + \frac{4306}{10^5} \cong \mathbf{90 \text{ K}}$$

This example shows the relatively high noise temperature of the receiver, which clearly cannot be its physical temperature! It also shows how the low-noise amplifier controls the noise temperature of the overall receiving system. In this example, the cable connecting the low-noise amplifier and the receiver is assumed to contribute negligible noise. In satellite receiving systems the connecting cable can contribute significantly to the noise, and this is discussed in Chapter 19.

4.19 Measurement of Noise Temperature and Noise Factor

Noise temperature (and noise factor) can be measured in a number of ways, the method selected depending largely on the range of values expected. For normal receiving systems, an avalanche diode noise source is commonly employed, and this method will be described. (Older noise-figure-meters made use of the shot-noise generated by a vacuum tube diode.)

When operated in the avalanche mode, the diode generates a comparatively large amount of noise and can be considered as a source of noise power at some equivalent “hot” temperature T_h . With the reverse bias switched off, the diode reverts to normal noise output and generates noise at some equivalent “cold” temperature T_c . The *excess noise ratio* ENR is defined as

$$\text{ENR (dB)} = 10 \log \frac{T_h - T_c}{T_c} \quad (4.19.1)$$

The cold temperature is normally taken as room temperature $T_c = T_o = 290$ K. The ENR for the source is normally printed on the diode enclosure and is specified by the manufacturer for a range of frequencies. Knowing the ENR and T_c , the hot temperature T_h can be found.

Now let the diode source be matched to the input of the amplifier under test, and let the (unknown) equivalent input noise temperature of the amplifier be denoted by T_e . The amplifier output noise is measured for two conditions, one with the diode in the avalanche mode, denoted by P_h , and one with the reverse bias switched off, denoted by P_c . The two equations for the noise output are

$$P_h = Gk(T_h + T_e)B_n \quad (4.19.2)$$

$$P_c = Gk(T_c + T_e)B_n \quad (4.19.3)$$

where G is the power gain of the amplifier under test. The power ratio, termed the *Y-factor*, is

$$Y = \frac{P_h}{P_c} \quad (4.19.4)$$

Solving these three equations for T_e yields

$$T_e = \frac{T_h - YT_c}{Y - 1} \quad (4.19.5)$$

Note, therefore, that the gain and noise bandwidth do not enter into the final equation; also, a noise power ratio Y is the measured quantity, which does not require an absolute measure of power.

EXAMPLE 4.19.1

In the measurement of noise temperature, an avalanche diode source is used, the ENR being 14 dB. The measured Y factor is 9 dB. Calculate the equivalent noise temperature of the amplifier under test.

SOLUTION The excess noise ratio, as a power ratio, is $\text{ENR} = 10^{1.4} = 25.12$. The Y factor expressed as a power ratio is $Y = 10^{1.9} = 7.94$. From the definition of ENR, the hot temperature is

$$T_h = T_o(\text{ENR} + 1) = 290(25.12 + 1) \cong 7575 \text{ K}$$

Substituting this in Eq. (4.19.5) gives

$$\begin{aligned} T_e &= \frac{7575 - 7.94 \times 290}{7.94 - 1} \\ &\cong 760 \text{ K} \end{aligned}$$

Knowing the equivalent noise temperature, the noise factor can be found. Alternatively, the equations can be rearranged to give

$$F = \frac{\text{ENR}}{Y - 1} \quad (4.19.6)$$

where ENR and Y are expressed as power ratios (not decibels).

In modern noise-measuring instruments, there is provision for making the measurements over a range of frequencies, and microprocessor control allows the noise temperature and noise factor to be automatically calculated and displayed as a function of frequency.

4.20 Narrowband Band-pass Noise

Band-pass filtering of signals arises in many situations, the basic arrangement being shown in Fig. 4.20.1. The filter has an equivalent noise bandwidth B_N (see Section 4.2) and a center frequency f_c . A narrowband system is one in which the center frequency is much greater than the bandwidth, which is the situation to be considered here.

The signal source is shown as a voltage generator of internal resistance R_s . System noise is referred to the input as a thermal noise source at a noise temperature T_s . The available power spectral density is, from Eq. (4.2.9),

$$G_a(f) = kT_s \quad (4.20.1)$$

For the ideal band-pass system shown, the spectral density is not altered by transmission through the filter, but the filter bandwidth determines the available noise power as $kT_s B_N$. So far, this is a result that has already been encountered in general. An alternative description of the output noise, however, turns out to be very useful, especially in connection with the modulation systems described in later chapters. The waveforms of input and output noise voltages are shown in Fig. 4.20.2.

The output waveform has the form of a modulated wave and can be expressed mathematically as

$$n(t) = A_n(t) \cos(\omega_c t + \phi_n(t)) \quad (4.20.2)$$

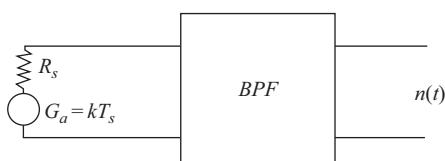


Figure 4.20.1 Noise in a band-pass system.

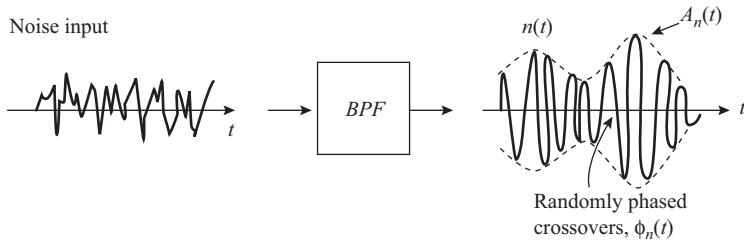


Figure 4.20.2 Input and output noise waveforms for a band-pass system.

This represents the noise in terms of a randomly varying voltage envelope $A_n(t)$ and a random phase angle $\phi_n(t)$. These components are readily identified as part of the waveform, as shown in Fig. 4.20.2, but an equivalent although not so apparent expression can be obtained by trigonometric expansion of the output waveform as

$$n(t) = n_I(t) \cos \omega_c t - n_Q(t) \sin \omega_c t \quad (4.20.3)$$

Here, $n_I(t)$ is a random noise voltage termed the *in-phase component* because it multiplies a cosine term used as a reference phasor, and $n_Q(t)$ is a similar random voltage termed the *quadrature component* because it multiplies a sine term, which is therefore 90° out of phase, or in quadrature with, the reference phasor. The reason for using this form of equation is that, when dealing with modulated signals, the output noise voltage is determined by these two components (this is described in detail in later chapters on modulation). The two noise voltages $n_I(t)$ and $n_Q(t)$ appear to modulate a carrier at frequency f_c and are known as the *low-pass equivalent noise voltages*. The carrier f_c may be chosen anywhere within the passband, but the analysis is simplified by placing it at the center as shown. This is illustrated in Fig. 4.20.3.

A number of important relationships exist between $n_I(t)$ and $n_Q(t)$ and $n(t)$, some of which will be stated here without proof. All three have similar noise characteristics and $n_I(t)$ and $n_Q(t)$ are uncorrected. Of particular importance in later work on modulation is that where the power spectral density of $n(t)$ is $G_a(f) = kT_s$ the power spectral densities for $n_I(t)$ and $n_Q(t)$ are

$$G_I(f) = G_Q(f) = 2kT_s \quad (4.20.4)$$

This important result, which is illustrated in Fig. 4.20.3, will be encountered again in relation to modulated signals.

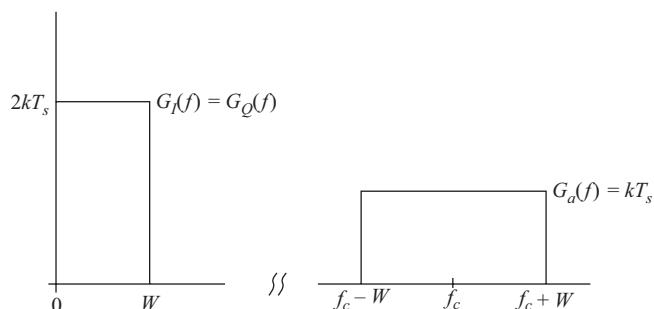


Figure 4.20.3 Noise spectral densities.

PROBLEMS

Assume that $q_e = 1.6 \times 10^{-19}$ C, $k = 1.38 \times 10^{-23}$ J/K and room temperature $T_o = 290$ K applies unless otherwise stated.

- 4.1.** Explain how thermal noise power varies (a) with temperature and (b) with frequency bandwidth. Thermal noise from a resistor is measured as 4×10^{-17} W for a given bandwidth and at a temperature of 20°C. What will the noise power be when the temperature is changed to (c) 50°C; (d) 70 K?
- 4.2.** Given two resistors $R_1 = 10$ kΩ and $R_2 = 15$ kΩ, calculate the thermal noise voltage generated by (a) R_1 , (b) R_2 , (c) R_1 in series with R_2 , and (d) R_1 in parallel with R_2 . Assume a 20-MHz noise bandwidth.
- 4.3.** Three resistors have values $R_1 = 10$ kΩ, $R_2 = 14$ kΩ, and $R_3 = 24$ kΩ. It is known that the thermal noise voltage generated by R_1 is 0.3 μV. Calculate the thermal noise voltage generated by (a) the three resistors connected in series and (b) connected in parallel.
- 4.4.** A 50-Ω source is connected to a T-attenuator, the two series resistors [R_1 and R_2 of Fig. 1.2.2(b)] each being 100 Ω, and the central parallel resistor [R_3 of Fig. 1.2.2(b)] being 150 Ω. Calculate the noise voltage appearing at the output terminals for a noise bandwidth of 1 MHz.
- 4.5.** The noise generated by a 1000-Ω resistor can be represented by a 4-nA current source in parallel with the 1000 Ω. Determine the equivalent emf source representation.
- 4.6.** Explain why inductance and capacitance do not generate noise.
- 4.7.** The available noise power spectral density at the input to an LC filter is kT_o joules. The filter has a transfer function that can be approximated by

$$\begin{aligned}
 |H(f)| &= 1 & 0 \leq f \leq 15 \text{ kHz} \\
 &= 1.75 - \frac{f}{2 \times 10^4} & 15 < f \leq 35 \text{ kHz} \\
 &= 0 & f > 35 \text{ kHz}
 \end{aligned}$$

Calculate the available output noise power.

- 4.8.** A 100-kΩ resistor at room temperature is placed at the terminals of the filter in Problem 4.7. Assuming that the resistor does not alter the response curve, calculate the mean-square voltage at the output.
- 4.9.** Explain how a capacitance C connected across a resistor R affects the thermal noise appearing at the terminals. What is the effective noise bandwidth of a 100-pF capacitor connected in parallel with a 10-kΩ resistor?
- 4.10.** A 100-kΩ resistor is connected in parallel with a 100-pF capacitor. Determine the effective noise bandwidth and the noise voltage appearing at the terminals of the combination.
- 4.11.** Calculate the equivalent noise bandwidth for the filter of Problem 4.7.
- 4.12.** A resistor has a self-capacitance of 3 pF. Calculate the mean-square noise voltage at its terminals at room temperature.
- 4.13.** Determine the mean-square noise voltage at the terminals of the resistor in Problem 4.12 if it is assumed that the self-capacitance is zero. Discuss the validity of this result.
- 4.14.** A single tuned circuit has a Q -factor of 70 and is resonant at 3 MHz with a 470 pF tuning capacitor. Calculate the equivalent noise bandwidth for the circuit.

- 4.15. The tuned circuit of Problem 4.14 is fed from a current source at resonance, the internal resistance of the source being equal to the dynamic impedance of the circuit. Calculate the equivalent noise bandwidth in this case.
- 4.16. A signal source having an internal resistance of $50\ \Omega$ is connected to a tap on the inductor of a tuned circuit, the tapping point being one-quarter up from the ground connection. The undamped Q -factor of the circuit is 75, and the tuning capacitance is 900 pF for resonance at 500 kHz . Calculate the equivalent noise bandwidth of the system, assuming that the tapping ratio can be treated as an ideal transformer coupling.
- 4.17. Write brief notes on the sources of noise, other than thermal, that arise in electronic equipment. Describe how the power spectral density varies with frequency in each case.
- 4.18. Calculate the shot noise current present on a direct current of 13 mA for a noise bandwidth of 7 MHz .
- 4.19. Calculate the mean-square spectral density of shot noise current accompanying a direct current of 5 mA .
- 4.20. An amplifier has an equivalent noise resistance of $300\ \Omega$ and an equivalent shot-noise current of $200\ \mu\text{A}$. It is fed from a signal source that has an internal resistance of $50\ \Omega$. Calculate the total noise voltage at the amplifier input for a noise bandwidth of 1 MHz .
- 4.21. Repeat Problem 4.20 for a source resistance of $600\ \Omega$.
- 4.22. A signal source has an emf of $3\ \mu\text{V}$ and an internal resistance of $450\ \Omega$. It is connected to an amplifier that has an equivalent noise resistance of $250\ \Omega$ and an equivalent shot noise current of $300\ \mu\text{A}$. Calculate the S/N ratio in decibels at the input. The equivalent noise bandwidth is 10 kHz .
- 4.23. An emf source of $1\ \mu\text{V}$ rms has an internal resistance of $600\ \Omega$. Calculate the S/N ratio at its terminals. Calculate the new S/N ratio at the terminals when the source is connected to a $600\text{-}\Omega$ load.
- 4.24. A telephone transmission system has three identical links, each having an S/N ratio of 50 dB . Calculate the output S/N ratio of the system. One of the links develops a fault that reduces its S/N ratio to 47 dB . Calculate the output S/N under these circumstances.
- 4.25. Three telephone circuits, each having an S/N ratio of 44 dB , are connected in tandem. Determine the overall S/N ratio. A fourth circuit is now added that has an S/N ratio of 34 dB . Determine the new overall value.
- 4.26. Define *noise factor* in terms of input and output signal-to-noise ratios of a network. The noise factor of an amplifier is given as $5:1$. If the input S/N is 50 dB , calculate the output S/N ratio in decibels.
- 4.27. The noise factor of a radio receiver is $15:1$. Calculate its noise figure. Determine the output S/N ratio when the input S/N ratio to the receiver is 35 dB .
- 4.28. The noise figure of an amplifier is 11 dB . Determine the fraction of the total available noise power contributed by the amplifier, referred to the input.
- 4.29. The available output noise power from an amplifier is 100 nW , the available power gain of the amplifier is 50 dB , and the equivalent noise bandwidth is 30 MHz . Calculate the noise figure.
- 4.30. The noise figure of an amplifier is 7 dB . Calculate the equivalent amplifier noise referred to the input for a bandwidth of 500 MHz .
- 4.31. Three amplifiers 1, 2, and 3 have the following characteristics:

$$\begin{array}{ll} F_1 = 9\text{ dB}, & G_1 = 48\text{ dB} \\ F_2 = 6\text{ dB}, & G_2 = 35\text{ dB} \\ F_3 = 4\text{ dB}, & G_3 = 20\text{ dB} \end{array}$$

The amplifiers are connected in tandem. Determine which combination gives the lowest noise factor referred to the input.

- 4.32.** An amplifier has an equivalent noise resistance of 350Ω and an equivalent shot noise current of $400 \mu\text{A}$. It is fed from a $1000\text{-}\Omega$ source. Calculate the noise factor.
- 4.33.** Calculate the optimum source resistance for the amplifier in Problem 4.32.
- 4.34.** A source has an internal resistance of 50Ω and is to be coupled into an amplifier input for which the equivalent noise generators are $R_n = 500 \Omega$ and $I_{EQ} = 300 \mu\text{A}$. Calculate the turns ratio of the input transformer required to minimize the noise factor, assuming that impedance is transformed according to the square of the turns ratio.
- 4.35.** A signal source is connected to an amplifier through a cable that is matched, but that introduces a loss of 2.3 dB . What is the noise factor of the cable?
- 4.36.** An attenuator has an insertion loss $IL = 0.24$. Determine its noise figure.
- 4.37.** An amplifier has a noise figure of 7 dB . Calculate its equivalent noise temperature.
- 4.38.** A cable has an insertion loss of 2 dB . Determine its equivalent noise temperature referred to the input.
- 4.39.** A cable that has a power loss of 3 dB is connected to the input of an amplifier, which has a noise temperature of 200 K . Calculate the overall noise temperature referred to the cable input.
- 4.40.** A mixer circuit has a noise figure of 12 dB . It is preceded by an amplifier that has an equivalent noise temperature of 200 K and a power gain of 30 dB . Calculate the equivalent noise temperature of the combination referred to the amplifier input.
- 4.41.** An amplifier has a gain of 12 dB and a noise temperature of 120 K . The amplifier may be connected into a receiving system at either end of the cable feeding the main receiver from the antenna. The cable has an insertion loss of 12 dB . Determine which connection gives the lowest overall noise figure. The noise characteristics of the main receiver may be ignored.
- 4.42.** A satellite receiving system consists of a low noise amplifier (LNA) that has a gain of 47 dB and a noise temperature of 120 K , a cable with a loss of 6.5 dB , and a main receiver with a noise factor of 9 dB . Calculate the equivalent noise temperature of the overall system referred to the input for the following system connections: (a) the LNA at the input, followed by the cable connecting to the main receiver; (b) the input direct to the cable, which then connects to the LNA, which in turn is connected directly to the main receiver.
- 4.43.** Using the information given in the text, derive Eq. (4.18.4). Two amplifiers are connected in cascade. The first amplifier has a noise temperature of 120 K and a power gain of 15 dB ; the second, a noise temperature of 300 K . Calculate the overall noise temperature of the cascaded connection referred to the input.
- 4.44.** The ENR for an avalanche diode is 13 dB . Given that the cold temperature is equal to room temperature ($= 290 \text{ K}$), determine the hot temperature of the source.
- 4.45.** In a noise measurement using an avalanche diode, the ENR was 14.3 dB . The noise power output with the diode on was 45 dBm , and with the diode off, 36 dBm . Calculate the noise temperature of the device under test.
- 4.46.** Derive Eq. (4.19.6) of the text. In the measurement of noise factor, the ENR = 13.7 dB and $Y = 7 \text{ dB}$. Calculate the noise figure and the equivalent noise temperature of the device under test.
- 4.47.** Explore MATLAB functions for generating random sequences. (Hint: `rand()`, and `randn()` functions.)
- 4.48.** Plot the thermal noise voltage generated by a $10k\Omega$ resistor, when the temperature is varied from $0K$ to $300K$. Let the bandwidth of interest be 10MHz . Use MATLAB/Mathematica.

- 4.49.** Explore the MATLAB functions to compute the power spectral density (PSD) of a noise signal.
- 4.50.** Write a MATLAB program to simulate the *rolling of a fair die*. (Hint: Use `randperm(6)`.)
- 4.51.** Generate a random noise vector containing 20 elements, with zero mean and variance 4. (Hint: Use `randn(1,20)*2`)
- 4.52.** Show that $\text{mean}(\text{randn}(1, x)) \rightarrow 0$ as $x \rightarrow \infty$.



Tuned Small-signal Amplifiers, Mixers, and Active Filters

5.1 Introduction

In this chapter, transistors are modeled using the hybrid- π small-signal equivalent circuit. This has the advantage that the same model, and hence the same methods of analysis, can be used for BJTs and FETs. Of importance is the ability to be able to set up the circuit equations from the equivalent circuit, and a number of examples are used to illustrate how this is done. Equally important is the ability to be able to interpret the equations, to see how various components are likely to influence the circuit performance.

Solving the equations is another matter, and the methods illustrated range from using approximate methods to get a “feel” for circuit performance, to “exact” solutions that really require the use of a personal computer or a good programmable calculator to implement. An effort has been made to steer clear of any one particular program, but Mathcad has been found to be very well suited to these applications and is available in a student’s edition.

Likewise, those students who have access to PSpice or Microcap may wish to use these to verify some of the example solutions given. Although computer simulation is not emphasized in the chapter, Microcap in particular is recommended as it utilizes on-screen capture of the circuit and provides excellent graphics output. A few examples of Microcap results are presented in the chapter.

5.2 The Hybrid- π Equivalent Circuit for the BJT

The hybrid- π equivalent circuit gets its name from the fact that the circuit configuration is π shaped, and the units are a mixture, or hybrid, containing a voltage-dependent current generator. A simplified version of the

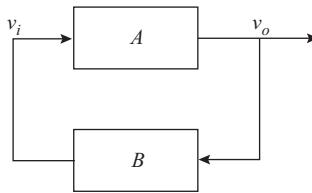


Figure 5.2.1 Simplified hybrid- π equivalent circuit for a BJT.

hybrid- π equivalent circuit for the bipolar junction transistor is shown in Fig. 5.2.1. The terminals marked B, E, and C are the external base, emitter, and collector terminals, available to the user. Terminal B' is internal to the transistor and is shown because the extrinsic base resistance $r_{b'b}$ must be taken into account in some situations at high frequencies.

The simplified circuit contains those elements that have most effect on the high-frequency response: the transconductance g_m ; the output resistance r_c ; the input resistance $r_{b'e}$; the collector output capacitance C_c ; the collector-to-base capacitance C_{cb} ; the base-to-emitter capacitance $C_{b'e}$; and the extrinsic base resistance $r_{b'b}$. In detail these are:

Transconductance: The transconductance is a function of collector current, given by

$$g_m = \frac{I_C}{V_T} \quad (5.2.1)$$

where $V_T = 26$ mV at room temperature. Thus the transconductance can be obtained immediately from a knowledge of the collector bias current I_C .

Output resistance: The output resistance is also a function of collector current and is given by

$$r_c = \frac{V_A}{I_C} \quad (5.2.2)$$

Here, V_A is known as the *early voltage*, a specified parameter for the transistor. Often, r_c is sufficiently large to have little effect on the circuit and can be ignored.

Input resistance: This is directly dependent on the g_m and is given by

$$r_{b'e} = \frac{\beta_o}{g_m} \quad (5.2.3)$$

Here, β_o is the low-frequency, short-circuit current gain, a specified parameter for the transistor.

Collector output capacitance C_c : In integrated circuits this is the depletion capacitance of the reverse-biased collector-to-substrate isolation junction. The value is a function of the reverse voltage. The value is normally small compared to other circuit capacitances and is usually specified for given operating conditions. In discrete devices it will include any stray capacitance from the collector circuit to ground.

Collector-to-base capacitance C_{cb} : This is the depletion capacitance of the reverse-biased collector-to-base junction. It is a function of the reverse voltage and is usually specified for given operating conditions. Although the value is normally small compared to other circuit capacitances, its effect can be magnified in a manner to be described shortly (see Miller effect).

Base-to-emitter capacitance $C_{b'e}$: This is the capacitance of the forward-biased base–emitter junction. It consists of two components, $C_{depl} + C_{diff}$. The depletion capacitance C_{depl} is a function of the forward bias on the junction and may be specified or estimated for given operating conditions. The diffusion capacitance C_{diff} is a function of the transconductance and is

$$C_{diff} = \tau_F g_m \quad (5.2.4)$$

where τ_F is the *forward transit time* for minority carrier movement through the base, a specified parameter for the transistor.

Extrinsic base resistance $r_{b'b}$: This is the resistance of the bulk base material, which effectively comes between the external terminal and the active part of the base–emitter junction. In many devices it may be considered negligible, and in others its value can be as high as 100Ω typically.

It will be seen therefore, that to obtain a reasonable picture of the operation of a BJT at high frequencies a number of parameters must be specified, these in summary being I_C , β_o , V_A , C_c , $C_{cb'}$, C_{depl} , $r_{b'b}$, and τ_F .

Exercise 5.2.1 Draw the hybrid- π equivalent circuit and show the values for a BJT for which $I_C = 1 \text{ mA}$, $\beta_o = 200$, $V_A = 600 \text{ V}$, $C_c = 0.6 \text{ pF}$, $C_{cb'} = 0.6 \text{ pF}$, $C_{dep} = 5 \text{ pF}$, $r_{b'b} = 30 \Omega$, and $\tau_F = 500 \text{ ps}$. (Ans. $g_m = 38.5 \text{ mS}$; $r_c = 60 \text{ k}\Omega$; $r_{b'e} = 5200 \Omega$; $C_{b'e} = 24.2 \text{ pF}$.)

The characteristics of discrete transistors are sometimes specified in terms of hybrid, or h -parameters. The term hybrid in this instance arises because these parameters are a mixture (or hybrid) of conductance and resistance. Table 5.2.1 shows the relationship between the h -parameters and the hybrid- π model:

TABLE 5.2.1

<i>h</i> -Parameter	Hybrid- π Model
h_{ie}	$r_{b'e}$
h_{fe}	β_o
h_{oe}	$1/r_c$
h_{fe}/h_{ie}	g_m

5.3 Short-circuit Current Gain for the BJT

The short-circuit current gain is a useful measure of how a transistor behaves with frequency. Referring to Fig. 5.3.1, which shows the transistor connected in the common-emitter configuration, the input or base current is given by

$$i_b = v_{b'e} \left(\frac{1}{r_{b'e}} + j\omega(C_{b'e} + C_{b'c}) \right) \quad (5.3.1)$$

Because the collector-to-emitter is short-circuited, the output resistance and capacitance of the equivalent circuit have no effect and these are not shown. The collector current, ignoring the small current that flows through $C_{cb'}$ is given by

$$i_c = g_m v_{b'e} \quad (5.3.2)$$

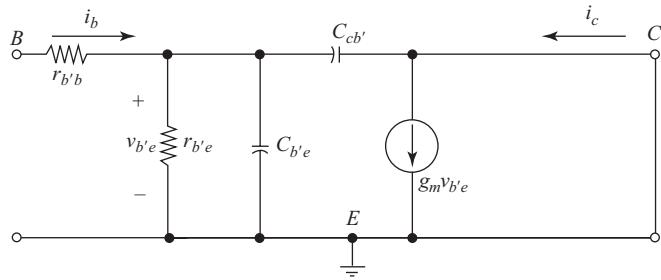


Figure 5.3.1 Circuit for determining the short-circuit current gain.

Hence the short-circuit current gain is

$$\begin{aligned} A_{isc} &= \frac{i_c}{i_b} \\ &= \frac{g_m}{\frac{1}{r_{b'e}} + j\omega(C_{b'e} + C_{bc})} \end{aligned} \quad (5.3.3)$$

The magnitude of this is

$$|A_{isc}| = \frac{g_m}{\sqrt{\frac{1}{r_{b'e}^2} + \omega^2(C_{b'e} + C_{bc})^2}} \quad (5.3.4)$$

Multiplying through by $r_{b'e}^2$ and recalling that $r_{b'e} = \beta_o/g_m$ allows the gain to be written as

$$|A_{isc}| = \frac{\beta_o}{\sqrt{(1 + r_{b'e}^2\omega^2(C_{b'e} + C_{bc})^2)}} \quad (5.3.5)$$

This is plotted in Fig. 5.3.2, and from the curve, two important parameters can be identified. These are the -3 -dB frequency ω_β and the unity-gain (0 dB) transition frequency ω_T .

The -3 -dB Frequency

The -3 -dB frequency is the frequency at which the gain magnitude drops by 3 dB from its maximum value. Denoting the -3 -dB frequency by ω_β , then

$$\frac{\beta_o}{\sqrt{2}} = \frac{\beta_o}{\sqrt{(1 + r_{b'e}^2\omega_\beta^2(C_{b'e} + C_{b'c})^2)}} \quad (5.3.6)$$

from which

$$\begin{aligned} \omega_\beta &= \frac{1}{r_{b'e}(C_{b'e} + C_{b'c})} \\ &= \frac{g_m}{\beta_o(C_{b'e} + C_{b'c})} \end{aligned} \quad (5.3.7)$$

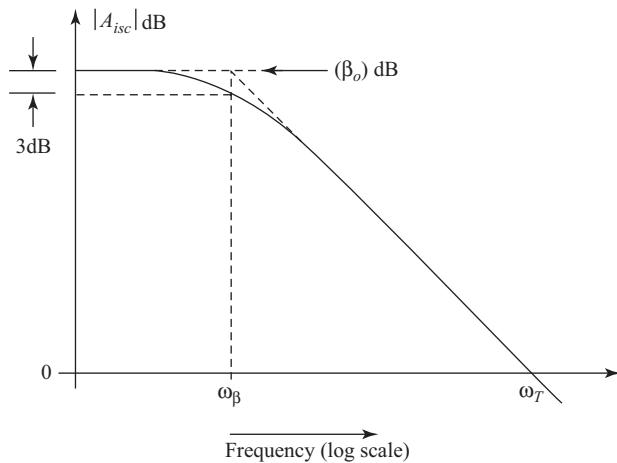


Figure 5.3.2 Short-circuit current gain as a function of frequency.

This shows the dependence of the -3-dB frequency on the transistor parameters. It must be kept in mind that this is for the transistor under short-circuit output conditions, but, even so, it gives a useful measure of how an amplifier using the transistor will behave at high frequencies.

Unity Gain Transition Frequency

This is the frequency at which the gain magnitude equals unity, or 0 dB. Again, making use of the gain magnitude equation gives

$$1 = \frac{\beta_o}{\sqrt{(1 + r_{b'e}^2 \omega_T^2 (C_{b'e} + C_{b'c})^2)}} \quad (5.3.8)$$

Now, since $\beta_o^2 \gg 1$, Eq. (5.3.8) is easily solved to give, to a very close approximation,

$$\omega_T = \frac{g_m}{C_{b'e} + C_{b'c}} \quad (5.3.9)$$

The transition frequency is seen to be independent of β_o , and for this reason it is relatively constant for a given transistor type under specified operating conditions. ω_T is the frequency parameter most often specified on transistor data sheets for a range of operating conditions.

For computer modelling, the *forward transit time* τ_F introduced in Eq.(5.2.4) is normally required. Unlike the transition frequency, the forward transit time is relatively independent of operating conditions. To find the forward transit time from the transition frequency equation, first find the diffusion capacitance. Thus, Eq.(5.3.9) can be written as

$$\omega_T = \frac{g_m}{C_{depl} + C_{diff} + C_{cb'}} \quad (5.3.10)$$

From this it is seen that

$$\frac{C_{diff}}{g_m} = \frac{1}{\omega_T} = \frac{C_{depl} + C_{cb'}}{g_m} \quad (5.3.11)$$

Hence, on substituting from Eq.(5.2.4)

$$\tau_F = \frac{1}{\omega_T} = \frac{C_{depl} + C_{cb'}}{g_m} \quad (5.3.12)$$

If the *bulk* resistance of the collector, denoted here by $r_{c'c}$ is significant, the effect of the base-collector capacitance is magnified by what is termed the *Miller effect* (this is described in Section 5.4) and Eq.(5.3.12) becomes

$$\tau_F = \frac{1}{\omega_T} - \frac{C_{depl} + C_{cb'}(1 + g_m r_{c'c})}{g_m} \quad (5.3.13)$$

5.4 Common-emitter (CE) Amplifier

The CE amplifier with tuned output and input circuits is shown in Fig. 5.4.1(a). C_3 and C_4 are dc blocking capacitors that have negligible reactance at high frequencies. The bias resistor R_{BIAS} supplies bias current to the base, and this can also be assumed to have negligible effect on the high-frequency performance. The signal source is shown as an equivalent current generator i_s and R_s . The equivalent circuit, using the hybrid- π equivalent circuit for the transistor, is shown in Fig. 5.4.1(b), where $r_{b'b}$ has been assumed negligible.

From the equivalent circuit of Fig. 5.4.1(b), it can be seen that the output resistance of the transistor and the load resistance are in parallel with the output tuned circuit. The output capacitance of the transistor, shown as C_c , is in parallel with the circuit tuning capacitance C_2 and will form part of the resonant circuit. Let the output inductor have a series resistance r_2 and inductance L_2 , then as shown in Fig. 5.4.1(c), the components on the output side can be grouped together in an admittance form as

$$Y_2 = \frac{1}{r_c} + \frac{1}{R_L} + \frac{1}{r_2 + j\omega L_2} + j\omega(C_c + C_2) \quad (5.4.1)$$

What is not immediately apparent from the equivalent circuit is the effect that the feedback capacitance $C_{cb'}$ has. To see the effect of this, the equivalent circuit in Fig. 5.4.1(c) can be analyzed. Admittance Y_1 represents the admittance of R_s , $r_{b'e}$ and $C_{b'e}$ in parallel with the input tuned circuit, and Y_2 the output admittance as previously defined. The feedback admittance is $Y_f = j\omega C_{cb'}$. The current equation for the output node is

$$\begin{aligned} 0 &= g_m v_i + (v_o - v_i) Y_f + v_o Y_2 \\ &= v_i(g_m - Y_f) + v_o(Y_2 + Y_f) \end{aligned} \quad (5.4.2)$$

From this the voltage gain is

$$\begin{aligned} A_v &= \frac{v_o}{v_i} \\ &= \frac{Y_f - g_m}{Y_f + Y_2} \end{aligned} \quad (5.4.3)$$

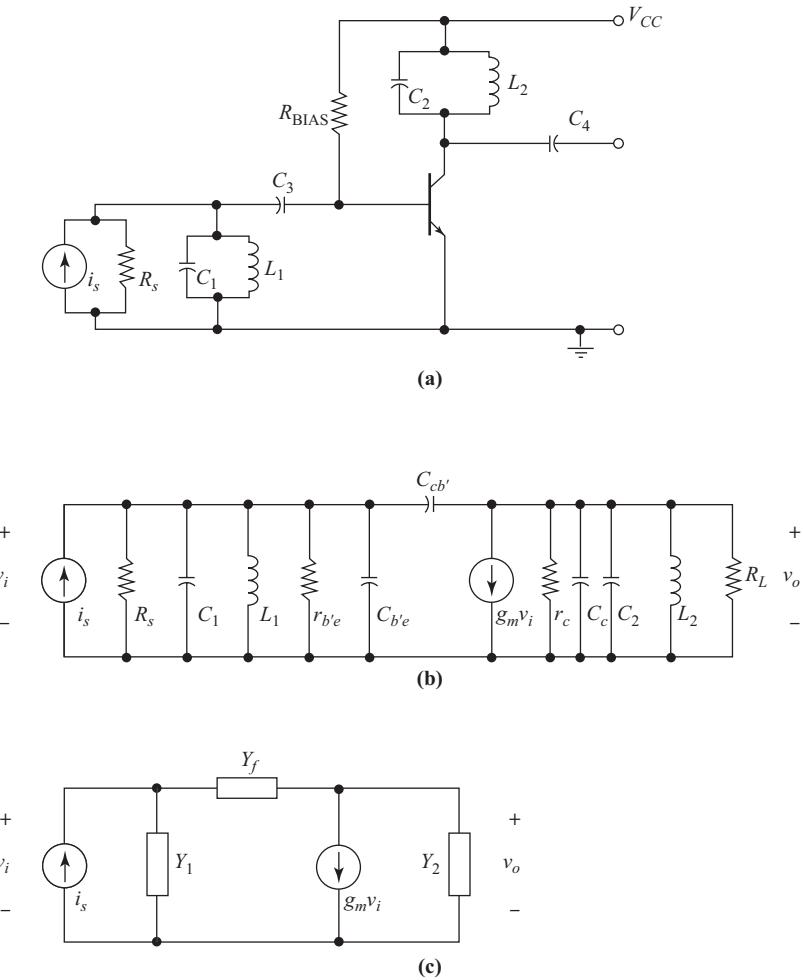


Figure 5.4.1 Tuned CE amplifier (a) circuit and (b) equivalent circuit. (c) Equivalent circuit for nodal analysis.

To compare Y_f and g_m , note that $Y_f = j\omega C_{cb'}$ and $g_m = \omega_T (C_{b'e} + C_{cb'})$. Since the transistor will be operated at a frequency $\omega \ll \omega_T$ to reduce the effects of feedback, and since $C_{b'e} \gg C_{cb'}$, then $g_m \gg |Y_f|$, and the expression for gain becomes

$$\begin{aligned} A_v &= -\frac{g_m}{Y_f + Y_2} \\ &= -\frac{g_m}{Y_o} \end{aligned} \quad (5.4.4)$$

where $Y_o = Y_2 + Y_f$. The gain is maximum when Y_o is resonant, which means that $C_{cb'}$ must be included in the output tuning. There is also a 180° phase shift in the gain under these conditions.

The output admittance can be written in the form from Eq.(1.4.2).

$$Y_o = \frac{1 + jyQ_{2\text{eff}}}{R_{D2\text{ eff}}} \quad (5.4.5)$$

where

$$\frac{1}{R_{D2\text{ eff}}} = \frac{1}{r_c} + \frac{1}{R_L} + \frac{(C_c + C_2 + C_{cb'})r_2}{L_2} \quad (5.4.6)$$

The last term on the right-hand side is the reciprocal of the dynamic resistance of the tuned circuit alone, but including the transistor capacitances. The damping effects of the transistor output resistance and load resistance are taken into account in the calculation of the effective dynamic resistance $R_{D2\text{ eff}}$. The effective Q -factor of the output circuit is

$$Q_2\text{ eff} = \omega_0(C_c + C_2 + C_{cb'})R_{D2\text{ eff}} \quad (5.4.7)$$

Hence the gain can be written as

$$A_v = -\frac{g_m R_{D2\text{ eff}}}{1 + jy Q_2\text{ eff}} \quad (5.4.8)$$

It should be noted that this is the voltage gain with reference to the input terminals, and for this reason the input admittance does not affect it.

Turning now to the input circuit, the equation for the input node is

$$\begin{aligned} i_s &= v_1(Y_1 + Y_f) - v_2Y_f \\ &= v_1(Y_1 + Y_f - A_v Y_f) \\ &= v_1(Y_1 + Y_f(1 - A_v)) \end{aligned} \quad (5.4.9)$$

Hence the input admittance is

$$\begin{aligned} Y_{\text{in}} &= \frac{i_s}{v_1} \\ &= Y_1 + Y_f(1 - A_v) \end{aligned} \quad (5.4.10)$$

Thus, to tune this to resonance, the effect of the feedback must be taken into account. In practice, this interaction with the output circuit can complicate the tuning procedure. The admittance term $Y_f(1 - A_v)$ is often referred to as the *Miller input admittance*, named after J. M. Miller, whose name is also given to a theorem dealing with the feedback case in general.

Substituting for gain expression gives

$$Y_{\text{in}} = Y_1 + j\omega C_{cb'} \left(1 + \frac{g_m R_{D2\text{ eff}}}{1 + jy Q_2\text{ eff}} \right) \quad (5.4.11)$$

At the resonant frequency of the output circuit, this becomes

$$Y_{\text{in}} = Y_1 + j\omega C_{cb'}(1 + g_m R_{D2\text{ eff}}) \quad (5.4.12)$$

Thus, for this situation, the Miller admittance can be represented by a capacitor C_M :

$$C_M = (1 + g_m R_{D2\text{ eff}})C_{cb'} \quad (5.4.13)$$

The Miller capacitance must be included in the tuning of the input circuit for this to be resonant at the same frequency as the output circuit. Note, however, that the frequency response of the input tuned circuit will not be that of a singly tuned circuit, because the Miller admittance is a function of frequency in general and, in fact, also introduces a conductance component at frequencies off resonance.

EXAMPLE 5.4.1

A CE amplifier has a tuned circuit in the collector, which resonates at 5 MHz with a total tuning capacity of 100 pF. The undamped Q -factor of the tuned circuit is 150. The amplifier feeds a load resistance of 5 k Ω , and the output resistance of the transistor is 40 k Ω . Calculate the voltage gain referred to the input terminals and the Miller capacitance at the input. The transistor operates a collector current of 500 μ A, and the collector-to-base capacitance is 0.6 pF.

SOLUTION

$$g_m = \frac{I_c}{V_T} = \frac{500 \times 10^{-6}}{26 \times 10^{-3}} = 0.019 \text{ S}$$

$$R_{D2} = \frac{Q}{\omega_o C} = \frac{150}{2\pi \times 5 \times 10^6 \times 10^{-10}} = 47.75 \text{ k}\Omega$$

At resonance the output admittance is purely conductive and is

$$\begin{aligned} Y_o &= \frac{1}{r_c} + \frac{1}{R_D} + \frac{1}{R_L} \\ &= \frac{1}{40 \times 10^3} + \frac{1}{47.75 \times 10^3} + \frac{1}{5 \times 10^3} \\ &= 246 \mu\text{S} \\ A_v &= -\frac{g_m}{Y_o} = -78 \end{aligned}$$

Therefore,

$$C_M = (1 + 78) \times 0.6 = 47 \text{ pF}$$

From this example, it is seen that the effect of the 0.6-pF capacitance translates to a Miller input capacitance of 47 pF, and this will be in addition to the already existing $C_{b'e}$ capacitance.

Returning now to the input circuit, at resonance its dynamic resistance is obtained from

$$\frac{1}{R_{D1\text{ eff}}} = \frac{1}{R_s} + \frac{1}{R_{D1}} + \frac{1}{r_{b'e}} \quad (5.4.14)$$

where R_{D1} is the dynamic resistance of the input tuned circuit by itself. Also, assuming the Miller admittance is constant over the frequency range of the -3-dB bandwidth, the effective Q -factor of the input circuit is

$$Q_{1\text{ eff}} = \omega_o(C_1 + C_{b'e} + C_M)R_{1\text{ eff}} \quad (5.4.15)$$

In many situations the input signal source is represented by a voltage equivalent generator, and the voltage gain referred to the source emf is of importance. In terms of the equivalent current generator source, the emf $v_s = i_s R_s$, or $i_s = v_s G_s$. Hence, from the input node equation,

$$\begin{aligned} i_s &= v_1 Y_{\text{in}} \\ \therefore v_s G_s &= v_1 Y_{\text{in}} \\ &= \frac{v_2}{A_v} Y_{\text{in}} \end{aligned} \quad (5.4.16)$$

The gain referred to the source emf is therefore

$$\begin{aligned} A_{vs} &= \frac{v_2}{v_s} \\ &= \frac{G_s}{Y_{\text{in}}} A_v \end{aligned} \quad (5.4.17)$$

EXAMPLE 5.4.2

For the amplifier of Example 5.4.1, the input tuned circuit has a Q -factor of 100 at a frequency of 5 MHz, the inductance being 2 μH . The source resistance is 1000 Ω . The transistor β is 200, and $C_{b'e} = 10 \text{ pF}$. Calculate the effective Q -factor of the input circuit and the voltage gain referred to the source emf.

SOLUTION From Example 5.4.1, $A_v \approx -78$ and $C_M \approx 47 \text{ pF}$. The dynamic resistance of the tuned circuit is

$$R_{D1} = Q\omega_o L = 4.71 \text{ k}\Omega$$

The effective dynamic conductance is

$$\frac{1}{R_{D1 \text{ eff}}} = \frac{1}{R_s} + \frac{1}{R_{D1}} + \frac{1}{r_{b'e}} = 1.31 \text{ mS}$$

Hence,

$$R_{D1 \text{ eff}} = 764 \Omega$$

The effective Q -factor is

$$Q_{\text{eff}} = \frac{R_{D1 \text{ eff}}}{\omega_o L} \approx 24$$

The voltage gain referred to source is

$$\begin{aligned} A_{vs} &= \frac{G_s}{Y_{\text{in}}} A_v \\ &= \frac{R_{D1 \text{ eff}}}{R_s} A_v \\ &= \frac{-78 \times 764}{1000} \\ &\approx -60 \end{aligned}$$

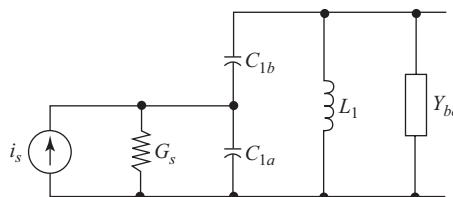


Figure 5.4.2 Capacitive input tap.

This example shows how the *Q*-factor can be severely reduced. In practice, to avoid this, the source and the transistor are usually connected through capacitive or inductive taps on the input tuned circuit, as discussed in Chapter 1. Figure 5.4.2 shows the source connected through a capacitive tap. This reduces the damping effect of the source conductance on the tuned circuit and, in addition, provides matching. As described in Section 4.16, the effective source resistance may be optimized for minimum noise factor through input coupling.

Output Coupling

The connections to the output tuned circuit may also be coupled in such a manner as to reduce damping. One such method utilizes mutual inductive coupling as shown in Fig. 5.4.3 and as discussed in Chapter 1. The voltage gain from the $b' - e$ terminals to the output is

$$A'_v = -g_m Z_T \quad (5.4.18)$$

where Z_T is the transfer impedance as given in Section 1.8. As before, A'_v is the gain referred to the internal terminals $b' - e$, which is also the gain from the external terminals if $r_{b'b}$ is negligible. The technique to follow in computing the transfer impedance is summarized here for convenience:

1. Compute the transformer impedances Z_p , Z_s , and Z_m [see Eqs. (1.8.1), (1.8.2), and (1.8.3)].
2. Compute the external impedances attached to primary and secondary, Z_1 and Z_2 [Eqs. (1.8.4), (1.8.5), and (1.10.6)].
3. Compute the system determinant Δ [see Example 1.8.1].
4. Compute the transfer impedance Z_T [Eq. (1.8.7)].

Evaluation of the gain using the transfer impedance is left as Problem 5.17. The circuit may also be analyzed using one of several computer analysis programs available, and Figure 5.4.4 shows the gain response for the circuit, obtained using Microcap.

Coupling may also be arranged with a tuned secondary, untuned primary, and the analysis procedure is similar to that outlined here.

Nodal analysis may be applied in a more formal way to analyse a complete circuit. Fig. 5.4.5 shows the small-signal equivalent circuit for a CE amplifier for which the transfer impedance is assumed known. The circuit also takes $r_{b'b}$ into account. With nodal analysis, it is more convenient to work with admittances rather than impedances and these are defined as

$$g_{b'b} = 1/r_{b'b}$$

$$Y_{b'e} = 1/r_{b'e} + j\omega C_{b'e}$$

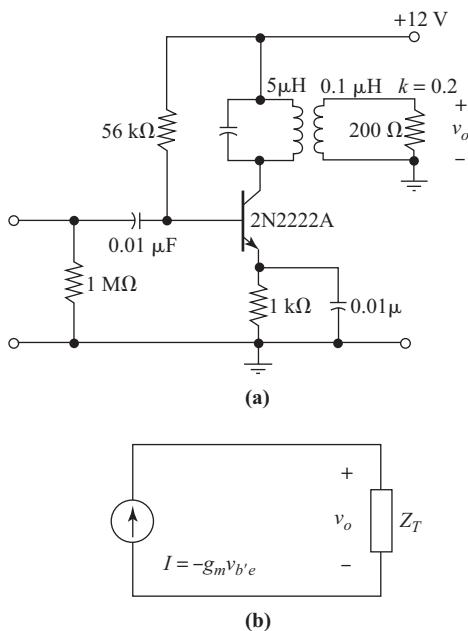


Figure 5.4.3 (a) Tuned primary, untuned secondary CE amplifier (component values used to obtain the Microcap response curve, Fig. 5.4.4). (b) Equivalent circuit.

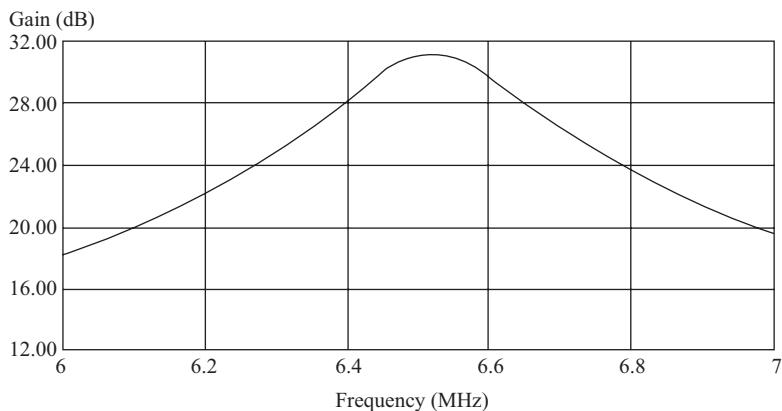


Figure 5.4.4 Gain/frequency response for the amplifier of Fig. 5.4.3, obtained using Microcap.

$$Y_f = j\omega C_{cb'}$$

$$Y_T = 1/Z_T$$

$$G_s = 1/R_s$$

The admittance Y_T includes the output capacitance and conductance of the transistor in parallel with the output coupling circuit, and the admittance Y_1 shown in Fig. 5.4.5 includes the source conductance in parallel with the input coupling circuit.

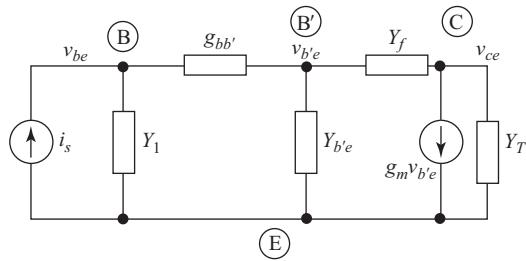


Figure 5.4.5 Small-signal equivalent circuit used in nodal analysis.

The nodal equations for the circuit of Fig. 5.4.5 are

$$\text{Node B: } i_s = (Y_1 + g_{b'b})v_{be} - g_{b'b}v_{b'e} + 0v_{ce}$$

$$\text{Node B': } 0 = -g_{b'b}v_{be} + (Y_{b'e} + g_{b'b} + Y_f)v_{b'e} - Y_f v_{ce}$$

$$\text{Node C: } 0 = 0v_{be} + (g_m - Y_f)v_{b'e} + (Y_T + Y_f)v_{ce}$$

These can be written in matrix form as

$$\begin{pmatrix} i_s \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} Y_1 + g_{b'b} & -g_{b'b} & 0 \\ -g_{b'b} & Y_{b'e} + g_{b'b} + Y_f & -Y_f \\ 0 & g_m - Y_f & Y_T + Y_f \end{pmatrix} \begin{pmatrix} v_{be} \\ v_{b'e} \\ v_{ce} \end{pmatrix}$$

or more concisely as

$$I = YV$$

The solution for the voltage matrix V is then

$$V = Y^{-1}I$$

where Y^{-1} is known as the *inverse* of the admittance matrix Y . Calculating the inverse matrix is no easy task if attempted “by hand,” but can be carried out literally “at the stroke of a key” with the appropriate software on a personal computer.

The voltage matrix will be a column matrix, and the last element in this, formally designated as V_{31} is identical to the output voltage v_{ce} . By setting the source voltage $v_s = 1V$, the element V_{31} also gives the voltage gain referred to the source emf. Setting the source emf equal to 1V means that the source current is numerically equal to G_s but in setting this up for computer solution it is necessary to maintain the correct units for all the quantities. Thus i_s would have to be entered as $1V \cdot G_s$. Since the other node voltages are contained in V , the voltage gain between any of the nodes can be determined.

EXAMPLE 5.4.3

The following parameters apply to a CE amplifier:

$$Z_T = 600\Omega; g_m = .04S; \beta_o = 200; r_{b'b} = 70\Omega; c_{b'e} = 7pF; C_{cb'} = 1pF; R_s = 50\Omega.$$

The input impedance, excluding the source is $Z'_1 = 300\Omega$. Determine the voltage gain, referred to the source emf, at a frequency of 5 MHz.

SOLUTION $r_{b'e} = \beta_o/g_m = 5K\Omega$

$$Y_{b'e} = 1/r_{b'e} + j\omega C_{b'e} = 0.2 + j.22mS$$

$$Y_f = j\omega C_{cb'} = j.031mS$$

$$g_{b'b} = 1/70 = 14.286 mS$$

$$G_s = 1/50 = 20mS$$

$$Y_T = 1/600 = 1.667mS$$

$$Y_1 = 1/Z'_1 + G_s = 23.333mS$$

In the following computations the values have been rounded off to the nearest decimal place. The Y-matrix is

$$Y = \begin{pmatrix} 37.6 & -14.3 & 0 \\ -14.3 & 14.5 + j.3 & 0 \\ 0 & 40 & 1.7 \end{pmatrix} mS$$

The inverse of this is

$$Y^{-1} = \begin{pmatrix} 42.3 - j1.7 & 41.3 - j4.6 & .1 + j.8 \\ 41.3 - j4.6 & 08.9 - j12.1 & .3 + j2 \\ -989.5 + j129.3 & -2600 + j340.5 & 592.4 - j60.3 \end{pmatrix} \text{ohms}$$

with $V_s = 1$ volt, the source current is $i_s = 1Y \cdot G_s = .02 A$. The I-matrix is therefore

$$I = \begin{pmatrix} .02 \\ 0 \\ 0 \end{pmatrix}$$

The matrix solution (obtained in this case using Mathcad) yields

$$V = \begin{pmatrix} .8 \\ .8 - j.1 \\ -19.8 + j2.6 \end{pmatrix}$$

Thus the voltage gain is $A_{vs} = V_{31} = -19.8 + j2.6$

The magnitude of this is $|A_{vs}| \approx 20$ and the phase shift is $\arg(A_{vs}) \approx 173^\circ$.

It should be noted that when using a program such as Mathcad, it is useful to evaluate key intermediate values as a check, although it is not necessary to know these for the computation to proceed. Some of the intermediate values are shown in the example.

For this example, a very quick estimate of the gain is obtained as follows. The input signal is attenuated approximately by $Z_1/(Z_1 + R_s) = 0.86$. The voltage gain of the stage referred to the input terminals is $-g_m Z_T = -24$. Hence the gain referred to source is $0.86 \times -24 \cong -21$.

5.5 Stability and Neutralization

A parallel tuned circuit is inductive at frequencies below the resonant frequency (see Chapter 1). There will be a frequency therefore where the input and output circuits of the tuned-input, tuned-output amplifier are inductive, as shown in Fig. 5.5.1(a). In this equivalent circuit, the input circuit is represented by a net inductance L_B in parallel with a resistance R_B ($r_{bb'}$ is assumed negligible), and the output tuned circuit by L_C in parallel with R_C and r_c .

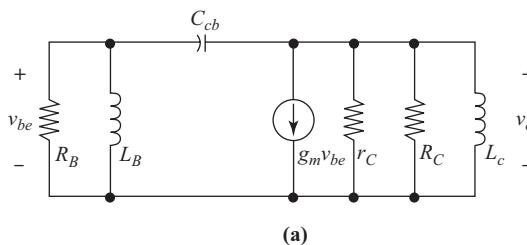
If the resistances R_B and R_C are large compared with the corresponding inductive reactance, the circuit reduces to that shown in Fig. 5.5.1(b). The loop consisting of L_B , L_C , and C_{cb} forms a resonant circuit for which the resonant frequency is

$$\omega_o = \frac{1}{\sqrt{(L_B + L_C)C_{cb}}} \quad (5.5.1)$$

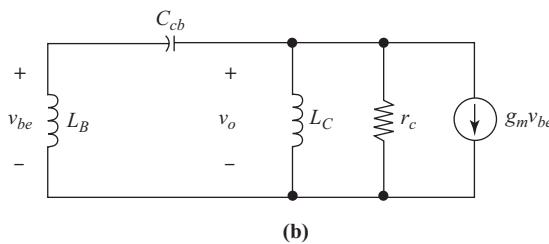
At the resonant frequency of the loop the system develops self-sustaining oscillations, the voltage v_{be} being provided through feedback from the output voltage v_o . Initially it requires only a random voltage, for example noise, to "kick-start" the system into the oscillatory mode. The detailed analysis will not be carried out here, but the approximate condition for maintenance of oscillations is

$$g_m r_c \geq \frac{L_C}{L_B} \quad (5.5.2)$$

An amplifier that can burst into oscillation is termed *unstable*, and of course instability is undesirable. One way of preventing instability is to provide damping through resistors R_B and R_C . If these are made sufficiently small compared to the corresponding inductive reactances, they will reduce the inductive currents,



(a)



(b)

Figure 5.5.1 (a) Equivalent circuit for the tuned-input, tuned-output amplifier at frequencies below resonance. (b) Circuit of (a) redrawn to emphasize the oscillatory loop.

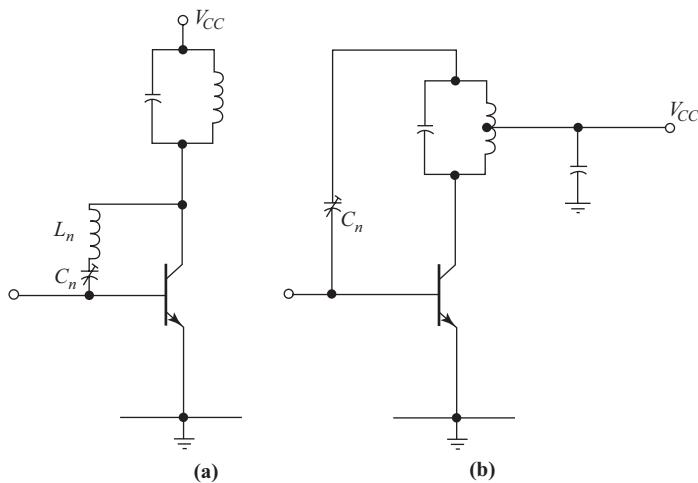


Figure 5.5.2 Neutralizing circuits.

and oscillations will not start. An amplifier can be made *unconditionally stable* in this way, but the price paid is a reduction in amplifier gain.

Stability can also be achieved by neutralizing the feedback signal. Thus an inductor may be connected in parallel with C_{cb} in order to form a parallel (high-impedance) circuit at the oscillatory frequency. This is shown in Fig. 5.5.2(a). The capacitor C_n in series with L_n is a dc blocking capacitor, which is adjustable so that the branch provides the net inductance required for parallel resonance with C_{cb} .

Figure 5.5.2(b) shows another method of neutralizing an amplifier. The signal fed back from the top of the tuned circuit is in antiphase to that at the collector end, and C_n can be adjusted to make the two signals equal in magnitude.

5.6 Common-base Amplifier

The effect of the feedback capacitor $C_{cb'}$ can be nullified completely by connecting the transistor in the common-base configuration, the small signal equivalent circuit for which is shown in Fig. 5.6.1. In this mode of operation, $C_{cb'}$ appears in parallel with the output capacitance C_c and therefore does not contribute to the input capacitance. The input resistance is α_o/g_m where $\alpha_o = \beta_o/(\beta_o + 1) \approx 1$. The input resistance for the CB circuit is therefore much smaller than that for the CE circuit which is given by β_o/g_m . The input capacitance is $C_{eb'} = C_{b'e}$. The output resistance for the CE circuit appears between collector and emitter. This is higher than the CE output resistance and can be shown to be given by $r_{cCB} \approx \beta_o r_{cCE}$. Because of its very high value the output resistance can be ignored for most practical purposes. The simplified equivalent circuit is shown in Fig. 5.6.1(b).

By applying a short-circuit to the output terminals of Fig. 5.6.1(b) and defining the currents as shown, the short-circuit output current is $i_c = -g_m v_{eb}$ and the input current is $i_b = (1/r_{eb} + j\omega C_{eb})v_{eb}$. Hence the short-circuit current gain for the CB amplifier is

$$A_{isc} = \frac{i_c}{i_b} = \frac{-g_m}{\frac{1}{r_{eb}} + j\omega C_{eb}} \quad (5.6.1)$$

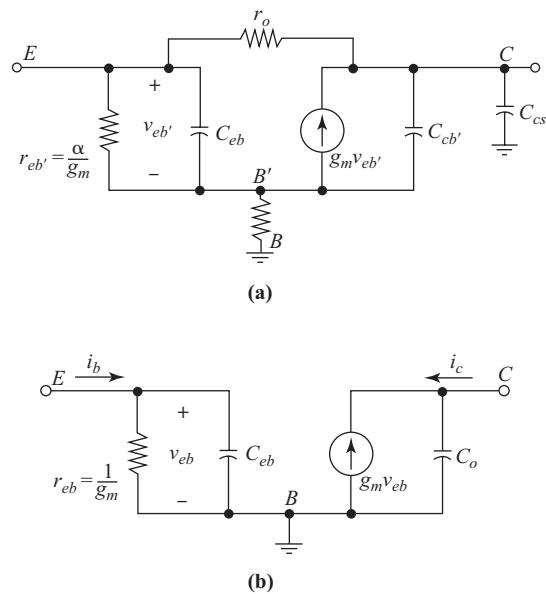


Figure 5.6.1 (a) Equivalent circuit for the common-base transistor. (b) Simplified version of (a) suitable for most practical purposes.

It is left as an exercise for the student to show that this can be expressed as

$$A_{isc} = \frac{\alpha_o}{1 + j \frac{\omega}{\omega_\alpha}} \quad (5.6.2)$$

where the -3 dB frequency for the short-circuit gain of the CB stage is $\omega_\alpha = 1/r_{eb}C_{eb}$.

Now, because $C_{eb} = C_{b'e}$ and $C_{b'e} \gg C_{cb}$, and $r_{eb} \cong 1/g_m$, it is easily shown that

$$\omega_\alpha \cong \omega_T \quad (5.6.3)$$

Thus, the -3 -dB frequency is, to a very close approximation, equal to f_T , the unity gain transition frequency. As shown previously, this is higher than the -3 -dB frequency for the CE connection by a factor β_o .

A basic CB amplifier circuit is shown in Fig. 5.6.2(a). From Fig. 5.6.2(c) the voltage gain referred to the e-b terminals is seen to be

$$\begin{aligned} A_v &= g_m Z_L \\ &= \frac{g_m R_D}{1 + jyQ} \end{aligned} \quad (5.6.4)$$

where Eq.(1.4.2) is used for the impedance of a parallel tuned circuit. Where a coupled circuit is used the gain is given by $g_m Z_T$ as before. It will be noticed that at resonance there is no phase shift with the CB amplifier, which contrasts with a 180 degree phase shift for the CE amplifier. The gain magnitude is approximately the same for both configurations.

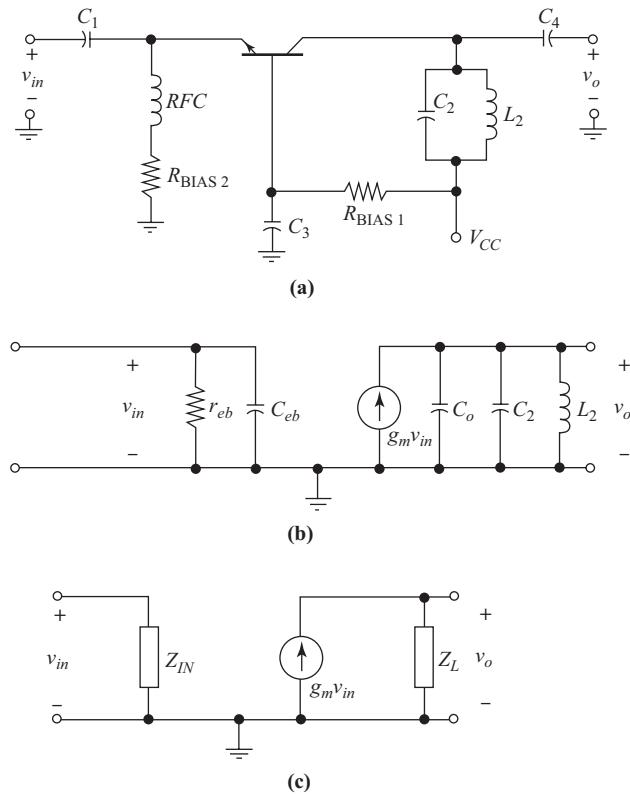


Figure 5.6.2 (a) CB amplifier with tuned collector load. (b) Equivalent circuit. (c) Equivalent circuit simplified.

As shown in the next section, the available power gain of the CB stage is lower than that for the CE stage, which limits its usefulness as a front-end amplifier.

5.7 Available Power Gain

In Section 4.15 it is shown that a high available power gain is needed to maintain a low noise factor with cascaded amplifiers (Friis's formula). An estimate of the available power gain of the CB and CE amplifiers can be made as follows.

Figure 5.7.1 shows a basic amplifier circuit. The available power from the source is

$$P_s = \frac{V_s^2}{4R_s} \quad (5.7.1)$$

The available power at the output is

$$P_o = \frac{I_o^2}{4G_o} = \frac{I_o^2 R_o}{4} \quad (5.7.2)$$

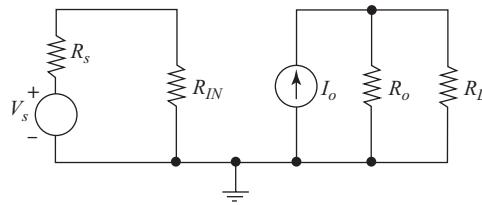


Figure 5.7.1 Equivalent circuit used to estimate the available power gain of an amplifier.

The available power gain is

$$\begin{aligned} G_{AV} &= \frac{P_s}{P_o} \\ &= \frac{I_o^2 R_s R_o}{V_s^2} \end{aligned} \quad (5.7.3)$$

For both the CB and the CE circuits, $I_0^2 = g_m^2 V_{in}^2$, and hence

$$G_{AV} = g_m^2 R_s R_o \left(\frac{R_{IN}}{R_s + R_{IN}} \right)^2 \quad (5.7.4)$$

Now, for the CE circuit, $R_{IN} = r_{be} = \beta_o/g_m$, and for the CB circuit, $R_{IN} = r_{eb} \cong 1/g_m$. Since the transistor output resistance is high in both cases, the output resistance in both cases will be that of the collector tuned circuit (with the actual load disconnected). The ratio of available power gains then becomes

$$\frac{G_{AVCE}}{G_{AVCB}} = \left(\frac{r_{be}}{r_{eb}} \right)^2 \cdot \left(\frac{R_s + r_{eb}}{R_s + r_{be}} \right)^2 \quad (5.7.5)$$

This can be simplified to

$$\frac{G_{AVCE}}{G_{AVCB}} = \left(\frac{1 + g_m R_s}{1 + g_m R_s / \beta_o} \right)^2 \quad (5.7.6)$$

This shows that the available power gain for the CE amplifier is greater than that for the CB amplifier. For this reason, the CE amplifier is preferred for the input stages of low-noise receivers. It should be noted that the underlying reason for the lower power gain of the CB amplifier is the low input resistance, which is $1/\beta_o$ times that of the CE amplifier.

5.8 Cascode Amplifier

The common-emitter and common-base amplifiers can be combined to form an amplifier unit that has high power gain and is stable. The combined unit is known as a cascode amplifier (the word is a relic from

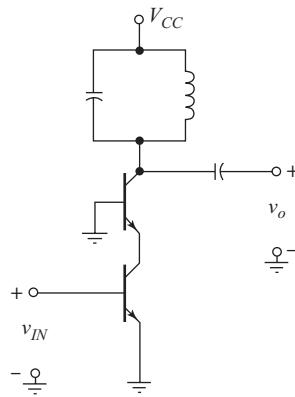


Figure 5.8.1 Basic cascode amplifier.

vacuum tube technology, where the original circuit employed *cascaded common-cathode* and *common-grid* stages).

A basic cascode amplifier is shown in Fig. 5.8.1, where the bias components are omitted for simplicity. Both transistors carry the same collector current and hence will have equal transconductances. The effective load seen by the CE stage is the input resistance of the CB stage, which is α_o/g_m . Hence the voltage gain of the CE stage is $g_m\alpha_o/g_m = \alpha_o$, or just slightly less than unity. This means that the feedback will be insufficient to cause oscillations. The voltage gain of the CB stage is g_mZ_L , and hence the overall voltage gain is $\alpha_o g_m Z_L \cong g_m Z_L$. The CB stage is inherently stable, as discussed previously, so the overall amplifier is stable.

The input resistance of the CE stage is r_{be} . Overall, therefore, the cascode amplifier has performance characteristics similar to those of a CE amplifier but with stability (and, note, no 180° phase change), and hence its available power gain is high.

5.9 Hybrid- π Equivalent Circuit for an FET

In many ways the field effect transistor (FET) is simpler than the bipolar junction transistor (BJT) because of the extremely high input impedance presented by the control gate. The hybrid- π equivalent circuit is shown in Fig. 5.9.1. In this, the external terminals are labeled *G* for gate, *S* for source, and *D* for drain. Analysis of circuits utilizing the FET then proceeds in a manner similar to that for the BJT using the hybrid π equivalent circuit.

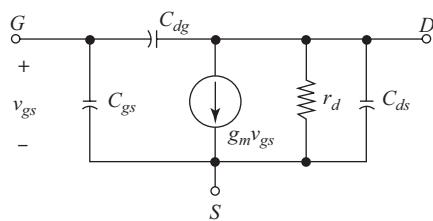


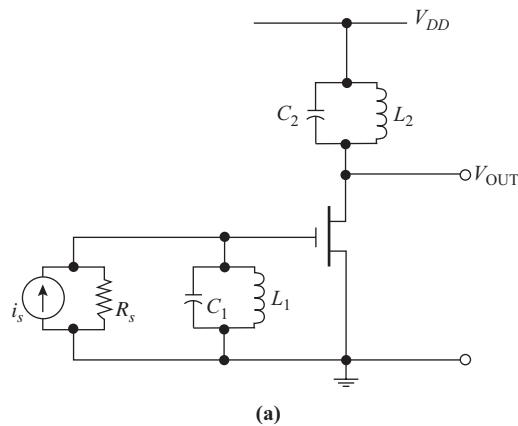
Figure 5.9.1 Hybrid- π equivalent circuit for an FET.

Figure 5.9.2(a) shows a simple CS amplifier, where for simplicity the bias components are omitted. The equivalent circuit is shown in Fig. 5.9.2(b), where it will be seen that the input tuning capacitance is $C_1 + C_{gs}$ (this excludes the Miller capacitance, which is taken care of in the analysis). Likewise, the output tuning capacitance is $C_2 + C_{ds}$. The circuit can be further reduced to that shown in Fig. 5.9.2(c). The nodal equations in matrix form are

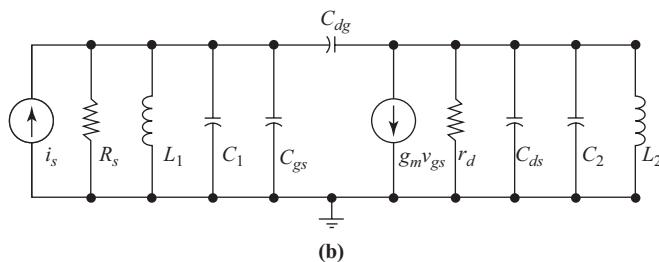
$$\begin{pmatrix} i_s \\ 0 \end{pmatrix} = \begin{pmatrix} Y_1 + Y_f & -Y_f \\ g_m - Y_f & Y_2 - Y_f \end{pmatrix} \begin{pmatrix} v_{gs} \\ v_{ds} \end{pmatrix} \quad (5.9.1)$$

or

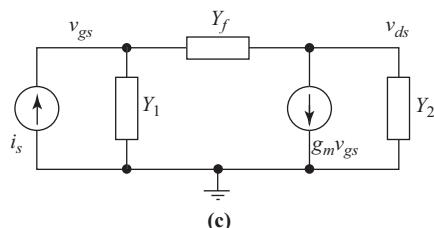
$$I = YV$$



(a)



(b)



(c)

Figure 5.9.2 (a) Basic common-source amplifier. (b) Equivalent circuit. (c) Equivalent circuit used in nodal analysis.

Hence the voltages are

$$V = Y^{-1}I \quad (5.9.2)$$

Again, the equations are presented in this manner so that a personal computer or suitable programmable calculator can be readily used to find the solutions.

EXAMPLE 5.9.1

An FET has the following parameters: $g_m = 2 \text{ mS}$; $C_{gs} = 5 \text{ pF}$; $C_{dg} = 1 \text{ pF}$; $C_{ds} = 1 \text{ pF}$; $r_d = 13 \text{ k}\Omega$. It is used in a tuned-input, tuned-output CS amplifier, both circuits being tuned to resonance at 10 MHz. Tuning includes the capacitances C_{gs} and C_{ds} , but not the effects of C_{dg} . The dynamic resistance of the input circuit is $3 \text{ k}\Omega$ (excluding R_s), and that of the output circuit is $10 \text{ k}\Omega$ (excluding r_d). The signal source has an internal resistance of 600Ω . Determine the voltage gain referred to the source emf.

SOLUTION At the resonant frequency

$$Y_1 = \frac{1}{R_s} + \frac{1}{R_{D1}} = 2 \text{ mS}$$

and

$$Y_2 = \frac{1}{r_d} + \frac{1}{R_{D2}} = 0.177 \text{ mS}$$

Also,

$$Y_f = j\omega_o C_{dg} = j0.063 \text{ mS}$$

The Y matrix in millisiemens evaluates as

$$Y = \begin{pmatrix} 2 + j0.063 & -j0.063 \\ 2 - j0.063 & 0.177 + j0.063 \end{pmatrix}$$

With $v_s = 1 \text{ V}$, the I matrix in milliamperes is

$$I = \begin{pmatrix} 1.667 \\ 0 \\ 0 \end{pmatrix}$$

The matrix solution (obtained in this case using Mathcad) yields

$$V = \begin{pmatrix} 0.68 - j0.21 \\ -5.94 + j4.7 \end{pmatrix}$$

Thus the voltage gain is

$$A_{vs} = V_{21} = -5.94 + j4.7$$

Note in this example that a quick estimate of the gain along the lines of that used following Example 5.4.3 does not yield an accurate result. The reason is that Y_f is larger and in this case is comparable in magnitude with the effective load admittance. In order to find the variation of voltage gain with frequency, the easiest way is to solve the nodal equations, and it is left as an exercise for the student to show that

$$\begin{aligned} A_{vs} &= \frac{v_{ds}}{v_s} \\ &= -\frac{g_m - Y_f}{R_s \Delta} \end{aligned} \quad (5.9.3)$$

where

$$\Delta = (Y_1 + Y_f) \cdot (Y_2 + Y_f) + Y_f(g_m - Y_f) \quad (5.9.4)$$

Plotting voltage gain as a function of frequency is left to problems 5.32 and 5.33.

5.10 Mixer Circuits

Mixers are used to change a signal from one frequency to another. There are a number of reasons why frequency changing is required, and in fact a number of mixing processes are used in specialized applications, which come under different names. Modulation, demodulation, and frequency multiplication are examples of these, which are covered in later chapters. The term *mixer* is generally reserved for circuits that change a radio-frequency signal to some intermediate value (known as the intermediate frequency or IF) and that require input from a local oscillator (LO) to do so. The general features of these circuits are covered in this section.

Some types of mixers (notably those used for microwaves) are available as packaged units, with input ports labeled RF and LO and an output port labeled IF. In certain receiver applications the oscillator circuit is an integral part of the mixer circuit, and only the RF input and IF output are readily identified.

All mixer circuits make use of the fact that, when two sinusoidal signals are multiplied together, the resultant consists of sum and difference frequency components. This can be demonstrated as follows. Let the oscillator signal be represented by

$$v_{osc} = V_{osc} \sin \omega_{osc} t \quad (5.10.1)$$

and the RF signal by

$$v_{sig} = V_{sig} \sin \omega_{sig} t \quad (5.10.2)$$

Multiplying these two signals together gives

$$\begin{aligned} v_{osc} v_{sig} &= V_{osc} \sin \omega_{osc} t V_{sig} \sin \omega_{sig} t \\ &= \frac{V_{osc} V_{sig}}{2} (\cos(\omega_{osc} - \omega_{sig})t - \cos(\omega_{osc} + \omega_{sig})t) \end{aligned} \quad (5.10.3)$$

The term containing the frequency $\omega_{osc} - \omega_{sig}$ is the one normally selected, by filtering, as the intermediate frequency (IF) signal (in certain specialized applications, the other, higher-frequency component may be selected). It will be noted that neither one of the two input frequencies is present in the output, only the sum and difference frequencies.

Diode Mixer

The circuit for a diode mixer is shown in Fig. 5.10.1. The two signals are connected in series, and a bias voltage may also be applied to optimize the working point on the diode. The diode V/I characteristic is non-linear, which results in the current having a term proportional to the product $v_{osc}v_{sig}$. This will develop a voltage across the output tuned circuit, which is resonant at the intermediate frequency.

An exact analysis of the diode mixer is complicated by the fact that the voltage across the diode is the sum of the input and output voltages. However, by assuming that the output circuit impedance is negligible at the input frequencies, the voltage across the diode is approximately

$$v_d \cong V_{bias} + v_{osc} + v_{sig} \quad (5.10.4)$$

Assuming that the diode characteristic curve can be expanded in a Taylor's series, and terms up to the second need only be taken into account, the diode current is

$$i_d \cong av_d + bv_d^2 \quad (5.10.5)$$

Expansion of the squared term shows that it contains a product term and substituting from Eq.(5.10.3) gives for the peak value of IF current

$$I_{IF} \cong bV_{osc}V_{sig} \quad (5.10.6)$$

Assuming the transfer impedance of the output circuit is known at the IF, the peak output voltage at IF is

$$V_{IF} \cong bV_{osc}V_{sig}Z_T \quad (5.10.7)$$

A disadvantage of the diode mixer is its high *conversion loss*. The conversion gain of a mixer is the ratio of output power at IF to input power at the signal frequency, and conversion loss is the reciprocal of this. Also, the oscillator and signal circuits are not isolated from one another, this giving rise to the problem of oscillator radiation from the signal input. The harmonics of the signal and oscillator, as well as other products known as intermodulation products, appear at the output. One advantage of the diode mixer is that it generates low noise compared to transistor mixers. However, unless advantage has to be taken of its low-noise properties, the single diode mixer is seldom used for normal receiver applications.

Balanced diode modulators are described in Chapter 8. These essentially provide for the multiplication of two signals and can therefore also be used as mixers.

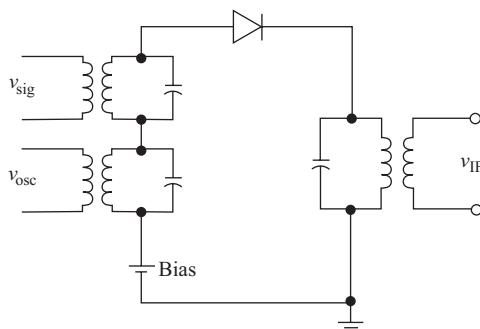


Figure 5.10.1 Diode mixer.

BJT Mixer

One circuit for the BJT mixer is shown in Fig. 5.10.2. Here, the signal voltage is applied between base and ground and the oscillator voltage between emitter and ground. The voltage/current relationship for the transistor is

$$I_c = I_s e^{V_{BE}/V_T} \quad (5.10.8)$$

where I_s is the transistor saturation current and V_{BE} the total base – emitter voltage, which is the algebraic sum of the dc bias, the signal, and oscillator voltages. As before, $V_T = 26$ mV at room temperature.

Expansion of the current equation shows that it contains a product term $v_{osc}v_{sig}$, which in turn contains the IF component of current. The expansion also shows that the dc level of collector current and hence the transconductance g_m is a function of both the signal and oscillator peak values. By keeping the signal amplitude small, the dependence on it can be made negligible, and by keeping the oscillator level constant, a constant effective g_m is achieved. Also, a large ($V_{osc} > 100$ mV) oscillator voltage is normally employed, and under these conditions the peak output current at the IF is given by

$$I_{IF} = g_c V_{sig} \quad (5.10.9)$$

where g_c is known as the *conversion transconductance* and is determined by the bias, and the oscillator peak voltage. Assuming that the transfer impedance of the collector output circuit is known at the IF, the output voltage is given by

$$\begin{aligned} V_{IF} &= I_{IF} Z_T \\ &= g_c V_{sig} Z_T \end{aligned} \quad (5.10.10)$$

The harmonics of the signal and oscillator frequencies and intermodulation terms also appear in the collector current as a result of the nonlinear transfer characteristic. Particularly troublesome are the components at frequencies $2f_{osc} - f_{sig}$ and $2f_{sig} - f_{osc}$. These are known as *third-order inter-modulation products*.

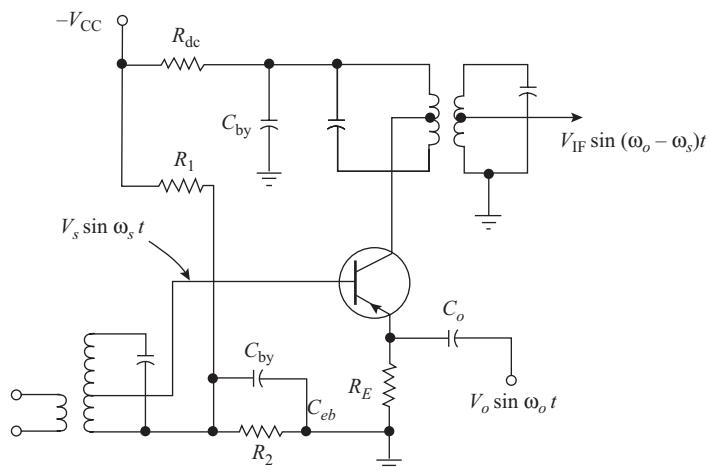


Figure 5.10.2 BJT mixer.

FET Mixers

For the ideal FET, the current/voltage transfer function for the constant current region (referred to as the saturation region for an FET) is given by

$$I_D = I_{DSS} \left(1 - \frac{V_{GS}}{V_P} \right)^2 \quad (5.10.11)$$

where I_D is the drain current, V_{GS} is the gate-source voltage, V_P the pinch-off voltage, and I_{DSS} the drain current for $V_{GS} = 0$. Both V_P and I_{DSS} are specified parameters for the transistor.

The square-law relationship for the ideal FET means that only terms up to the second order will be present in the output. These will contain a product term $v_{osc}v_{sig}$, which results in the IF component as before. One major advantage of the FET mixer over the BJT mixer is the very low level of third-order intermodulation products (for the *ideal* FET these would be absent). Also, the FET can handle a much wider range of input voltage, compared to the BJT. The circuit for an FET mixer is shown in Fig. 5.10.3.

The circuit for a dual-gate MOSFET is shown in Fig. 5.10.4(a). Good isolation between signal and oscillator circuits is provided with this arrangement since they are connected to different gates. The signal is

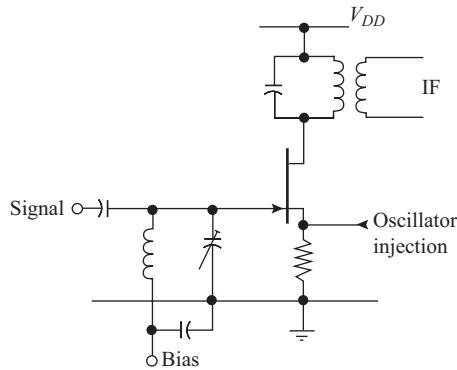


Figure 5.10.3 FET mixer.

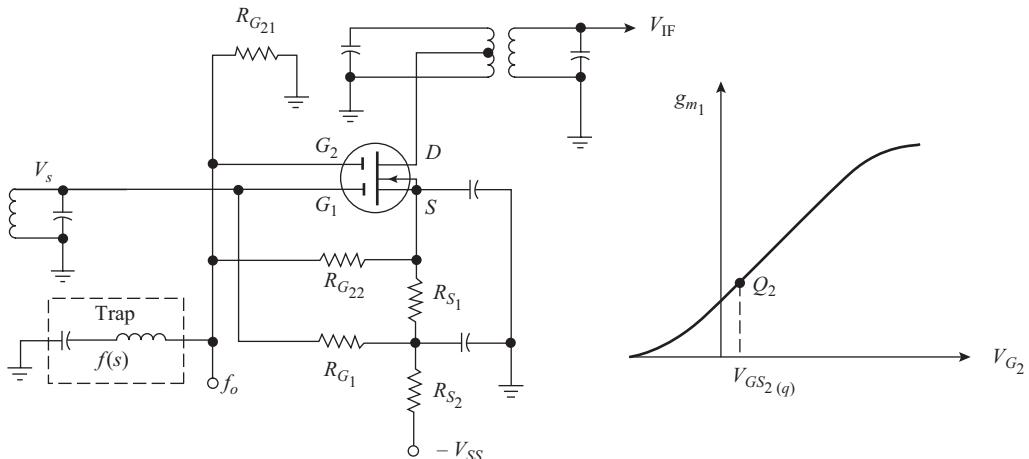


Figure 5.10.4 (a) Dual-gate FET mixer. (b) g_m1 transconductance as a function of gate 2 voltage.

normally applied to gate 1 because this provides greater gain. The oscillator voltage is applied to gate 2, through which it controls the transconductance referred to gate 1. The gate 1 transconductance is a function of the gate 2 voltage between cutoff and the saturation level, as shown in Fig. 5.10.4(b). To get some idea of the mixing process, assume that the function is approximately linear, of the form $g_{m1} = a + bv_{osc}$, where a and b are constants; then the ac component of drain current is $i_c = g_{m1}v_{sig}$. This is seen to contain the product term $bv_{osc}v_{sig}$ and hence an IF component of current. Some intermodulation products occur, and the operating point is chosen as a compromise between obtaining high conversion gain and low intermodulation products.

IC Balanced Mixer

The integrated-circuit (IC) balanced mixer is widely used in receiver ICs, as well as being available as a separate integrated circuit. The IC versions are usually described as *balanced modulators* since the modulation function is basically the same as the mixing function. Such circuits are well typified by the Motorola-type MC 1596. Figure 5.10.5(a) shows the basic circuit. The circuit may be operated in a variety of modes, and its operation here will be illustrated with reference to the large oscillator signal mode.

With a large oscillator level, the top transistors are switched alternately on and off. Transistors $Q1$ and $Q4$ are switched on and off together, at the same time as transistors $Q2$ and $Q3$ are switched off and on. The equivalent circuits for these on – off conditions are shown in Figs. 5.10.5(b) and (c).

The voltage equation for the signal input loop is $v_{sig} = V_{BE5} - V_{BE6} + i_R R_E$. Provided $I_Q \gg i_R$, the base-emitter voltages are approximately equal and

$$v_{sig} \approx i_R R_E \quad (5.10.12)$$

The current through the load impedance is $i_L = i_R S(t)$, where $S(t)$ is the switching function generated by the oscillator voltage. This has unity amplitude and can be expressed as a Fourier series (see Chapter 2):

$$S(t) = \frac{4}{\pi} \left(\sin \omega_{osc} t + \frac{1}{3} \sin 3\omega_{osc} t + \dots \right) \quad (5.10.13)$$

Thus

$$i_L = \frac{4V_{sig} \sin \omega_{sig} t}{\pi R_E} \left(\sin \omega_{osc} t + \frac{1}{3} \sin 3\omega_{osc} t + \dots \right) \quad (5.10.14)$$

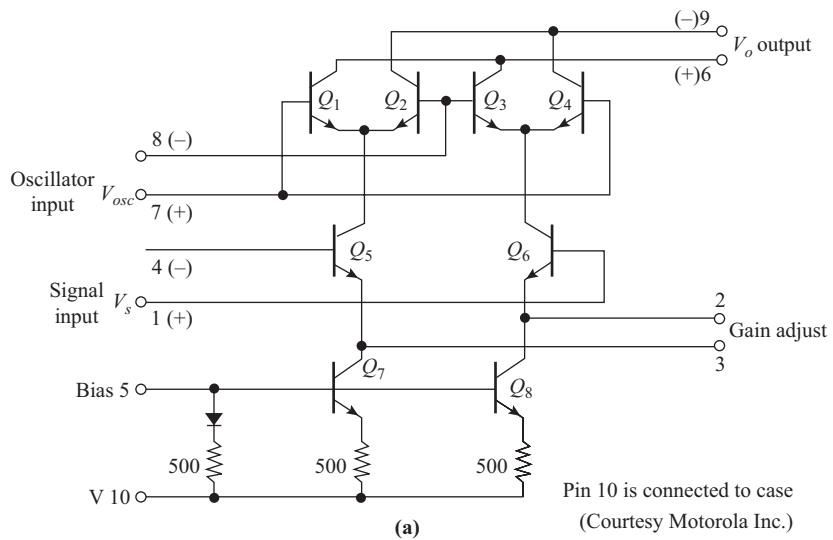
The load current i_L therefore contains the product $(\sin \omega_{sig} t)(\sin \omega_{osc} t)$, which can be expanded into sum and difference components as shown previously. The difference or IF component has a peak value

$$I_{IF} = \frac{2V_{sig}}{\pi R_E} \quad (5.10.15)$$

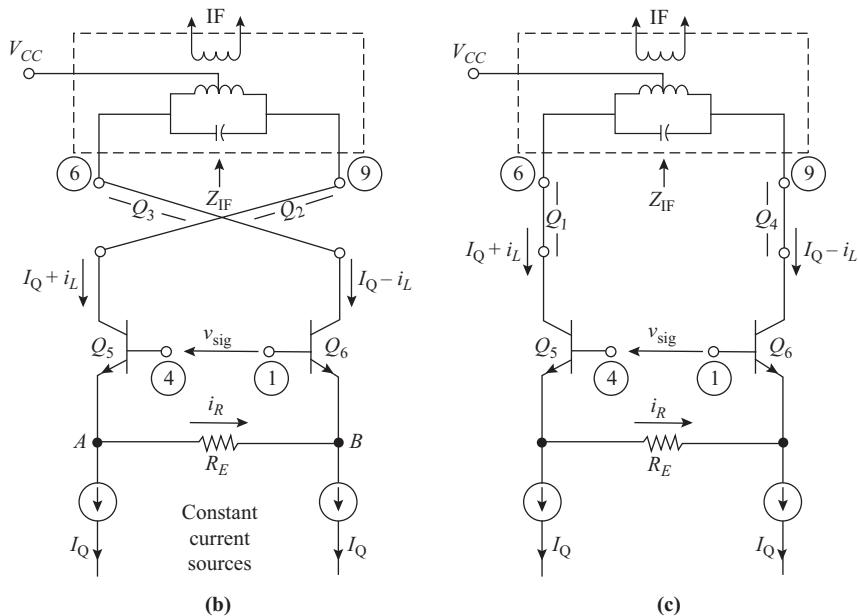
With the output circuit tuned to the IF and presenting a transfer impedance Z_T at this frequency, the output voltage is

$$V_{IF} = \frac{2V_{sig} Z_T}{\pi R_E} \quad (5.10.16)$$

The main advantage of the balanced circuit is that there are no components of output at signal and oscillator frequencies. Also, no intermodulation products occur, and the sum-frequency component, along with the higher harmonic terms associated with the switching signal, is easily filtered out.



(a)

Pin 10 is connected to case
(Courtesy Motorola Inc.)

(b)

(c)

Figure 5.10.5 (a) Balanced IC mixer. (b) and (c) The circuit when the oscillator voltage is a large switching signal.

5.11 Active Filters

Active filters make use of amplifiers along with resistors and capacitors in order to achieve frequency selective characteristics. They offer a number of advantages over the passive (RLC) filters described in Chapter 1. Active filters do not require inductors, which are physically large at low (for example, audio) frequencies and are thus unsuitable for use with compact designs utilizing integrated circuits. In addition,

active filters offer great versatility in design, programmable control of the characteristics being possible if required.

The disadvantages are that they require power supplies, they can introduce noise into the system, and, in the case of switched capacitor filters, the clocking signals can cause interference and distortion. However, these potential problems can all be avoided through proper design techniques.

RC Low-pass Filters

A simple *RC* low-pass filter is shown in Fig. 5.11.1(a) and an active filter version in Fig. 5.11.1(b). The transfer function for the *RC* filter of Fig. 5.11.1(a) is

$$\begin{aligned} H(f) &= \frac{V_{\text{out}}}{V_{\text{in}}} \\ &= \frac{1}{1 + j\omega RC} \end{aligned} \quad (5.11.1)$$

To be able to take the output voltage from the capacitor without loading it with the following circuit, a voltage follower is used, as shown in Fig. 5.11.1(b). Here, the operational amplifier circuit provides a very high input impedance (in the megohm range), a very low output impedance (a few tens of ohms), and a voltage gain of unity. The resistor R in the feedback path is to compensate for the dc offset developed across the input R . The response curve for the active filter is shown in Fig. 5.11.1(c).

The filter shown in Fig. 5.11.1 is a first-order Butterworth, a comparison with Eq. (1.13.2) showing that $m = 1$ and $f_c = 1/2\pi RC$, where f_c is the -3 -dB frequency.

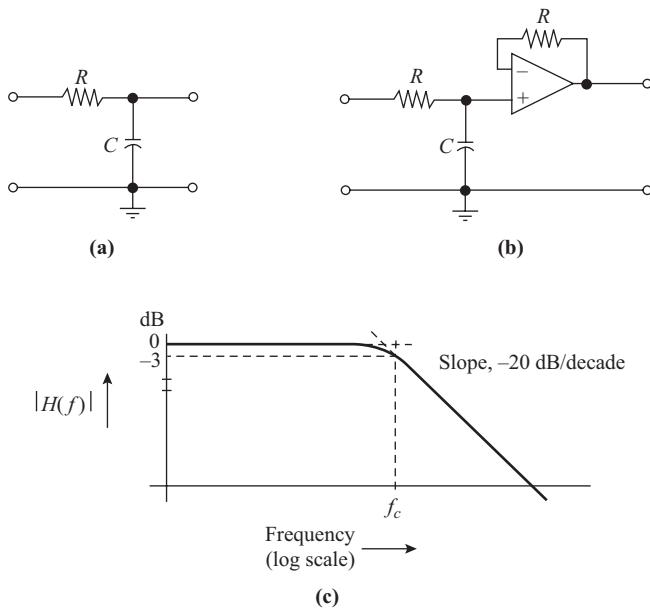


Figure 5.11.1 (a) First-order low-pass filter and (b) an active filter version. (c) Frequency response (magnitude) curve.

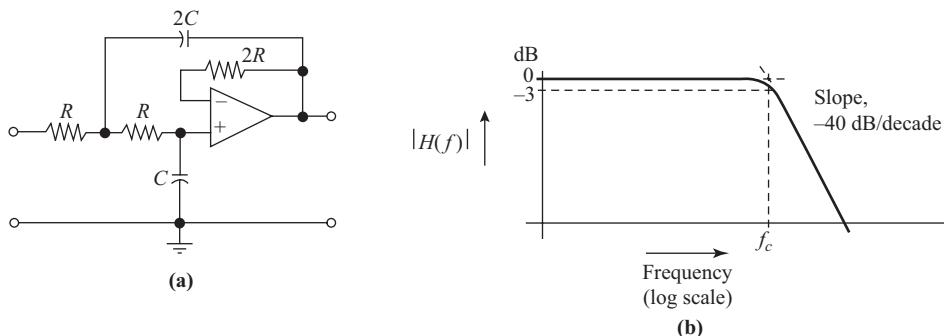


Figure 5.11.2 (a) Second-order low-pass filter and (b) frequency response (magnitude) curve.

A second-order Butterworth response can be achieved using the circuit of Fig. 5.11.2(a). For this filter, the -3-dB frequency is $f_c = 1/\sqrt{8}\pi RC$, and the transfer function is

$$H(f) = \frac{1}{1 - (f/f_c)^2 + j\sqrt{2}f/f_c} \quad (5.11.2)$$

The magnitude of the transfer function is

$$|H(f)| = \frac{1}{\sqrt{1 + (f/f_c)^4}} \quad (5.11.3)$$

Again, comparison with Eq. (1.13.2) shows $m = 2$.

In integrated-circuit realization of active filters, physical limits restrict resistor values to about a maximum of $10\text{ k}\Omega$, and capacitors to about 100 pF . Thus, for the second-order filter, the limit on the -3-dB frequency would be about 113 kHz . Reducing this to about 4 kHz as required for telephony applications would require larger values of R and/or C , which could not be implemented in IC form. The IC approach is solved by utilizing switched capacitor filters.

Switched Capacitor Filters

Figure 5.11.3(a) shows a capacitor connected to a voltage v through an MOS switch that is switched on and off by a square-wave clocking signal. Provision is also made to discharge the capacitor to ground through another MOS switch, which is operated 180° out of phase with the first one; that is, when one is switched on the other is off, and vice versa. The on period is denoted by T_c , which is also equal to the off period, and the clock frequency is $f_c = 1/2T_c$.

Consider the circuit with C initially discharged, switch 1 closed, and switch 2 open [Fig. 5.11.3(b)]. By making the clock frequency very much greater than the highest component of signal frequency, the voltage v will not change appreciably during the time T_c (even though it is a function of time), and the capacitor charges up to the voltage v . At the instant switch 1 is opened, switch 2 is closed, and the capacitor is discharged to zero in time T_c , ready for the next cycle [Fig. 5.11.3(c)].

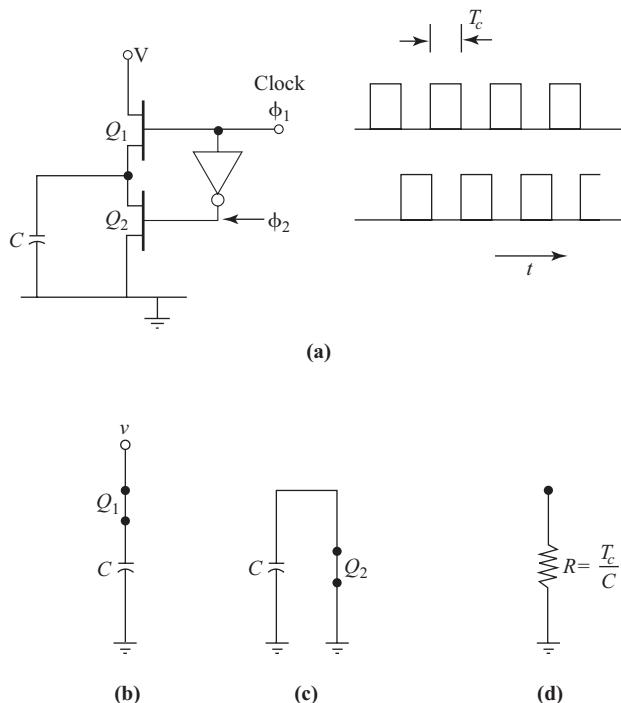


Figure 5.11.3 (a) Switched capacitor. (b) Switch 1 on, switch 2 off. (c) Switch 1 off, switch 2 on. (d) Equivalent resistance $R = T_C/C$.

Two equations can be written for the charge received by the capacitor during the charging period, $q = Cv$ and $q = iT_C$, where i is the average charging current during this period. Equating these two expressions for q gives the relationship between v and i as

$$v = i \frac{T_c}{C} \quad (5.11.4)$$

This may be interpreted as a form of Ohm's law, where resistance R is equivalent to

$$R = \frac{T_c}{C} = \frac{1}{2f_c C} \quad (5.11.5)$$

The factor 2 enters into this equation because the clock frequency is defined as $f_c = 1/2T_c$. (Some texts define this as $f_c = 1/T_c$ so that the equivalent R is shown as $R = 1/Cf_c$.)

This may seem like an unduly elaborate way of making a resistor R , but in IC technology capacitors and MOS switches can be made much smaller than physical resistors, and the circuit for the clocking signal is also readily fabricated as part of an IC.

An example of a first-order low-pass, switched capacitor filter is shown in Fig. 5.11.4(a). The discharge paths for the capacitors C_1 and C_2 are through the MOS switches Q_2 and Q_3 to the virtual ground at the

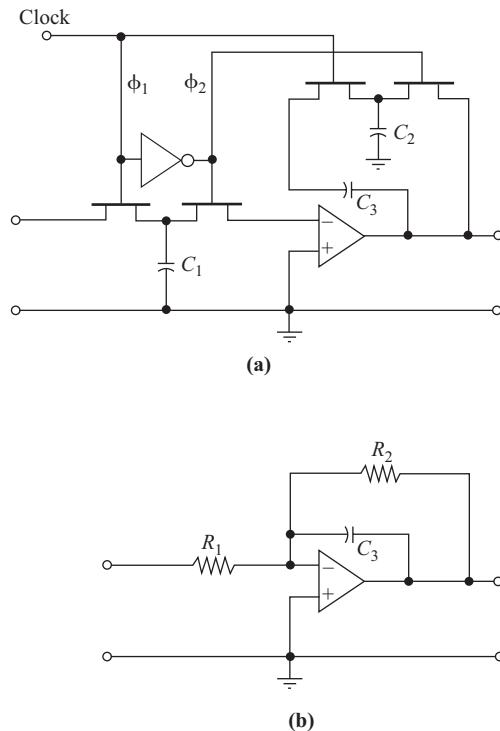


Figure 5.11.4 (a) Simple low-pass filter utilizing switched capacitors. (b) Equivalent circuit.

operational amplifier input. The switched capacitor branches can therefore be analyzed in the manner shown previously, resulting in an equivalent resistance in series with the input of $R_1 = T_c/C_1$ and in the feedback path $R_2 = T_c/C_2$. The equivalent circuit is shown in Fig. 5.11.4(b).

Analysis of the equivalent circuit yields for the transfer function

$$H(f) = -\frac{R_2}{R_1} \cdot \frac{1}{1 + j\omega R_2 C_3} \quad (5.11.6)$$

This is seen to provide an inverting gain of magnitude R_2/R_1 at low frequencies. By making $R_1 = R_2$ a first-order Butterworth response is obtained, but with the additional 180° phase shift resulting from the inverting gain.

The switched capacitor filter is a *sampled data* system. The Nyquist sampling theorem states in part that the sampling frequency must be at least twice the highest frequency in the analog signal being sampled. (The Nyquist theorem is discussed in Chapter 17, and applications have already been encountered in Sections 2.12 and 2.14 in connection with waveform analysis.) For present purposes, it should be noted that commercial switched capacitor circuits usually operate with a sampling frequency well above the Nyquist lower limit, 50 to 100 times being common. The Motorola MC145414 is an example of an IC switched capacitor filter. This unit contains two separate low-pass switched capacitor filters in addition to other circuitry, as shown in the block diagram of Fig. 5.11.5(a). A filter schematic is shown in Fig. 5.11.5(b), which includes one of the uncommitted operational amplifiers being utilized as a 60-Hz reject filter at the input of

BLOCK DIAGRAM

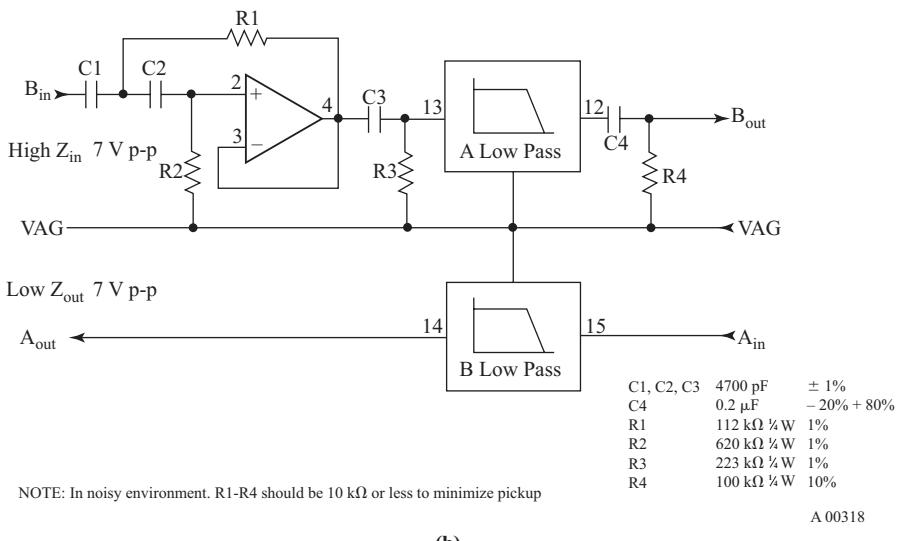
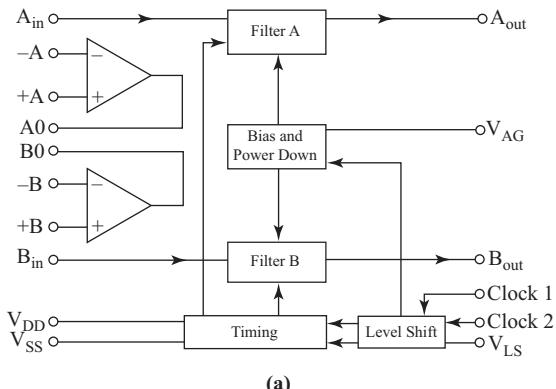


Figure 5.11.5 (a) Block diagram for the Motorola MC145414 switched capacitor filter. (b) Filter schematic, showing a 60-Hz reject filter at the input of switched capacitor filter A. (From Motorola Telecommunications Device Data DL 136, Rev. 2.)

the switched capacitor filter A. The components for this reject filter are added externally. The 60-Hz reject filter need not be added if not required, and switched capacitor filter B is shown without any external filtering.

The filters are fifth-order elliptic with response characteristics as shown in Fig. 5.11.6. The clock terminals 1 and 2 are tied together in practical circuits, and the response curves show the effects of changing the clock frequency. With the clock frequency at one of 256 kHz, 128 kHz, or 64 kHz, the low-frequency band limit is 7.2 kHz, 3.6 kHz, or 1.8 kHz, respectively.

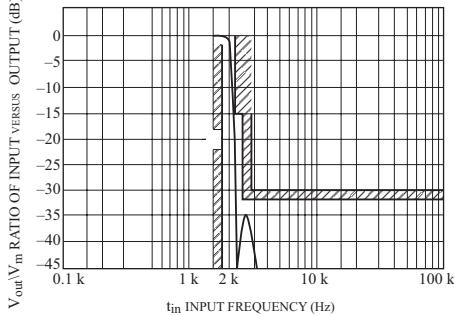
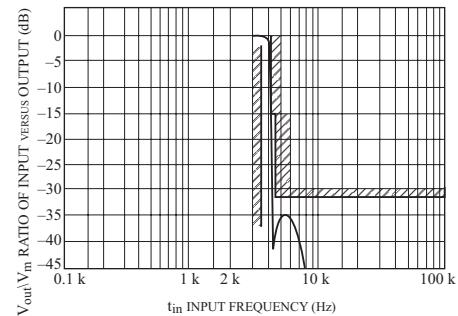
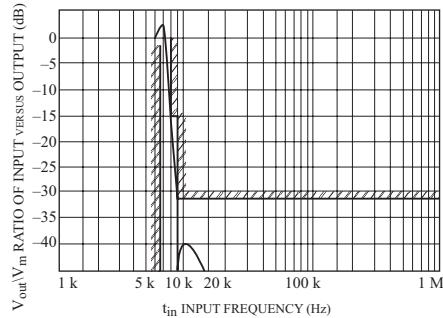
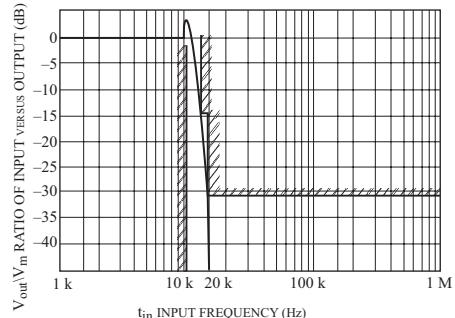
FIGURE 1 – FILTER A AND B LOWPASS CHARACTERISTICS
WITH CLOCK 1 AND 2 AT 64 KHZFIGURE 2 – FILTER A AND B LOWPASS CHARACTERISTICS
WITH CLOCK 1 AND 2 AT 128 KHZFIGURE 3 – FILTER A AND B LOWPASS CHARACTERISTICS
WITH CLOCK 1 AND 2 AT 256 KHZFIGURE 4 – FILTER A AND B LOWPASS CHARACTERISTICS
WITH CLOCK 1 AND 2 AT 400 KHZ

Figure 5.11.6 Response curves for the Motorola MC145414 switched capacitor filter. (From Motorola Telecommunications Device Data DL 136, Rev. 2.)

PROBLEMS

Assume $V_T = 26$ mV.

- 5.1. The following values apply for a BJT at $I_C = 2\text{mA}$: $\beta_o = 200$; $V_A = 750$ V; $\tau_F = 300$ ps; $C_{\text{depl}} = 3.5$ pF. Find the hybrid- π parameters g_m , r_c , $r_{b'e}$, C_{diff} , and $C_{b'e}$.
- 5.2. Repeat Problem 5.1 for $I_C = 5$ mA, assuming β_o increases to 230, while all the other transistor values remain constant.
- 5.3. Given that $C_{cb'} = 3$ pF, $C_c = 3$ pF, and $r_{bb'} = 70$ Ω , draw the hybrid- π equivalent circuit (corresponding to Fig. 5.2.1) for the transistors in Problems 5.1 and 5.2.
- 5.4. For a BJT working an $I_C = 500$ μA , the following values apply: $\beta_o = 150$; $C_{b'e} = 10$ pF; $C_{b'c} = 1.5$ pF. Plot the magnitude of the short-circuit current gain for the frequency range zero to 2 MHz. From this obtain an estimate of the -3 -dB bandwidth and compare with the calculated value.

- 5.5.** For the transistor in Problem 5.4, plot the phase shift as a function of frequency over the same range. Obtain an estimate of the phase shift at the -3 -dB frequency and compare with the value you expect from the theory.
- 5.6.** For the short-circuit current gain curve of Fig. 5.3.2, show that the slope at high frequencies is given by -6 dB per octave or, equivalently, -20 dB per decade.
- 5.7.** For a given transistor, $\beta_o = 150$ and $\omega_B = 10$ MHz. Determine the unity gain transition frequency.
- 5.8.** For a given transistor, $|A_{isc}| = 30$ dB at $f = 54$ MHz. Determine the unity gain transition frequency.
- 5.9.** The unity gain transition frequency for a BJT is 400 MHz at a mutual conductance of 30 mS. Given that the base – emitter depletion capacitance is 2 pF and the base – collector capacitance is 0.7 pF, calculate the forward transit time (a) assuming the bulk collector resistance is negligible, and (b) given that it is equal to $70\ \Omega$.
- 5.10.** Referring to Fig. 5.3.1, given that $r_{b'e} = 1.6\ k\Omega$, $C_{b'e} = 7.5\ pF$, $C_{cb'} = 1\ pF$, and $g_m = 25\ mS$, find i_c for $i_b = 3\ \mu A$.
- 5.11.** Repeat Problem 5.10 assuming $C_{cb'} = 0$.
- 5.12.** The following data apply to the tuned CE amplifier of Fig. 5.4.1(a): $r_c = 200\ k\Omega$, $R_L = 10\ k\Omega$, $r_2 = 5\ \Omega$, $L_2 = 50\ \mu H$, $C_2 = 100\ pF$, $C_{cb'} = 2\ pF$, $C_c = 1\ pF$, $g_m = 2\ mS$. Using Eq. 5.4.8, determine the voltage gain at resonance.
- 5.13.** Determine the Miller capacitance for the amplifier in Problem 5.12.
- 5.14.** For the amplifier of Problem 5.12, the input inductor is $50\ \mu H$ and $C_{b'e} = 12\ pF$. Determine the tuning capacitor C_1 required to make the input circuit resonant at the same frequency as the output circuit.
- 5.15.** For the circuit of Fig. 5.4.1(a): $R_s = 5\ k\Omega$, $r_{b'e} = 65\ k\Omega$, $r_1 = 5\ \Omega$, $L_1 = 50\ \mu H$, $C_1 = 100\ pF$, $C_{b'e} = 12\ pF$, $C_{cb'} = 1\ pF$, $g_m = 2\ mS$, $R_{D2\ eff} = 8\ k\Omega$. Determine, for the input circuit, (a) the resonant frequency, and (b) the effective Q -factor.
- 5.16.** For the circuit in Problem 5.15, determine for resonance (a) the terminal voltage gain, and (b) the voltage gain referred to the source emf.
- 5.17.** Transistor 2N2222A is used as a common-emitter amplifier with a tuned collector circuit mutually inductively coupled to a secondary load of $200\ \Omega$. The tuning capacitance (including the transistor output capacitance) is 119 pF, the primary inductance is $5\ \mu H$, the secondary inductance is $0.1\ \mu H$ and the coefficient of coupling is 0.2. The unloaded Q -factor of the primary tuned circuit is 100, and secondary resistance may be assumed negligible. The transistor operates at a collector current of 8.24 mA, and the other parameters are $V_A = 45.6$ V and $\beta_o = 100$. Assuming $r_{b'b}$ is negligible, calculate the voltage gain of the stage referred to the input terminals, and compare with the results presented in Fig. 5.4.4.
- 5.18.** Rework Example 5.4.3 for a frequency of 20 MHz, assuming all other values remain unchanged.
- 5.19.** Rework Example 5.4.3 for $Z_T = 300 + j250\ \Omega$, assuming all other values remain unchanged.
- 5.20.** Discuss the factors that can give rise to instability in a CE amplifier and how the instability may be avoided.
- 5.21.** The input circuit for a CE amplifier can be represented by the parallel circuit impedance given by Eq. (1.4.2). The Q -factor is 70, and the dynamic resistance is $2\ k\Omega$. The resonant frequency of the circuit is 5 MHz. Calculate the equivalent parallel inductance of the circuit at a frequency of 4 MHz.
- 5.22.** Repeat the calculations of Problem 5.21 for the output impedance of the amplifier, given that its Q -factor is 85 and dynamic impedance is $5\ k\Omega$.
- 5.23.** For the amplifier in Problem 5.22, the transistor output resistance is $r_c = 75\ k\Omega$. Calculate the minimum value of g_m necessary to start oscillation at 4 MHz.

- 5.24.** For a certain BJT, the bulk base resistance may be assumed negligible, and the other component values are: $\beta_o = 150$, $C_{be} = 13 \text{ pF}$, $C_{cb} = 1.5 \text{ pF}$, and $C_{cs} = 1.5 \text{ pF}$. The transistor is operated at a collector current of $750 \mu\text{A}$. For short-circuit output conditions, calculate (a) the -3-dB frequency for the CE current gain, (b) the unity-gain transition frequency, and (c) the -3-dB frequency for the CB current gain. Compare the results of (b) and (c).
- 5.25.** For the transistor in Problem 5.24, calculate the short-circuit current gains for both the CE and CB connections at frequencies $\omega = 0.003 \omega_T$ and $0.03 \cdot \omega_T$.
- 5.26.** The transistor and output circuit of Problem 5.12 are reconnected as a CB amplifier, the specified values remaining unchanged, and $\beta_o = 150$. Calculate (a) the CB output resistance of the transistor, and (b) the voltage gain at resonance. Compare the results with those of Problem 5.12.
- 5.27.** A BJT is operated at $g_m = 2 \text{ mS}$ and $\beta_o = 150$. Calculate the available power gains for (a) the CE and (b) the CB configurations, given that the source resistance is 300Ω and the output resistance is $10 \text{ k}\Omega$ for each.
- 5.28.** A BJT is operated at $g_m = 0.5 \text{ mS}$ and $\beta_o = 150$. Calculate the ratio of available power gains for CE and CB for the following values of source resistance: (a) 0Ω ; (b) 50Ω ; (c) $5 \text{ k}\Omega$.
- 5.29.** A cascode amplifier employing BJTs works at a collector current of $500 \mu\text{A}$ and $\beta_o = 200$ for each transistor. The output tuned circuit has a Q -factor of 100, and the total tuning capacitance is 47 pF . Determine (a) the voltage gain, (b) the input resistance, and (c) the power gain. The source resistance is 50Ω and the resonant frequency is 148 MHz.
- 5.30.** An FET CS amplifier has identical tuned input and output circuits for which $L = 10 \mu\text{H}$ and $Q = 100$ at the resonant frequency of 10.7 MHz without the transistor in circuit. The signal source resistance is $R_s = 1000 \Omega$ and the transistor values are $g_m = 1.3 \text{ mS}$, $r_d = 25 \text{ k}\Omega$, $C_{ds} = 2 \text{ pF}$, and $C_{gs} = 5 \text{ pF}$. Assuming $C_{dg} = 0$, calculate (a) the external capacitors C_1 and C_2 required to maintain the amplifier resonant frequency at 10.7 MHz, (b) the dynamic resistance of the input and output circuits, excluding the damping effects of the transistor and the source, (c) the input and output admittances including the damping effects of the transistor and the source, and (d) the voltage gain referred to the source emf.
- 5.31.** Repeat Problem 5.30 with $C_{dg} = 0.7 \text{ pF}$, assuming that the input and output circuits are tuned to 10.7 MHz.
- 5.32.** An FET CS amplifier utilizes identical tuned input and output circuits for which $R_D = 67 \text{ k}\Omega$ and $Q = 100$ at the resonant frequency of 10.7 MHz. Resonance includes C_{gs} and C_{ds} , but excludes the damping effects of source and transistor. For the transistor, $r_d = 25 \text{ k}\Omega$ and $g_m = 1.3 \text{ mS}$, and it may be assumed that $C_{dg} = 0$. The source resistance is 1000Ω . Plot the magnitude and phase angle of the voltage gain referred to source emf for a frequency range of approximately $\pm 1\%$ about the resonant frequency.
- 5.33.** Repeat Problem 5.32 for $C_{dg} = 0.7 \text{ pF}$, but use a frequency range of 10.2 to 10.4 MHz. From the graphs, determine the new resonant frequency for this situation and explain why it is different from that in Problem 5.32.
- 5.34.** For a diode mixer, the b coefficient is 0.03 S/V, the peak oscillator voltage is 10 mV, and the peak signal voltage is 1 mV. The transfer impedance of the output circuit at IF is 1000Ω . Determine the peak output voltage at IF.
- 5.35.** For a BJT mixer, the conversion transconductance is 5 mS and the transfer impedance of the output circuit at IF is 1000Ω . Determine the peak output voltage at IF for an input signal voltage of 1 mV.
- 5.36.** Given that the gate-source voltage to an FET mixer consists of sinusoidal signal and oscillator voltages in series, determine an expression for the peak IF current in terms of the peak voltages and I_{DSS} and V_P of the transistor. Hence, derive an expression for the conversion transconductance of the mixer.

- 5.37.** For a FET mixer, $I_{DSS} = 15 \text{ mA}$ and $V_P = -4 \text{ V}$. Plot the output transfer characteristic for a gate–source voltage range of -4 V to 2 V . Calculate the conversion transconductance given that the peak oscillator voltage is one-half the magnitude of the pinch-off voltage. Calculate also the peak IF current for a peak signal voltage of 1 mV .
- 5.38.** For the balanced mixer of Fig. 5.10.5, $R_E = 1 \text{ k}\Omega$. Calculate the peak IF current for (a) fundamental and (b) third harmonic mixing for a peak signal of 1 mV .
- 5.39.** For the filter transfer function of Eq. (5.11.2), given that $R = 5 \text{ k}\Omega$ and $C = 57 \text{ pF}$, determine (a) the value of the transfer function at the -3-dB frequency. Calculate also the magnitude and phase angle of the transfer function at (b) 400 kHz and (c) 800 kHz .
- 5.40.** Derive Eq. (5.11.6). Given that $T_C = 100 \mu\text{s}$, $C_1 = 57 \text{ pF}$, $C_2 = 100 \text{ pF}$, and $C_3 = 75 \text{ pF}$, calculate the values of R_1 and R_2 . Plot the magnitude and phase of the filter response for f in the range $0.1 f_c$ to $10 f_c$ where f_c is the -3-dB frequency.
- 5.41.** Derive an expression for the -3-dB frequency for the switched capacitor filter of Fig. 5.11.4. For the filter, $T_C = 100 \mu\text{s}$, $C_1 = 47 \text{ pF}$, $C_2 = 75 \text{ pF}$, and $C_3 = 47 \text{ pF}$. Calculate (a) the -3-dB frequency. Calculate also the magnitude and phase of the relative response at frequencies (b) 400 Hz and (c) 800 Hz .
- 5.42.** Applying the *hybrid- π* equivalent circuit, plot the frequency response of a common emitter transistor amplifier using MATLAB.
- 5.43.** Applying the *hybrid- π* equivalent circuit, plot the frequency response of an FET amplifier.
- 5.44.** Derive the transfer function of the second order low pass filter given in Figure 5.11.2(a). Plot its frequency response using MATLAB.
- 5.45.** Show how the circuit given in Figure 5.11.2(a) can be modified to make it a second order high pass filter. Plot the response.
- 5.46.** Derive the transfer function of the circuit shown in Figure 5.11.4(b). Show that it is an LPF.
- 5.47.** Draw the circuit diagram of a *Universal filter* using three OPAMPS. Show how it can be converted to an oscillator.



Oscillators

6.1 Introduction

Electronic communications systems could not operate without sources of sinusoidal electrical waves. Many types of oscillator circuits are used to produce these sinusoids, and a few of the more commonly used ones are analyzed on the following pages, in order to illustrate the general method.

Feedback oscillators, *RC* and tuned *LC* types, are discussed first. These circuits are applicable to frequencies from the audio range up to the VHF range and may use any convenient three-terminal amplifying device. Discussion here is limited to the bipolar transistor and the field-effect transistor. The crystal-controlled oscillator is included as a tuned *LC* type, and a discussion of the factors affecting the frequency stability of oscillators is included.

The principles of voltage-controlled oscillators (VCOs) are presented.

Although it is not an oscillator in its own right, the frequency synthesizer is included here because it is revolutionizing the frequency-control scene in communications, making possible complex systems that until recently have not been economically feasible.

6.2 Amplification and Positive Feedback

The oscillators to be considered in this chapter can be modeled as amplifiers with positive feedback, illustrated in Fig. 6.2.1. Any small disturbance at the input to the amplifier, such as caused by noise, or a switching-on transient, will be amplified, and part of the amplified signal is fed back to the input. Providing the feedback signal has sufficient amplitude and is of the correct phase, the process can result in the buildup of a self-sustaining signal, or oscillation. The purpose of analysis is to establish the amplitude and phase conditions necessary for oscillation.

In Fig. 6.2.1 the input v_i is multiplied by the forward gain A to give the output v_o . A fraction B of this is fed back to provide the input v_i . Thus, $ABv_i = v_i$ or $AB = 1$ is the condition required to sustain oscillation. This is known as the *Barkhausen criterion*.

In practice, the usual conditions are that the amplification A is frequency independent and incorporates a 180° phase change. In the closed-loop condition, the magnitude of A must equal the magnitude of $1/B$ in order to sustain oscillations. The feedback network B is made up of passive components, which are the

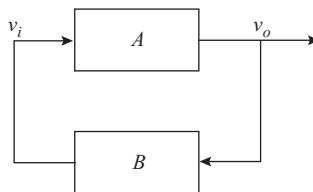


Figure 6.2.1 Oscillator block schematic.

frequency-determining elements. The feedback network also introduces a further 180° phase shift, so the total phase shift is 360° around the closed loop.

Small-signal analysis is generally used to establish the starting conditions for oscillation and the frequency at which oscillation takes place. It will be recalled that small-signal analysis utilizes impedance and admittance concepts defined for sinusoidal waveforms. Thus the analysis yields the sinusoidal frequency of oscillation.

It will also be recalled that sinusoidal analysis implies steady-state conditions. In practice, the oscillations will pass through a transient stage from start-up to final steady state. In the final steady state, the transistor is usually working under large-signal conditions, so the small-signal parameters have no real meaning under these conditions. What is important is that the small-signal analysis yields the minimum conditions necessary for oscillations to be maintained and shows the dependence of the frequency on the circuit parameters.

It should be noted that since the feedback circuit is closed the Barkhausen equation $AB = 1$ is valid at all times. The feedback network is a passive circuit, and so the amplifier gain must change automatically to maintain $A = 1/B$ as the oscillation builds up to the steady-state conditions.

In modeling the amplifier for small-signal conditions, one or other of the circuits shown in Fig. 6.2.2 will be used. These are based on the small-signal hybrid- π model introduced in Chapter 6. The input impedance Z_i will be the input impedance of the transistor in parallel with the input bias components. The output impedance Z_o is the output impedance of the transistor in parallel with the output bias components.

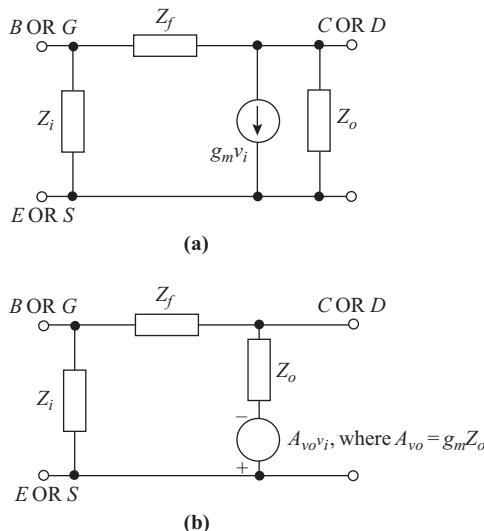


Figure 6.2.2 Equivalent small-signal amplifier circuits utilizing (a) a voltage-dependent current generator and (b) a voltage-dependent voltage generator.

The feedback impedance Z_f is that which exists within the amplifier, along with that provided by the external feedback network B .

The circuit of Fig. 6.2.2(b) is readily obtained from that of Fig. 6.2.2(a) by converting the current (Norton) source to the Thevenin equivalent voltage source.

6.3 RC Phase Shift Oscillators

The circuit for an RC phase shift oscillator utilizing a BJT is shown in Fig. 6.3.1(a), and the equivalent small-signal circuit in Fig. 6.3.1(b). The amplifier portion is a common-emitter amplifier that introduces a 180° phase shift, and the function of the RC network is to introduce a further 180° phase shift so that the $AB = 1$ condition can be realized.

Because this type of oscillator is used mainly in the audio-frequency range, the feedback path in the BJT, consisting of the collector to base capacitance, presents a very high impedance and can be ignored. Likewise, the input capacitance and the input biasing network are ignored, the input impedance being essentially the base to emitter small-signal resistance r_{be} . Also, the output capacitance of the amplifier may be ignored, and the output impedance R_o is the collector biasing resistor R_C in parallel with the small-signal output resistance r_c .

For ease of design, the RC sections are made identical so that the final resistance consists of $R_1 + r_{be} = R$. In practice, R is made sufficiently large for R_1 to be much greater than r_{be} , in order to minimize the effect of r_{be} on the oscillations.

As discussed in Chapter 5, for the BJT $r_{be} = \beta_o/g_m$, and since $v_{be} = i_b r_{be}$, the current source can be rewritten as $g_m v_{be} = \beta_o i_b$. Figure 6.3.1(c) shows the equivalent circuit used for the small-signal analysis.

Denoting $Z = R - j/\omega C$, the three mesh equations for Fig. 6.3.1(c) are

$$-\beta_o i_b R_o = (R_o + Z)i_1 - Ri_2 + 0i_b \quad (6.3.1)$$

$$0 = -Ri_1 + (Z + R)i_2 - Ri_b \quad (6.3.2)$$

$$0 = 0i_1 - Ri_2 + (Z + R)i_b \quad (6.3.3)$$

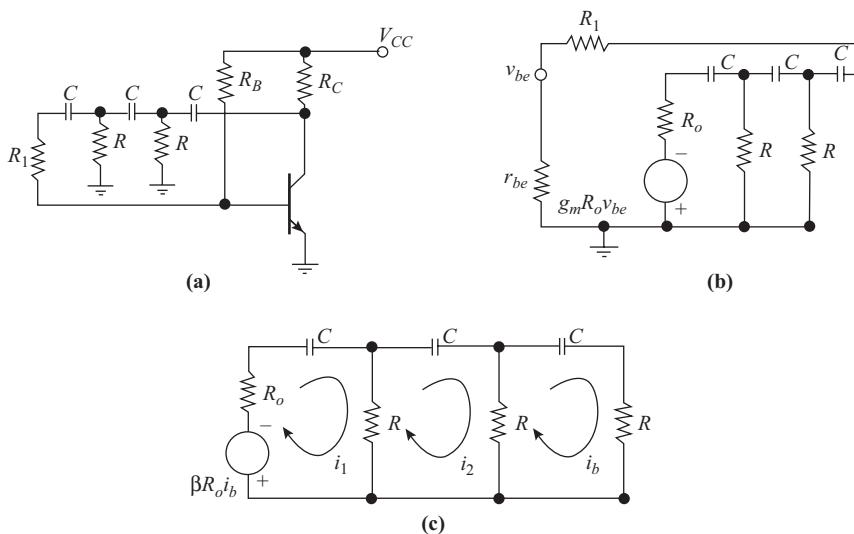


Figure 6.3.1 (a) RC phase shift oscillator, (b) equivalent circuit, and (c) simplified equivalent circuit.

Solving these equations for i_b gives

$$i_b = \frac{-\beta_o i_b R^2 R_o}{\Delta} \quad (6.3.4)$$

where

$$\Delta = (R_o + Z)(Z^2 + 2RZ) - R^2(Z + R) \quad (6.3.5)$$

From equation 6.3.4 on, cancelling i_b from both sides gives the equivalent equation to the $AB = 1$ condition for the RC phase shift oscillator as $1 = -\beta_o R^2 R_o / \Delta$. Expanding this gives

$$(R_o + Z)(Z^2 + 2RZ) - R^2(Z + R) + R^2 R_o \beta_o = 0 \quad (6.3.6)$$

Because complex impedances are involved, this is really two equations in which real and imaginary parts must be equated separately. The imaginary parts yield, after some lengthy algebraic manipulation,

$$\omega = \frac{1}{CR\sqrt{6 + 4R_o/R}} \quad (6.3.7)$$

Equating the real parts and substituting for ω from Eq. (6.3.7) gives

$$\beta_o = 23 + \frac{4R_o}{R} + \frac{29R}{R_o} \quad (6.3.8)$$

This is the minimum value that the transistor beta must have in order for oscillations to start.

It will be seen that the beta value required is dependent on the ratio R_o/R and its reciprocal R_o/R . It is left as an exercise for the student to show that β_o has a minimum value when $R_o/R = 2.7$.

EXAMPLE 6.3.1

Determine the minimum beta required for a BJT RC phase shift oscillator, for which the small-signal output resistance is $40 \text{ k}\Omega$, and the collector bias resistor is $10 \text{ k}\Omega$. The required operating frequency is 400 Hz . Determine also the individual R and C values.

SOLUTION For minimum beta,

$$\frac{R_o}{R} = 2.7$$

Hence

$$\beta_{o \min} = 23 + 4 \times 2.7 + \frac{29}{2.7} = 44.5$$

$$R_o = \frac{r_c R_c}{r_c + R_c} = 8 \text{ k}\Omega$$

$$\therefore R = 2.7 \times R_o = 21.6 \text{ k}\Omega$$

From the frequency equation

$$C = \frac{1}{2 \times \pi \times 400 \times 21,600 \times \sqrt{6 + 4 \times 2.7}} \cong 450 \text{ pF}$$

It will be seen from this example that R is not very much greater than R_o , so the frequency of operation will be influenced by changes that might occur in R_o . Also r_{be} will be a significant part of the final R value in the feedback network, which means that the feedback will be affected by changes in r_{be} . Choosing a larger value of R minimizes these effects, as shown in the following example.

EXAMPLE 6.3.2

Determine suitable component and device parameter values for an RC phase shift oscillator required to oscillate at a frequency of 800 Hz. The active device is a BJT, and the output resistance, including the collector bias resistor, is 18 k Ω .

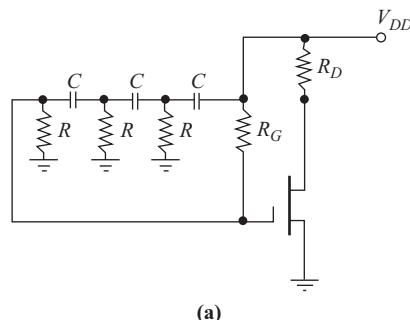
SOLUTION $R \gg R_o$ should be chosen to minimize the effect of R_o on frequency. Also, large R should result in a large R_1 for typical values of r_{be} , thus minimizing any dependence of oscillation conditions on r_{be} . A number of values for R can be tried, and it will be found that $R = 100$ k Ω is reasonable. C may now be determined from the frequency equation as

$$C = \frac{1}{2 \times \pi \times f \times R \times \sqrt{6 + 4R_o/R}} = 767 \text{ pF}$$

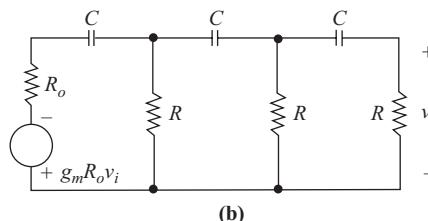
The required transistor beta is

$$\beta_o = 23 + 4 \frac{R_o}{R} + 29 \frac{R}{R_o} \approx 185$$

A similar analysis technique can be applied for FETs, one difference being that, because of the high input resistance of an FET, the final resistor in the RC chain is simply R taken to ground, as shown in Fig. 6.3.2(a). Also, in this case it is best to work in terms of the device g_m since current gain has no particular meaning for an FET.



(a)



(b)

Figure 6.3.2 (a) RC phase shift oscillator utilizing an FET as the active device. (b) Equivalent circuit.

The analysis for the FET circuit follows along the same lines as for the BJT version, and the result for the frequency of oscillation is the same. The starting conditions require that

$$g_m R = 23 + \frac{4R_o}{R} + \frac{29R}{R_o} \quad (6.3.9)$$

This can be rearranged as

$$g_m R_o = 29 + 23 \frac{R_o}{R} + 4 \left(\frac{R_o}{R} \right)^2 \quad (6.3.10)$$

The reason for writing the equation in this latter form is that $g_m R_o$ is the open-circuit voltage gain of the amplifier section. Again, keeping in mind that because the oscillator is a closed circuit, if oscillations occur at all, the equation applies at all phases of the oscillation buildup and steady state. What happens is that the bias currents will automatically adjust so that the equation conditions are met, but this means that the transistor parameters g_m and r_o will change as the oscillation level changes.

EXAMPLE 6.3.3

An FET RC phase shift oscillator is required to generate a frequency of 1000 Hz. The output resistance of the FET amplifier is $5\text{ k}\Omega$. Determine the values of the other components.

SOLUTION

Choose $R \gg R_o$ to minimize the effects of R_o on frequency. Try $R = 100\text{ k}\Omega$ to get

$$C = \frac{1}{2 \times \pi \times 10^3 \times 10^5 \times \sqrt{6 + 4 \times \frac{5}{100}}} = 639 \text{ pF}$$

The required open-circuit voltage gain is

$$A_o = 29 + 23 \times \frac{5}{100} + 4 \times \left(\frac{5}{100} \right)^2 \approx 30$$

Hence

$$g_m = \frac{30}{5000} \approx 6 \text{ mS}$$

6.4 LC Oscillators

LC circuits provide the most convenient means of achieving oscillation at high (radio) frequencies. Before analyzing a few of the more common circuits, the general form will be examined. Figure 6.4.1 shows a BJT with a tuned circuit feedback network consisting of Z'_1 , Z'_2 , and Z'_3 . The small-signal equivalent circuit is shown in Fig. 6.4.1(b). It will be seen that the transistor impedances can be combined in parallel with the feedback impedances, resulting in the simplified equivalent circuit of Fig. 6.4.1(c).

The nodal equation for the output node is

$$-g_m v_1 = v_2 \left(\frac{1}{Z_3} + \frac{1}{Z_2} \right) - \frac{v_1}{Z_3} \quad (6.4.1)$$

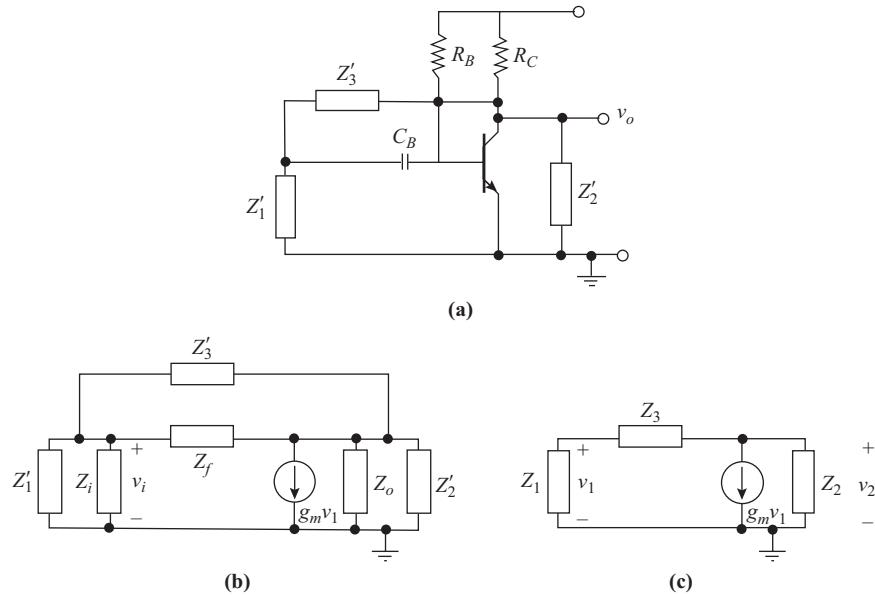


Figure 6.4.1 (a) General form of an LC oscillator. (b) Equivalent circuit. (c) Simplified equivalent circuit.

Also, v_1 is seen to be

$$v_1 = v_2 \frac{Z_1}{Z_1 + Z_3} \quad (6.4.2)$$

Combining these two equations and simplifying yields

$$-g_m Z_1 Z_2 = Z_1 + Z_2 + Z_3 \quad (6.4.3)$$

This is equivalent to the general condition stated previously, $AB = 1$.

In some situations it is easier to work with admittances Y_1 and Y_2 rather than impedances Z_1 and Z_2 . Dividing through Eq. (6.4.3) by $Z_1 Z_2$ gives

$$-g_m = Y_1 + Y_2 + Z_3 Y_1 Y_2 \quad (6.4.4)$$

Colpitt Oscillator

The circuit for the Colpitt's oscillator utilizing a BJT is shown in Fig. 6.4.2(a) and the equivalent circuit in Fig. 6.4.2(b). For the Colpitt's oscillator, since the components at the input and output are connected in parallel, Eq. (6.4.4) is the more convenient one to use [see Fig. 6.4.2(c)]. For the moment, the admittances will be expressed generally as $Y_1 = G_1 + jB_1$ and $Y_2 = G_2 + jB_2$, while the feedback impedance is $Z_3 = r + j\omega L$. It is assumed that feedback through G_{cb} is negligible in comparison to that through Z_3 . Thus Eq. (6.4.3) becomes

$$-g_m = G_1 + jB_1 + G_2 + jB_2 + (r + j\omega L)(G_1 + jB_1)(G_2 + jB_2) \quad (6.4.5)$$

The real and imaginary parts of this have to be equated separately. The imaginary part is

$$B_1 + B_2 + \omega L(G_1 G_2 - B_1 B_2) + r(B_1 G_2 + B_2 G_1) = 0 \quad (6.4.6)$$

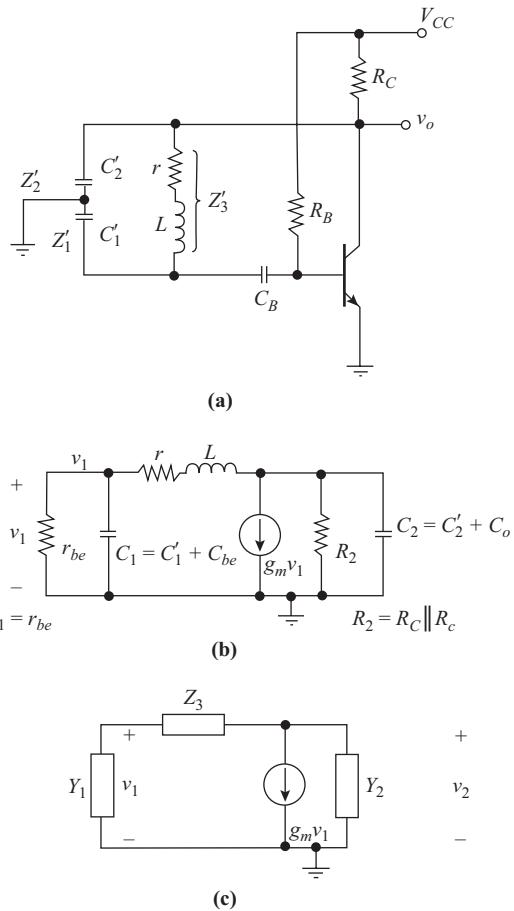


Figure 6.4.2 (a) Colpitt's oscillator. (b) Small-signal equivalent circuit. (c) Simplified small-signal equivalent circuit.

Now, $B_1 = \omega C_1$ and $B_2 = \omega C_2$ and algebraic manipulation of the imaginary part yields

$$\omega^2 L C_1 C_2 = C_1 \left(1 + \frac{r}{R_2} \right) + C_2 \left(1 + \frac{r}{R_1} \right) + \frac{L}{R_1 R_2} \quad (6.4.7)$$

The substitutions $G_1 = 1/R_1$ and $G_2 = 1/R_2$ have been made. Now, since $R_1 (= r_{be})$ and $R_2 (= R_o = R_c \parallel r_c)$ are generally in the kilohm range while r is a few tens of ohms at most, it is seen that this equation reduces to

$$\omega^2 L C_1 C_2 = C_1 + C_2 + \frac{L}{R_1 R_2} \quad (6.4.8)$$

This can be put into the form

$$\omega^2 = \frac{1}{LC_s} + \frac{1}{R_1 R_2 C_1 C_2} \quad (6.4.9)$$

Where C_s is the series combination of C_1 and C_2 . Further simplification requires a knowledge of component values. In general, values will be chosen to reduce the dependence of the frequency of oscillation on the transistor parameters so that the second term on the right-hand side becomes negligible. Hence

$$\omega^2 \cong \frac{1}{LC_s} \quad (6.4.10)$$

The real part of the general expression is

$$\begin{aligned} -g_m &= G_1 + G_2 + r(G_1G_2 - B_1B_2) - \omega L(B_1G_2 + B_2G_1) \\ &= G_1 + G_2 + r(G_1G_2 - \omega^2 C_1C_2) - \omega^2 L(C_1G_2 + C_2G_1) \end{aligned} \quad (6.4.11)$$

This gives the maintenance conditions in terms of the frequency. The equation can be solved in the form given, or the substitution $\omega^2 \cong 1/LC_s$ may be made. Also, the dynamic impedance of the tuned circuit is given by $R_D = L/C_s r$, and substituting this, along with $G_1 = 1/R_1$ and $G_2 = 1/R_2$, yields, after considerable manipulation,

$$g_m = \frac{1}{R_1} \left(\frac{C_2}{C_1} - \frac{r}{R_2} \right) + \frac{1}{R_2} \frac{C_1}{C_2} + \frac{1}{R_D} \left(\frac{C_1}{C_2} + \frac{C_2}{C_1} + 2 \right) \quad (6.4.12)$$

Again, for the usual condition that $R_2 \gg r$, this reduces to

$$g_m = \frac{1}{R_1} \left(\frac{C_2}{C_1} \right) + \frac{1}{R_2} \frac{C_1}{C_2} + \frac{1}{R_D} \left(\frac{C_1}{C_2} + \frac{C_2}{C_1} + 2 \right) \quad (6.4.13)$$

For the BJT, $R_1 \cong r_{be} = \beta_o/g_m$, and hence the condition for start of oscillation becomes

$$g_m \left(1 - \frac{1}{\beta_o} \frac{C_2}{C_1} \right) R_2 = \frac{C_1}{C_2} + \left(\frac{C_1}{C_2} + \frac{C_2}{C_1} + 2 \right) \frac{R_2}{R_D} \quad (6.4.14)$$

Since C_1 and C_2 are of the same order of magnitude, then assuming the transistor beta is reasonably large (> 50), this reduces to

$$g_m R_2 \cong \frac{C_1}{C_2} + \frac{R_2}{R_D} \left(\frac{C_1}{C_2} + \frac{C_2}{C_1} + 2 \right) \quad (6.4.15)$$

EXAMPLE 6.4.1

Circuit values for a BJT Colpitt's oscillator are $L = 400 \mu\text{H}$, $C_1 = 100 \text{ pF}$, and $C_2 = 300 \text{ pF}$. The inductor Q -factor is 200, and for the transistor amplifier, $R_o = 5 \text{ k}\Omega$ and $\beta_o = 100$. Determine the transistor g_m for startup conditions and the frequency of oscillation.

SOLUTION The tuning capacitance is

$$C_s = \frac{C_1 C_2}{C_1 + C_2} = 75 \text{ pF}$$

An estimate of the frequency of oscillation is obtained as

$$f = \frac{1}{2 \times \pi \times \sqrt{LC_s}} \cong 0.92 \text{ MHz}$$

The dynamic impedance of the tuned circuit is

$$R_D = \frac{Q}{\omega_o C_s} \cong 2.9 \text{ M}\Omega$$

The coil series resistance is

$$r = \frac{\omega_o L}{Q} \cong 11.6 \text{ }\Omega$$

The capacitor ratio is $C_1/C_2 = \frac{1}{3}$, and therefore $1 - C_2/\beta_o C_1 = 1$. The starting value of g_m is therefore given by

$$g_m = \frac{1}{5000} \times \frac{1}{3} + \left(\frac{1}{3} + 3 + 2 \right) \times \frac{1}{2.9 \times 10^6} \cong 66.7 \mu\text{S}$$

Assuming the input resistance is that of the transistor alone,

$$R_1 = r_{be} = \frac{\beta_o}{g_m} \cong 1.5 \text{ M}\Omega$$

The actual starting frequency is therefore obtained from

$$\begin{aligned} \omega_o^2 &= \frac{1}{LC_s} + \frac{1}{R_1 R_2 C_1 C_2} \\ &= \frac{10^{14}}{3} + \frac{10^{10}}{2.25} \\ &\cong \frac{10^{14}}{3} (\text{rad/sec})^2 \end{aligned}$$

Hence the frequency is

$$f = \frac{10^7}{2 \times \pi \times \sqrt{3}} \cong 0.92 \text{ MHz} \text{ as estimated above.}$$

For the FET circuit, R_1 may be assumed very high, approaching infinity, so that the first term in Eq. 6.4.13 may be neglected. Hence the condition for start of oscillation is

$$g_m R_2 \cong \frac{C_1}{C_2} + \frac{R_2}{R_D} \left(\frac{C_1}{C_2} + \frac{C_2}{C_1} + 2 \right) \quad (6.4.16)$$

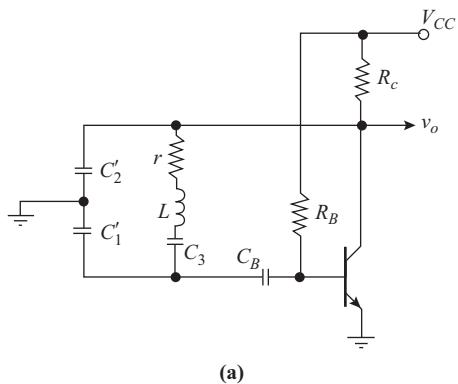
This is seen to be the same as that for the BJT. As before, $g_m R_2$ is the open-circuit gain of the amplifier section. Again, keeping in mind that because the oscillator is a closed circuit the operating point on the transistor will change to maintain $AB = 1$, and because the transistor moves into the large-signal mode, the small-signal parameters have no real meaning for the steady-state condition.

Clapp Oscillator

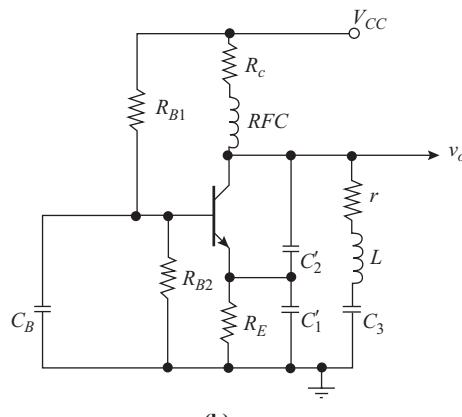
The Clapp oscillator is a modified version of the Colpitt's circuit, where the inductance is replaced by a series LC circuit, as shown in Fig. 6.4.3(a). Figure 6.4.3(b) shows an alternative coupling arrangement, which can also be used with the Colpitt's oscillator. In the circuit of Fig. 6.4.3(b), capacitor C_B effectively grounds the base of the transistor, and the feedback voltage is developed across the emitter resistor R_E .

The series LC_3 branch can be replaced by an effective inductance L_{eff} . The reactance of this branch is given by

$$\omega L_{\text{eff}} = \omega L - \frac{1}{\omega C_3} \quad (6.4.17)$$



(a)



(b)

Figure 6.4.3 (a) Clapp oscillator. (b) Alternative coupling arrangement.

The effective inductance is therefore given by

$$L_{\text{eff}} = L - \frac{1}{\omega^2 C_3} \quad (6.4.18)$$

With the Clapp circuit, the feedback capacitive tap formed by C_1 and C_2 is fixed, so the starting conditions are independent of oscillator tuning. Also, relatively large values of these capacitors can be chosen so that the transistor capacitances have less effect on the frequency. The reactance of the inductive branch changes more rapidly with frequency than that of an inductor alone, which improves frequency stability, assuming the inductance remains constant. However, for the same reason, variations in L , resulting for example from temperature variations, will have a greater effect on frequency than would occur in the Colpitt's circuit.

The analysis is similar to that for the Colpitt's circuit, but with the complication that L_{eff} , which is frequency dependent, must be used in place of L . The results of the analysis are

$$\omega^2 = \frac{1}{LC_s} + \frac{1}{R_1 R_2 C_1 C_2} \left(1 - \frac{1}{\omega^2 LC_3} \right) \quad (6.4.19)$$

where

$$\frac{1}{C_s} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} \quad (6.4.20)$$

The equation for ω is a quartic (to the fourth power) and is difficult to solve. However, in practice the component values are chosen to make the second term on the right-hand side negligible, so the frequency of oscillation is given by

$$\omega^2 \cong \frac{1}{LC_s} \quad (6.4.21)$$

The small-signal analysis also yields for the starting conditions

$$g_m R_2 \cong \frac{1}{R_1} \left(\frac{C_2}{C_1} \right) + \frac{1}{R_2} \frac{C_1}{C_2} + \frac{1}{R_D} \frac{C_1 C_2}{C_s^2} \quad (6.4.22)$$

Hartley Oscillator

The circuit for the Hartley oscillator is shown in Fig. 6.4.4. The analysis of this circuit is complicated by the fact that mutual inductance exists between the two sections of the tapped inductor, and only the results of the small-signal analysis will be summarized here. The frequency of oscillation is given by

$$\omega^2 = \frac{1}{LC} \quad (6.4.23)$$

where $L = L_1 + L_2 + 2M$ is the total inductance of the coil. The starting condition is

$$g_m R_2 \cong \frac{L_2}{(L_1 + M)(L_2 + M)} \frac{R_2}{R_D} + \frac{L_2 + M}{L_1 + M} \quad (6.4.24)$$

As before, $R_1 = r_{be}$, $R_2 = R_0 = R_c \parallel r_c$, and $R_d = L/Cr$ is the dynamic resistance of the tuned circuit.

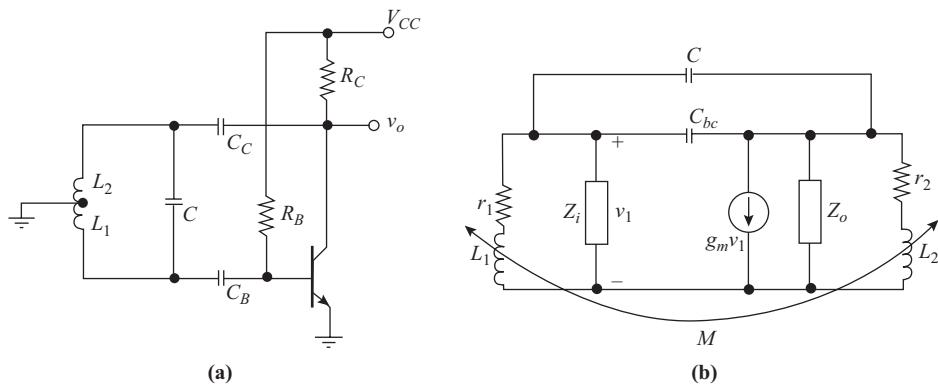


Figure 6.4.4 (a) Hartley oscillator. (b) Small-signal equivalent circuit.

It is assumed that R_1 is large enough to be ignored and that $R_2 \gg r$, the coil resistance. Since the inductance terms in Eq. (6.4.24) are roughly of the same order, then for $R_D \gg R_2$ the starting conditions reduce to

$$g_m R_2 \cong \frac{L_2 + M}{L_1 + M} \quad (6.4.25)$$

6.5 Crystal Oscillators

The characteristics of quartz piezoelectric crystals were discussed in Section 13 on filters. Their use as the frequency control elements in oscillator circuits is discussed here. Crystals can be used to control any of the tuned LC oscillators that were discussed previously by appropriate connections. The crystal may be used to replace an entire LC tank circuit, or it may be used to replace one of the reactances in a tank circuit. The Pierce crystal oscillator circuit illustrates the method.

The Pierce oscillator, like the Clapp oscillator, is basically a Colpitts oscillator in which the inductor is replaced by the crystal. The circuit is shown in Fig. 6.5.1(a), with an equivalent circuit shown in Fig. 6.5.1(b) in which the crystal has been replaced by its equivalent circuit. The resonant frequency of the circuit is determined

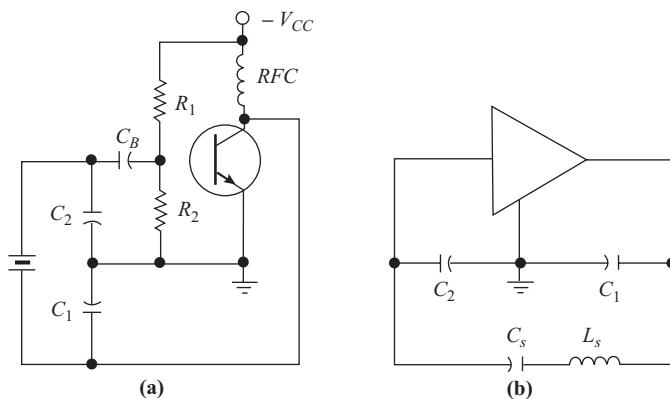


Figure 6.5.1 (a) Pierce crystal oscillator. (b) Equivalent circuit.

by the series resonance of the circuit made up of C_1 , C_2 , C_s and L_s . C_1 and C_2 are both very much larger than C_s , so the resonant frequency is almost entirely dependent on the value of C_s , with little dependence on the input capacity of the amplifier, C_1 , and C_2 . The resonant frequency is almost the series value of the crystal itself.

Any energy withdrawn from the circuit to drive successive amplifier stages is the equivalent of spoiling the Q of the crystal, and special coupling circuits must be used to minimize the loading effect. In practice, a parallel tank circuit tuned to the desired frequency is placed in the collector circuit, and the next stage is transformer coupled through this tank. The load impedance presented to the amplifier can thus be increased to the point where it does not greatly affect the crystal Q .

6.6 Voltage-controlled Oscillators (VCOs)

Voltage-controlled oscillators (VCOs) are found in many applications, such as in automatic frequency control, preset tuning of radios, and phase locked loops (PLLs). These applications are discussed later in the text, but the general principles are similar in all cases. The oscillator is designed so that its frequency can be varied by means of a control voltage, which may for example be applied through operation of a switch or automatically as part of a feedback loop.

Figure 6.6.1(a) shows how the frequency of a Clapp oscillator may be controlled by means of a voltage applied to a varactor diode, which forms part of the tuning circuit. The varactor diode is a reversed biased pn

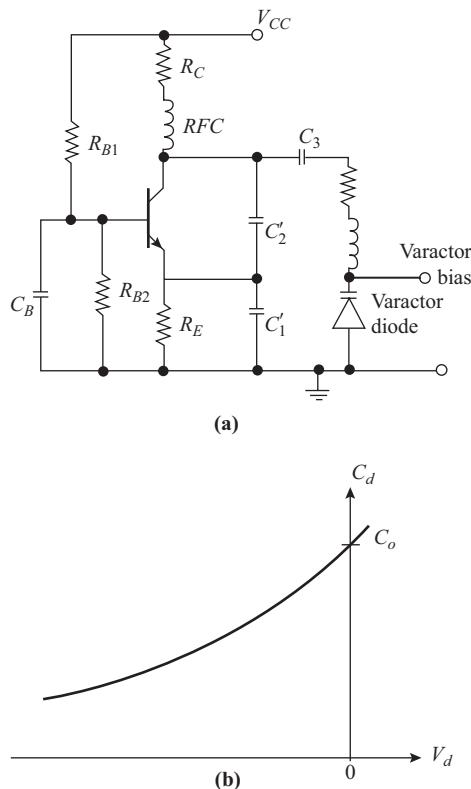


Figure 6.6.1 (a) Voltage-controlled Clapp oscillator. (b) Capacitance/reverse voltage curve for a varactor diode.

junction diode, the name varactor being coined from *variable reactor*. The capacitance of a *pn* junction is a function of the applied reverse bias voltage, the relationship being given by

$$C_d = \frac{C_o}{\left(1 - \frac{V_d}{\psi}\right)^\alpha} \quad (6.6.1)$$

C_o is the diode capacitance at zero bias ($V_d = 0$), ψ is the contact potential of the junction, which may be assumed constant at approximately 0.5 V, and the index α depends on the type of junction. For an abrupt junction, $\alpha = 1/2$ and for a linearly graded junction, $\alpha = 1/3$. This parameter is under the control of the manufacturer, so diodes with different characteristics are available commercially. V_d is the voltage applied across the diode, negative values representing reverse bias.

EXAMPLE 6.6.1

The zero bias capacitance for an abrupt junction varactor diode is 20 pF. Calculate the capacitance when a reverse bias of -7 V is applied.

SOLUTION

$$C_d = \frac{20}{\left(1 - \frac{(-7)}{0.5}\right)^{0.5}} = 5.16 \text{ pF}$$

By altering the bias across the diode, the frequency of the oscillator is changed.

EXAMPLE 6.6.2

The varactor diode of Example 6.6.1 is connected in a Clapp oscillator as shown in Fig. 6.6.1(a). The other values for the oscillator are $C'_1 = 300$ pF, $C'_2 = 300$ pF, $C_C = 20$ pF, and $L = 100 \mu\text{H}$. Calculate (a) the frequency for zero bias and (b) -7 -V reverse bias.

SOLUTION With zero applied bias, the total tuning capacity is

$$C_s = \frac{1}{\frac{1}{300} + \frac{1}{300} + \frac{1}{20} + \frac{1}{20}} \cong 9.38 \text{ pF}$$

Hence the frequency of oscillation is

$$f = \frac{1}{2 \times \pi \times \sqrt{10^{-4} \times 9.38 \times 10^{-12}}} \cong 5.2 \text{ MHz}$$

With a reverse bias of -7 V, the new tuning capacity becomes

$$C_s = \frac{1}{\frac{1}{300} + \frac{1}{300} + \frac{1}{20} + \frac{1}{5.16}} \cong 4 \text{ pF}$$

Hence

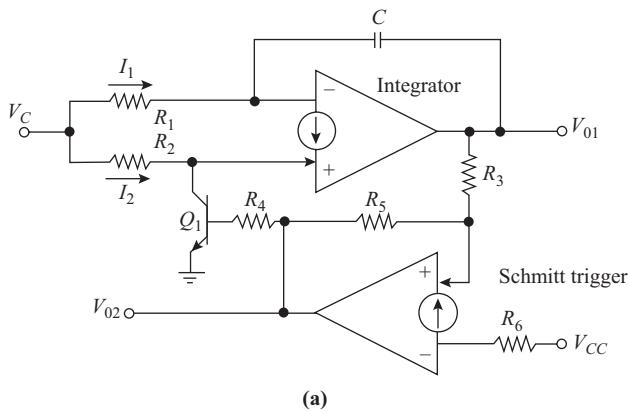
$$f = \frac{1}{2 \times \pi \times \sqrt{10^{-4} \times 4 \times 10^{-12}}} \cong 7.97 \text{ MHz}$$

VCOs can be designed using operational amplifiers, and indeed the complete VCO circuit can be fabricated as a single integrated circuit. The circuit shown in Fig. 6.6.2(a) illustrates a VCO circuit utilizing LM3900 operational amplifiers manufactured by National Semiconductor Corporation. The VCO consists of an inverted integrator, a noninverting Schmitt trigger, both using the LM3900, and a switching transistor Q_1 .

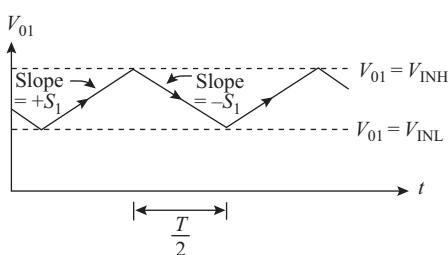
Amplifier type LM3900 is known as a *Norton amplifier* because it works on current inputs rather than voltage inputs. Each terminal, inverting (−) and noninverting (+), is *always* at one diode voltage drop $V_{BE} \cong 0.5$ V above ground. The arrow pointing into the noninverting terminal is to indicate that a controlling current may be fed in at this terminal. The current into the inverting terminal is *forced* to mirror the input current to the noninverting terminal. Assuming for the moment that transistor Q_1 is in the off state so that current I_2 flows into the noninverting terminal, the equations for the integrating amplifier are

$$I_1 = \frac{V_C - V_{BE}}{R_1} \quad (6.6.2)$$

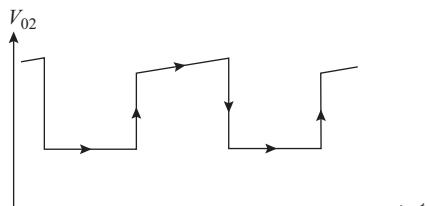
$$I_2 = \frac{V_C - V_{BE}}{R_2} \quad (6.6.3)$$



(a)



(b)



(c)

Figure 6.6.2 (a) VCO utilizing Norton amplifiers. (b) and (c) Output waveforms.

where V_C is the control voltage. While the output from the Schmitt trigger is high, transistor Q_1 is turned on, and the current I_2 is shunted to ground through Q_1 . Current I_1 therefore charges the capacitor, and the rate of change of the output voltage from the inverting integrator is

$$\frac{dV_{o1}}{dt} = -\frac{I_1}{C} \quad (6.6.4)$$

Since I_1 is constant as shown by Eq. (6.6.2), the rate of change gives the slope of the curve, or

$$S_1 = -\frac{I_1}{C} \quad (6.6.5)$$

The output voltage from the integrator is sketched in Fig. 6.6.2(b). When the integrator output voltage reaches the lower switching limit of the Schmitt trigger V_{INL} , the Schmitt trigger switches to a low output, transistor Q_1 turns off, and current I_2 flows into the noninverting terminal of the integrator. The current mirror in the LM3900 requires an equal current to flow into the inverting terminal, and this is supplied from the output of the integrator through C . Thus the charging current for C becomes $(I_1 - I_2)$, and the rate of change of integrator output voltage is

$$\frac{dV_{o1}}{dt} = -\frac{I_1 - I_2}{C} \quad (6.6.6)$$

Current I_2 is also constant, as given by Eq. (6.6.3), and by making $R_1 = 2R_2$, then $I_2 = 2I_1$ and the slope becomes

$$S_2 = \frac{I_1}{C} \quad (6.6.7)$$

Thus the output voltage from the integrator is a symmetrical triangular waveform, and the magnitude of the slope can be written as

$$S = |S_1| = |S_2| = \frac{I_1}{C} \quad (6.6.8)$$

The half-period of the triangular wave is

$$\begin{aligned} \frac{T}{2} &= \frac{V_{INH} - V_{INL}}{S} \\ &= \frac{C(V_{INH} - V_{INL})}{I_1} \end{aligned} \quad (6.6.9)$$

Hence the frequency of the VCO is

$$\begin{aligned} f &= \frac{1}{T} \\ &= \frac{I_1}{2C(V_{INH} - V_{INL})} \end{aligned} \quad (6.6.10)$$

I_1 is a linear function of the control voltage as shown by Eq. (6.6.2), and hence the frequency is also a linear function of the control voltage. The high and low voltages of the Schmitt trigger are set by the resistor

values R_3 , R_5 , and R_6 , and design details will be found in Application Note AN-72, published by the National Semiconductor Corporation.

The multivibrator circuit of Fig. 6.6.3 is another form of integrated circuit VCO. In the steady-state condition, as Q_1 and Q_2 alternately switch on and off, the capacitor charging current alternates between the steady values $\pm I_1$ since the constant current source connected to the emitter of the off transistor must draw its current through C . In Fig. 6.6.3(a) for example, Q_1 is shown in the off state. The rate of change of capacitor voltage, which is the slope of the capacitor-voltage/time curve is therefore

$$S = \pm \frac{I_1}{C} \quad (6.6.11)$$

Denoting by V_D the voltage required to keep a transistor biased on, detailed analysis (see, for example, *Analog Integrated Circuits*, by Paul R. Gray and Robert G. Meyer, 2nd ed., 1984, published by John Wiley & Sons) shows that in the steady state, the capacitor voltage alternates between $\pm V_D$ as shown in

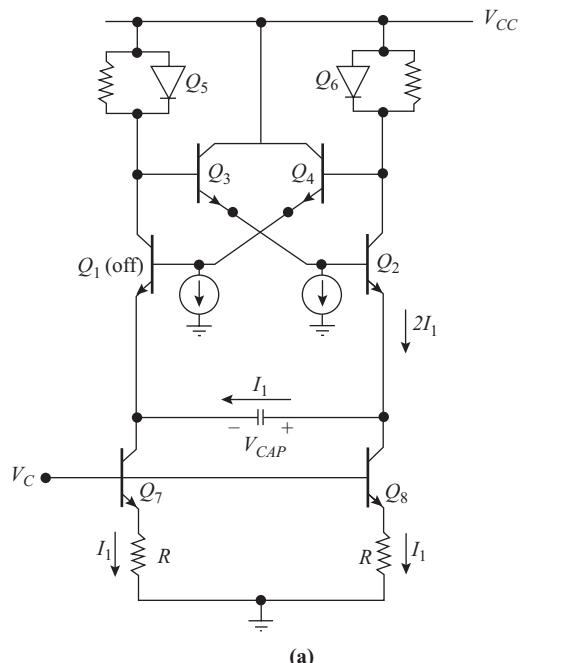


Figure 6.6.3 (a) Multivibrator VCO. (b) Its switching characteristics.

Fig. 6.6.3(b). Over one half-cycle, the capacitor voltage changes by a total amount $2V_D$ and hence the half-period is given by

$$\begin{aligned}\frac{T}{2} &= \frac{2V_D}{S} \\ \therefore T &= \frac{4V_D C}{I_1}\end{aligned}\quad (6.6.12)$$

The frequency of oscillation is given by

$$\begin{aligned}f &= \frac{1}{T} \\ &= \frac{I_1}{4CV_D}\end{aligned}\quad (6.6.13)$$

The collector current of transistor Q_7 , which forms a constant current source, is given by

$$I_1 = I_s e^{v_{BE7}/V_T} \quad (6.6.14)$$

where $V_T = 26$ mV at room temperature, and I_s is the saturation current, a transistor parameter. The base-emitter voltage is given by

$$v_{BE7} = V_C - I_1 R \quad (6.6.15)$$

Thus the relationship between the control voltage V_C and the current I_1 is

$$V_C = I_1 R + V_T \ln \frac{I_1}{I_s} \quad (6.6.16)$$

This is a nonlinear relationship; however, the linear term ($I_1 R$) varies much more rapidly than the logarithmic term, and the circuit parameters can be chosen to make the linear term dominant over the required range of operation, so the relationship becomes $V_C \approx I_1 R + \text{constant}$ (see Problem 6.21). Thus I_1 is linearly dependent on V_C and hence the frequency, as given by Eq. (6.6.13), is also linear in V_C .

6.7 Stability

The stability of oscillator operation can be discussed from a number of points of view. First, will the oscillator produce the desired frequency? Will the oscillator start up by itself and maintain oscillations under all normal load conditions? Will it maintain oscillations at the desired level under all load conditions? Will it produce a sinusoidal output, or will it contain harmonics? These topics will each be discussed in turn.

Frequency Stability

Although the inductance and capacitance in the feedback network are the main frequency determining elements, as the analyses in the previous sections show, other circuit components also affect the frequency. Amplifier output and input resistances and circuit resistances spoil the Q and broaden out the impedance curve near resonance. Also the transistor-capacitances can vary over a wide range, depending on the amplifier supply voltage, temperature, and loading, and unless some effort is made to reduce these effects, they will cause changes to occur in the resonant frequency.

The amount of change occurring in the amplifier parameters can be reduced by providing power to the amplifier from a voltage-regulated source and by using a buffer amplifier with a high input impedance to isolate the oscillator from successive stages. Temperature effects can be minimized by operating the oscillator at low power levels and providing thermal isolation against ambient temperature variations. (The whole oscillator circuit may be mounted in a temperature-regulating oven, for instance.)

Isolation of the tank circuit from the amplifier parasitics can be improved by means of impedance transformation. In the Hartley oscillator this may be accomplished by providing a second tap point on the coil L_1 so that the amplifier input capacity only appears in parallel with a fraction of L_1 . Furthermore, the ratio L/C can be increased (within limits) so that the reactance of the coil L_1 is very much larger than the parasitic reactance and its detuning effect will not be so pronounced. For the Colpitts circuit, the ratio C/L can be increased so as to swamp the parasitic capacity, but the degree of improvement is limited by the small size of coil that may result.

The Clapp circuit provides a greater degree of stability, because the oscillating frequency is almost entirely due to the series reactance of Z_3 , and any convenient values of L_3 and C_3 may be used to yield the correct frequency. Isolation from the amplifier may be improved by several orders of magnitude. The price that must be paid for this improvement, however, is that more gain is required from the amplifier to maintain oscillations, and the frequency is more susceptible to changes in L_3 .

The tuned circuit itself may experience changes of parameters that will cause a change of oscillator frequency. Temperature changes can cause significant changes of parameters. These temperature changes may be partially compensated by choosing elements whose temperature coefficients cancel out in the circuit, and the whole circuit may be placed in a temperature-controlled environment. Each component in the tank circuit must be a rigid entity and must be protected from mechanical vibrations. Even very small distortions of a coil will cause a shift of the oscillating frequency. Microphonics, or changes of the tank-circuit parameters by sound waves or mechanical vibrations, cause frequency modulation of the oscillator and are a common and serious problem. Acoustic shielding and shock mounting are a must for the oscillator.

A low Q in the tank circuit (because of coil resistance or loading) results in a broadening of the notch that appears in the impedance-versus-frequency curve for the tank circuit. The result is that the frequency of oscillation can shift a considerable amount near the true oscillation frequency. The higher the Q becomes, the less this shift is allowed to become. Circuit Q 's from 10 to 1000 are practical in ordinary LC circuits and, with careful design, yield frequency stabilities to about 1 part in 10^4 . For stabilities greater than this, crystals must be used. A quartz crystal run at ambient temperature can give stabilities of about 1 in 10^5 , while a temperature-controlled oscillator can give about 1 in 10^6 .

Self-starting

Self-starting is ensured by guaranteeing that the amplifier is biased to a point before oscillation begins where its gain provides several decibels more than the minimum according to the Barkhausen criterion. If self-starting is to be guaranteed under all operating conditions, it is necessary to make sure that the amplifier is unaffected by temperature and supply variations.

Amplitude Stability

The amplitude of oscillations is determined by factors that reduce the amplifier gain to the point where the loop gain is unity. These factors include an inherent nonlinearity of the amplifying device, such as the onset of saturation or self-bias, and an externally provided amplitude clipper. In any case, the amplitude is subject to variation with load, operating conditions, and supply variations.

A more stable mode of operation can be obtained by using an amplifier whose gain can be controlled by the application of an external bias voltage. An amplitude detector is used to provide a bias voltage that increases with increasing signal amplitude, and this increasing bias is used to decrease the gain of the amplifier (automatic gain control). Very good amplitude stability can be obtained by these means, even under varying load conditions.

Linearity

Some harmonic distortion will result in the output waveform of any oscillator that relies on the amplifier nonlinearity to limit oscillation amplitude. The larger the signal amplitude is allowed to become, the more pronounced is the nonlinearity of operation, and the more harmonic content that will be produced.

Linear operation with a minimum of harmonic generation can be obtained by operating the oscillator at signal levels well below the total range of the amplifier used. The degree to which this can be done is limited by the amount of gain margin necessary to start and maintain oscillations, because the reduced amplitude is obtained by reducing the loop gain. A more reasonable way to do this is to use an amplifier that does operate linearly over a wide range of amplitudes, and to use a feedback system to limit the amplitude of oscillations to well within the linear region of operation. Practically pure sinusoidal output can be obtained by this latter method.

6.8 Frequency Synthesizers

The frequency synthesizer is not a frequency generator in the same sense as an oscillator, but is a frequency converter, which uses a phase-locked loop and digital counters in a phase-error feedback system to keep the output running in a fixed phase relation to the reference signal. Output frequency stabilities are determined by the stability of the reference oscillator, which is typically a crystal-controlled oscillator circuit.

The principles of the frequency synthesizer were developed about 1930 but only found application in very sophisticated equipment because of the cost of the components. Microcircuit chips designed especially for this application are available now at very low cost, and frequency synthesizers are finding increasing application for channel selection in communications equipment.

Phase-locked Loop (PLL)

The heart of the frequency synthesizer is the phase-locked loop. A simple phase-locked loop is illustrated in Fig. 6.8.1, and its operation may be described as follows. A stable oscillator produces a square-wave reference frequency f_r , which provides one of the inputs to the phase-detector circuit. This reference frequency may be any convenient value, but is usually chosen so that a crystal oscillator circuit may be used. A voltage-controlled oscillator (VCO) generates the final output frequency f_o and is designed so that it will tune over the whole range from the minimum frequency to the maximum frequency desired. Its output is fed directly to the load and also is used to drive a programmable binary counter that provides the function of frequency division ($\div N$, where N is the number programmed into the counter). The output of the counter is a square wave at the reference frequency which provides the second input to the phase-comparator circuit.

The phase comparator is a circuit which produces a dc signal whose amplitude is proportional to the phase difference between the reference signal f_r and the counter output f_o/N . This dc signal is filtered to smooth out noise and slow the response of the circuit to prevent overshoot or oscillations and is applied as the control input to the VCO. When the phase difference between the two signals f_r and f_o/N is zero, the dc output from the phase comparator is just exactly that needed to tune the VCO to the frequency Nf_r . If a phase difference exists between the two, the bias applied to the VCO will change in a direction to raise or lower the frequency f_o just sufficiently so that the phase difference will disappear. Once the VCO output reaches the value Nf_r , it will "lock onto" that frequency, and the feedback loop will prevent it from drifting.

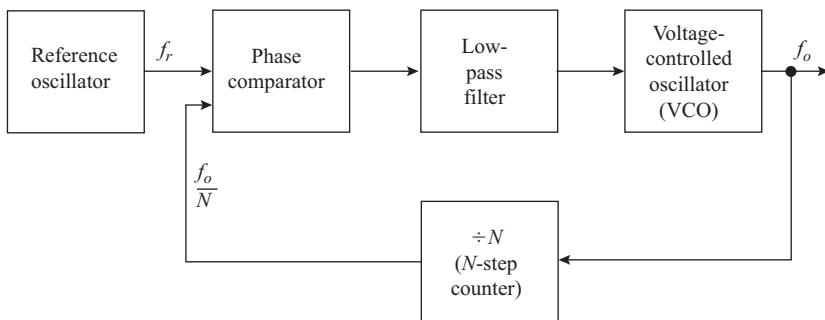


Figure 6.8.1 Basic phase-locked loop frequency synthesizer.

The output frequency f_o is adjusted to a new value by changing the number by which the counter divides. This is accomplished by means of thumbwheel switches or by means of a register into which a new number for N can be entered to control the set point of the counter. The number N is the number of pulses that the counter will count before it recycles, coded in binary.

Prescaling

The simple frequency synthesizer described will only produce output frequencies that are integer multiples of the reference frequency f_r . If other frequencies intermediate between these values are desired, prescaling must be used. Another reason for the use of prescaling is because at high frequencies (above 100 MHz) programmable counters are not available. Fixed-modulus prescale counters are used to count down to a frequency below the 100-MHz limit, and then the prescaler output can drive a low-frequency programmable counter, which is readily available.

Figure 6.8.2 shows how a prescaler circuit can be used to allow division by a noninteger number (a number that contains a fractional part). The prescaler circuit is a two-modulus counter; that is, in one mode it produces an output for every P input pulses, and in the other mode an output for every $P + 1$ pulses. Two low-frequency programmable counters count the output pulses from the prescaler circuit, the main counter counting B pulses and the second counter counting A pulses.

At the beginning of a cycle, both counters are set to their programmed numbers (that is, B and A). As long as the A counter contains a nonzero number, the prescaler will be conditioned to count in the $P + 1$ mode, so the counter chain will count down for $(P + 1)A$ pulses until the A counter goes to zero. At this point, the prescaler circuit will be forced to count in the P mode, and also the input to the A counter will be turned off so that the A counter will remain in the zero state until the B counter completes its count. At the point where the A counter has reached the zero state, the B counter will contain the number $(B - A)$ and will then proceed to count down from $(B - A)$ on every P th pulse from the output. When the B counter reaches zero, both counters reset to their programmed numbers, and the cycle begins again.

The result of this prescaling procedure is shown in Eq. (6.8.1), which relates the output frequency to the reference frequency in terms of the three counter moduli.

$$\begin{aligned}
 f_o &= Nf_r \\
 &= \left(B + \frac{A}{P} \right) Pf_r \\
 &= [(B - A)(P) + (A)(P + 1)]f_r
 \end{aligned} \tag{6.8.1}$$

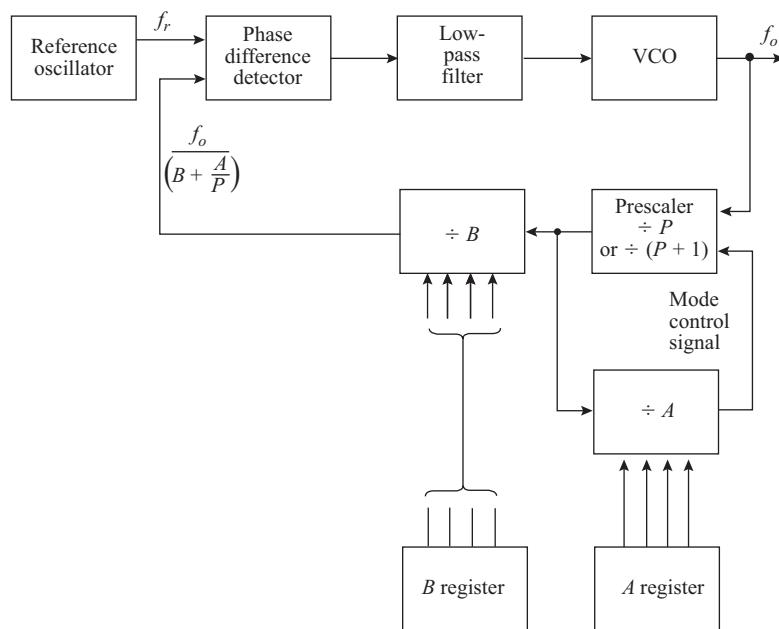


Figure 6.8.2 Frequency synthesizer using prescaling.

Since any fractional number may be stated to a very close approximation as the ratio of two integers, the number of precise frequencies that may be “dialed up” on this frequency synthesizer is very much expanded. A further advantage is that only the prescaler circuit need operate at very high frequencies, and the programmable counters can be made from readily available low-frequency components.

Applications

The most obvious application of the frequency synthesizer is as a digitally programmed (manual or remote) signal generator that may be used for testing purposes, frequency measuring at radio-monitoring stations, or in laboratory frequency-measuring apparatus. Because of the ease of remote setting of the digital numbers to select the output frequency, the synthesizer is increasingly being used in computer-controlled testing stations and other computer-controlled apparatus.

The frequency synthesizer is also used in multichannel communication link transceivers, where it is necessary to frequently switch from one channel to another. The frequency synthesizer generates the local oscillator frequency for the receiver mixer and also the primary frequency source for the transmitter. The switching is usually done manually, but may also be done automatically, as would be the case in a frequency-diversion system. Since the output frequency will have a stability comparable to that of the reference, which may be a crystal oscillator, the flexibility of a variable-frequency oscillator is obtained with the stability of the crystal oscillator. Since most of the components used are inexpensive microcircuits, the result is high-quality but inexpensive communication equipment.

PROBLEMS

- 6.1.** Calculate the frequency of the *RC* phase shift oscillator of Fig. 6.3.1(a) given that $CR = 10^{-4}$ s and $R_o/R = 5$. Calculate also the minimum value of β_o required for these values.
- 6.2.** By plotting Eq. (6.3.8) for $2 \leq R_o/R \leq 3$, verify that the minimum in β_o occurs at $R_o/R = 2.7$. Plot also the frequency as a function of R_o/R given that $CR = 10^{-4}$ s.
- 6.3.** For an FET *RC* phase shift oscillator, $g_m = 12$ mS and $R_o = 34$ k Ω . Calculate R and C required for a frequency of 800 kHz.
- 6.4.** Does the ratio R_o/R have a value that minimizes the open-circuit voltage gain of the FET device in an *RC* phase shift oscillator, similar to that for the β_o of the BJT oscillator? Give reasons for your answer.
- 6.5.** Determine the minimum value which β_o can have for a *BJT* used in the *RC* oscillator of Fig. 6.3.1. Given that in an actual oscillator $\beta_o = 50$, determine the two possible values of R_o/R .
- 6.6.** For the *LC* oscillator of Fig. 6.4.1(c), $Z_1 = Z_2 = -j100$ Ω , and $Z_3 = 10 + j200$ Ω . Find the minimum value of g_m required to sustain oscillation.
- 6.7.** For a Colpitt's BJT oscillator, $C_1 = C_2 = 70$ pF, $r = 5$ Ω , $L = 15$ μH ; $R_1 = 10$ k Ω , and $R_2 = 10$ k Ω . Calculate and compare the frequency obtained using Eqs. (6.4.7), (6.4.9) and (6.4.10).
- 6.8.** For a Colpitt's BJT oscillator, $C_1 = 120$ pF, $C_2 = 70$ pF, $r_{ce} = 25$ k Ω , $L = 3$ mH, and $R_C = 10$ k Ω . The *Q*-factor of the tuned circuit alone is 75. Calculate the transconductance required to sustain oscillation, and the frequency.
- 6.9.** For a Colpitt's BJT oscillator, $C'_1 = C'_2 = 100$ pF, $C_{be} = 10$ pF, $C_o = 7$ pF, $r_{ce} = 25$ k Ω , $r = 7$ Ω , $L = 100$ μH , and $R_C = 5$ k Ω . Calculate the transconductance needed to sustain oscillation, and the frequency.
- 6.10.** For a Colpitt's FET oscillator, $C_1 = C_2 = 200$ pF, $r = 4$ Ω , $L = 50$ μH , $r_d = 25$ k Ω , and $R_C = 15$ k Ω . Find the open-circuit voltage gain and the frequency.
- 6.11.** The Colpitt's oscillator of Problem 6.7 is converted to a Clapp oscillator by the addition of a capacitor $C_3 = 5$ pF in series with the inductor, the other values remaining at $C_1 = C_2 = 70$ pF, $r = 5$ Ω , $L = 15$ μH , $R_1 = 10$ k Ω , and $R_2 = 10$ k Ω . Determine the frequency of oscillation and the transconductance required to start oscillation.
- 6.12.** Rearrange Eq. (6.4.19) as a function of frequency and, using the values from Problem 6.11, plot this function over the range from 19.64 to 19.65 MHz in steps of 0.001 MHz. Determine from the graph the oscillation frequency and compare with the result obtained in Problem 6.11.
- 6.13.** For a Hartley oscillator, $L_1 = 15$ μH , $L_2 = 20$ μH , $M = 10$ μH , $C = 70$ pF, $R_2 = 10$ k Ω , and $R_D = 80$ k Ω . Calculate (a) the coefficient of coupling k , (b) the frequency, and (c) the open-circuit voltage gain using both Eqs. (6.4.24) and (6.4.25).
- 6.14.** For a Hartley oscillator the inductor is tapped at exactly the midway point. Determine the open-circuit voltage gain required for start-up of oscillation. Assume $R_D \gg R_2$.
- 6.15.** A crystal oscillator has the crystal connected between the amplifier input terminals, and a parallel-resonant tank circuit connected between the amplifier output terminals. A small capacitor provides the feedback path between the input and output. The crystal is series resonant at 1.50000 MHz and parallel-resonant at 1.50001 MHz. (a) To what frequency must the output tank circuit be tuned to guarantee oscillations? What will the frequency of oscillation be? How stable is this in parts per million? (b) If the tank circuit were tuned to the second harmonic, would the circuit still oscillate? (c) At what frequency would the crystal vibrate under these conditions?

- 6.16.** On a common set of axes, plot the capacitance – voltage curves for abrupt and linear graded junctions for the voltage range $-7 \text{ V} \leq V_D \leq 0$. A built-in potential of $\psi = 0.5 \text{ V}$ may be assumed.
- 6.17.** A Clapp oscillator has $C_1 = C_2 = 70 \text{ pF}$, $L = 15 \mu\text{H}$, and C_3 is a varactor diode for which $\psi = 0.5 \text{ V}$, $\alpha = 0.5\text{V}$, $C_o = 7\text{pF}$, and bias = -7 V . A low-frequency sinusoidal signal of peak value 10 mV is applied in series with the bias. Plot the resultant frequency over one cycle of the signal voltage.
- 6.18.** The low-frequency signal of problem 6.17 is changed to a linear function of time, ranging from 10 to 15 mV . Plot the oscillator frequency as a function of this voltage, and use linear regression to determine how close the plot is to a straight line.
- 6.19.** For a VCO, $V_{\text{INH}} = 8 \text{ V}$, $V_{\text{INL}} = 2 \text{ V}$, $C = 500 \text{ pF}$, $V_{BE} = 0.5 \text{ V}$, $R_1 = 0.5 \text{ k}\Omega$, $V_C = 3.5 \text{ V}$, and $R_2 = 0.5R_1$. Calculate the slope S of the output waveform and the periodic time T .
- 6.20.** For the VCO of Problem 6.19, given that the control voltage consists of a bias of 3.5 V in series with a low-frequency sinusoidal signal voltage of 0.5 V peak, plot the frequency over one cycle of the signal voltage.
- 6.21.** For the circuit of Fig. 6.6.3, the values are $R = 500 \Omega$, $I_s = 10^{-14} \text{ A}$, and $V_T = 26 \text{ mV}$. Given that $-0.5 \text{ mA} \leq I_1 \leq 0.5 \text{ mA}$, plot the control voltage V_C as a function of I_1 in 0.5-mA steps. Using linear regression, find the best straight-line fit to the plot. State, with reasons, the parts of Eq. (6.6.16) that contribute to the intercept and to the slope.
- 6.22.** For the circuit of Problem 6.21, the control voltage consists of a fixed bias $V_{\text{BIAS}} = 3 \text{ V}$ in series with a small-signal voltage. Calculate the frequency when the small-signal voltage is zero. Plot the change in frequency as a function of small-signal voltage in steps of 1mV , when this varies between $\pm 5 \text{ mV}$.
- 6.23.** The frequency synthesizer in Fig. 6.8.2 is run from a 1.0000-MHz crystal reference oscillator to produce an output of 146 MHz . The main counter will only work up to 80 MHz , and a prescaler P that divides by 10 is used. Find the values of B and A that are appropriate to these conditions.
- 6.24.** Show how an oscillator can be simulated using MATLAB. (Hint: Invoke Simulink in MATLAB, by typing *simulink* at MATLAB command prompt. Use the step function block as the signal source, a linear transfer function block such as $1/(s^2 + 25)$ as the system block (any system with poles on the imaginary axis would do), and the scope as the signal sink.)
- 6.25.** Determine the value of $\frac{R}{R_o}$ in Equation 6.3.8 so that β_o has its minimum value. What is the minimum value of β_o ? (Hint: Let $R/R_0 = x$. Now substitute $d\beta_o/dx = 0$, and solve for x .)
- 6.26.** Plot Equation 6.6.1 for an abrupt junction ($\alpha = 1/2$) and a linear graded junction ($\alpha = 1/3$), using MATLAB.
- 6.27.** Determine the value of $\frac{R_o}{R}$ in Equation 6.3.10 so that $g_m R_o$ has its minimum value. Assume $R_o = 4k\Omega$. Obtain the minimum value of $g_m R_o$.
- 6.28.** Derive the condition for oscillation in the *Clapp Oscillator*.



Receivers

7.1 Introduction

A radio receiver must perform a number of functions. First, the receiver must separate a wanted radio signal from all other radio signals that may be picked up by the antenna and reject the unwanted ones. Next, the receiver must amplify the desired signal to a usable level. Finally, the receiver must recover the information signal from the radio carrier and pass it on to the user. This chapter will examine the operating principles of some of the more commonly used radio receivers.

7.2 Superheterodyne Receivers

Early receivers used for the reception of amplitude-modulated signals or interrupted-carrier telegraph signals used the tuned radio frequency (TRF) principle. This type of receiver was simply a chain of amplifiers, each tuned to the same frequency, followed by a detector circuit. These receivers suffered from poor adjacent signal rejection, especially when required to tune over wide frequency ranges where the Q of the tuned circuits change with frequency.

The superheterodyne receiver was developed to improve the adjacent channel selectivity by placing most of the frequency selectivity in the fixed tuned *intermediate frequency* (IF) stages after the frequency conversion. *Superheterodyne* action takes place when two signals of different frequencies are *mixed* together. Mixing involves adding and passing the result through a nonlinear device (or multiplying together) so that the output contains the product of the two signals as well as the two original signals. The product term can be separated into two signals, one at the sum frequency and one at the difference frequency.

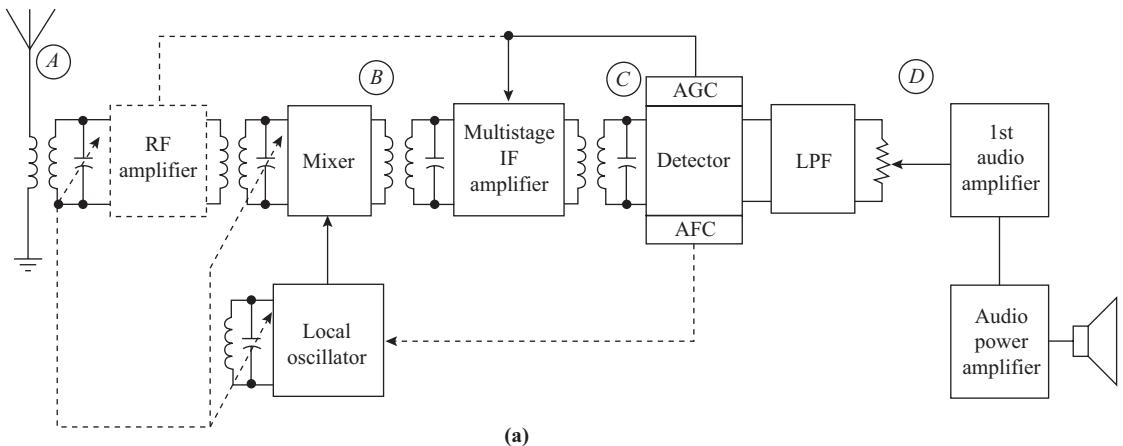
In the frequency conversion process, the oscillator frequency may be placed above or below the signal frequency, and either the sum or the difference frequency may be used as the output. For an up-conversion, the sum frequency is used as the output, with the oscillator either above or below the signal. For a down-conversion, the difference frequency is used as the output, with the oscillator either above or below the signal frequency. In the superheterodyne receiver, a down-conversion is usual, where the received radio signal at

frequency f_s is mixed with the signal from a local oscillator at f_o (usually located above f_s), and the difference frequency produced is taken as the intermediate frequency or IF as

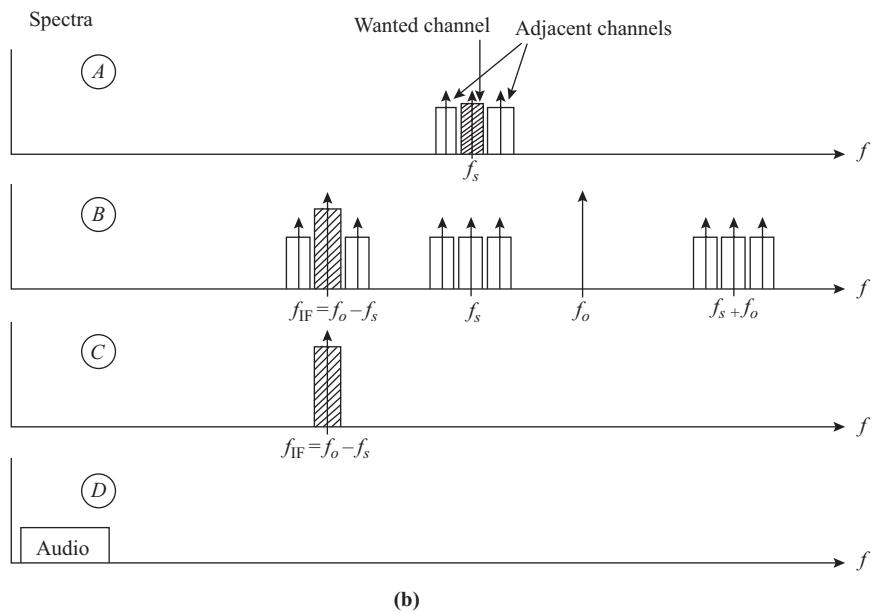
$$\text{IF} = |f_o - f_s| \quad (7.2.1)$$

The superheterodyne broadcast receiver was the original application of this principle and is still one of the largest. The name superheterodyne is a contraction of the term *supersonic heterodyne*, or the production of beat frequencies above the range of hearing.

The basic superheterodyne receiver is illustrated in Fig. 7.2.1(a). The first stage is a tuned RF amplifier, using two variable tuned circuits that track each other and the local oscillator. The two tuned RF circuits form a band-pass filter to pass the desired RF signal frequency while blocking others. This stage acts to boost the weak



(a)



(b)

Figure 7.2.1 (a) Superheterodyne receiver. (b) Signal spectra in a superheterodyne receiver.

signal level from the antenna above the noise level to provide some signal selectivity and to prevent reradiation of the local oscillator signal. Cheaper receivers may omit the RF amplifier and the second tuning circuit.

The output signal from the RF amplifier is fed to one input of the mixer circuit and the local oscillator signal to the other. While separate circuits may be used for the mixer and oscillator, the two functions are frequently combined in the same circuit. The oscillator is also variably tuned so as to track the incoming signal frequencies. In some receivers (especially older ones) the variable tuning is done with a multigang variable capacitor especially cut to provide proper frequency tracking. Newer receivers mostly use varactor diode tuning, which allows remote control and very compact circuits.

The mixer output (the difference frequency for down-conversion) is fed to two cascaded tuned IF amplifiers, which are fixed-tune and provided with sufficient selectivity to reject adjacent channel signals. Older receivers typically used tuned transformers for the filtering action, but many recent receivers use inexpensive ceramic resonator filters with a high-gain integrated-circuit amplifier.

The output from the IF amplifier chain is fed to the detector circuit, where the audio signal is extracted from the IF carrier, or demodulated. The detector also provides signals for automatic gain control (AGC) and for automatic frequency control (AFC) in FM receivers. The AGC signal is used as a bias signal to reduce the gain of the RF and the IF amplifiers to prevent detector overload on strong signals. The AFC signal is used to adjust the frequency of the local oscillator so that it “locks” to the average of the received signal frequency and to counteract minor mistuning problems.

The audio signal from the detector is passed through a low-pass filter to remove unwanted high-frequency components and then through a volume control to an audio amplifier. The audio amplifier is usually one low-level audio stage followed by a power amplifier and a speaker.

Figure 7.2.1(b) illustrates the spectra of signals at various points in the receiver. The spectrum of the RF signal obtained from the antenna is shown in ①, with the desired channel and two adjacent channels. The unfiltered output from the mixer ② includes the RF signal frequencies, the oscillator frequency, and repeats of the RF signals at the sum and at the difference frequencies. The spectrum of the output from the IF band-pass filters ③ shows the desired channel at the IF, with all other frequencies, including adjacent channels, removed. Finally, the spectrum at the output of the demodulator low-pass filter ④, shows only the baseband modulation frequencies.

7.3 Tuning Range

Many radio receivers are fixed-tuned to a specific signal frequency, while others are designed to be continuously adjustable over a range (or band) of frequencies. Tuning of the RF amplifiers and the oscillator is accomplished by varying the capacitance (or sometimes the inductance) in resonant circuits that act as band-pass filters. The tuning range is usually limited by the range over which the capacitance can vary, typically a maximum of about 10 : 1. The resonant frequency of a high-*Q* tuned circuit is given by

$$f_o = \frac{1}{2\pi\sqrt{LC}} \quad (7.3.1)$$

The circuit frequency tuning range ratio R_f is defined as the ratio of its maximum frequency to its minimum frequency, and the corresponding capacitance tuning range ratio R_C is the ratio of maximum capacity to minimum capacity. Applying these ratios to the resonant equation (1.3.4) gives

$$R_C = \frac{C_{\max}}{C_{\min}} \quad (7.3.2)$$

$$R_f = \frac{f_{\max}}{f_{\min}} = \sqrt{R_C} \quad (7.3.3)$$

If the oscillator frequency is chosen to be above the received signal frequency, then the tuning range of the oscillator tuned circuit will be smaller than that of the RF amplifier tuned circuits. If the oscillator frequency is below the signal, then its tuning range will be larger than that of the RF circuits, and also its harmonics may fall within the signal range to cause interference. This is particularly true if the intermediate frequency IF is made much smaller than the signal frequency, where the oscillator is located very near the signal frequency. While direct interference may not occur, the oscillator signal may desensitize the receiver. Thus it is usual to choose the oscillator frequency to be well above the signal frequency, with the IF just below the minimum signal frequency to be used.

EXAMPLE 7.3.1

A receiver tunes signals from 550 to 1600 kHz with an IF of 455 kHz. Find the frequency tuning ranges and capacitor tuning ranges for the oscillator section and for the RF section.

SOLUTION For the oscillator section,

$$f_{o \min} = f_{s \min} + \text{IF} = 550 + 455 = 1005 \text{ kHz}$$

$$f_{o \max} = f_{s \max} + \text{IF} = 1600 + 455 = 2055 \text{ kHz}$$

$$R_f = \frac{f_{o \max}}{f_{o \min}} = \frac{2055}{1005} = 2.045$$

$$R_C = R_f^2 = 2.045^2 = 4.182$$

For the RF section,

$$R_f = \frac{f_{s \max}}{f_{s \min}} = \frac{1600}{550} = 2.909$$

$$R_C = R_f^2 = 2.909^2 = 8.463$$

This example illustrates the usual design for medium-wave broadcast receivers. Typically, these receivers tune from 550 to 1600 kHz, with an IF of 455 kHz. The oscillator tunes from 1005 to 2055 kHz. The bandwidth of the RF tuned circuits is typically less than 100 kHz, so the oscillator signal falls outside the passband and no interference takes place. The tuning range ratio of the RF circuits is 2.91 and that of the oscillator is 2.05. The corresponding capacitor tuning ranges are 8.46 (RF) and 4.18 (osc).

7.4 Tracking

A *scanning receiver* is one that is designed to tune continuously over a range or band of frequencies. For this type of receiver to function properly, the local oscillator tuning circuit and the RF amplifier tuning circuits must track each other so that at all points across the band the RF circuits tune exactly to the signal, and the oscillator is offset from this by exactly the IF. As noted previously, the signal capacitance tuning ratio for an MF receiver is 8.46 while that for the oscillator is 4.18. Also, the oscillator frequency is above that of the

signal. Thus the oscillator tuning section must have a smaller capacitor with a smaller tuning range than that of the RF amplifier circuits.

One way to obtain the different capacitances is to use a specially made tuning capacitor that has two or more sections, with one cut specially to provide the proper capacitance to tune the oscillator. Two-gang and three-gang capacitors were common in broadcast receivers built up until recently.

If the receiver is to tune more than one band, then usually the sections of the tuning capacitor are made identical to each other, and the value of the oscillator section is altered by adding series *padder* capacitors and/or parallel *trimmer* capacitors. The tracking that results is not perfect, and adjustment of the circuit to minimize tracking error is tedious. Furthermore, for each band of frequencies to be covered, a different set of trimmers and padders must be switched into the circuits, further compromising the design. Figure 7.4.1 shows the three possible arrangements of trimmers and padders to provide oscillator tracking. Note that with a single trimmer or single padder, the tracking error can be zero at two points within the band, while if both padders and trimmers are used, the error is zero at three points, and the maximum error can be made smaller with careful adjustment.

For tracking with only a padder C_p in series with the oscillator section main capacitor C_s , the oscillator capacitor C_o is given by

$$C_o = \frac{C_s \cdot C_p}{C_s + C_p} \quad (7.4.1)$$

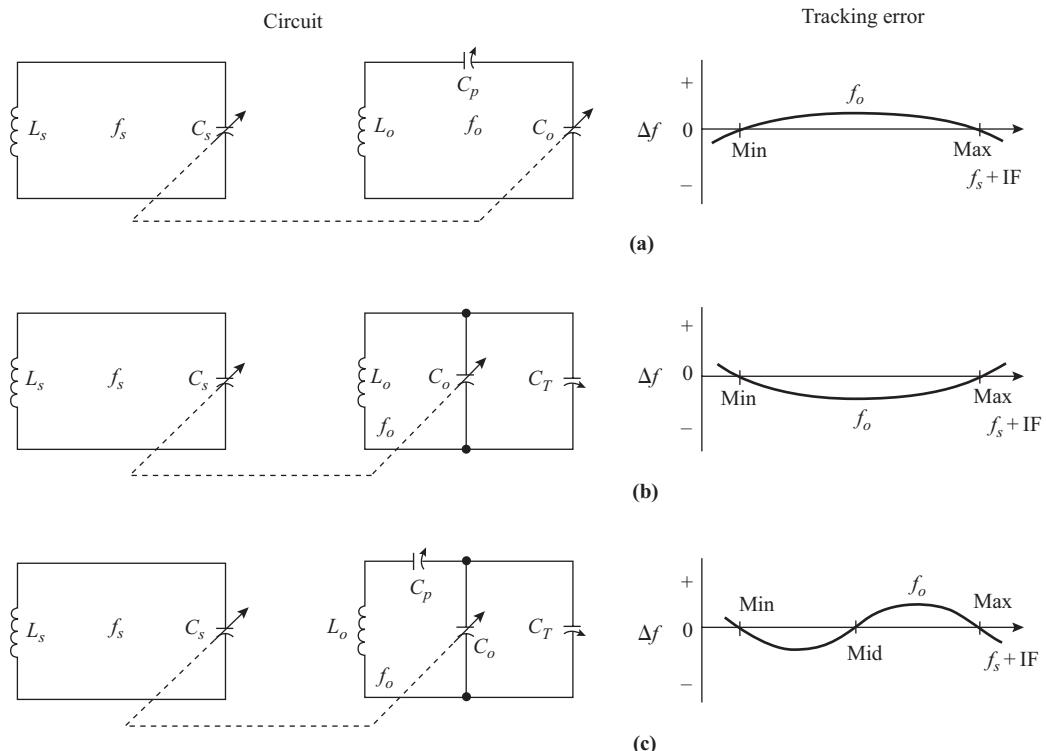


Figure 7.4.1 Superheterodyne receiver tracking methods: (a) padder tracking, (b) trimmer tracking, and (c) combination or three-point tracking. In each case both sections of the tuning capacitor have the same value and the resulting tracking error is shown.

Substituting minimum and maximum values of C_o into the max/min capacitor range equation allows C_p to be found uniquely.

If a trimmer capacitor is used in the oscillator section, then C_o is the parallel combination, given by

$$C_o = C_s + C_t \quad (7.4.2)$$

and again C_t can be found by substituting in the max/min capacitor range equation.

For the combined padder – trimmer combination, tracking coincides at a midrange frequency f_{mid} as well as the two ends, and a second capacitor range $C_{o\ min}/C_{o\ max}$ can be obtained at this frequency. The oscillator capacitor is found to be

$$C_o = \frac{C_p(C_s + C_t)}{C_p + C_s + C_t} \quad (7.4.3)$$

Substituting the min-, mid-, and max-values of C_o into the mid/min and max/min oscillator capacitance ratios allows simultaneous solution for C_p and C_t .

EXAMPLE 7.4.1

For the receiver in Example 7.3.1, given a two-section tuning capacitor with a maximum 350 pF/section and a capacitance ratio as in Example 7.3.1, find the required padder capacitor required in the oscillator section. Assume that tuning error is zero at the extreme ends of the tuning range.

SOLUTION From Example 7.3.1,

$$\begin{array}{lll} R_{Co} = 8.463, & R_{fo} = 2.909, & f_{o\ max} = 2055 \text{ kHz} \\ R_{Cs} = 4.182, & R_{fo} = 2.045, & f_{o\ min} = 1005 \text{ kHz} \end{array}$$

For the RF section,

$$C_{s\ max} = 350 \text{ pF}, \quad C_{s\ min} = \frac{C_{s\ max}}{R_{Cs}} = \frac{350}{8.463} = 41.36 \text{ pF}$$

From Eq. (7.4.1), the oscillator max/min ratio is

$$\begin{aligned} \frac{C_{o\ max}}{C_{o\ min}} &= \left(\frac{C_{s\ max}}{C_{s\ min}} \right) \left(\frac{C_{s\ min} + C_p}{C_{s\ max} + C_p} \right) \\ 4.182 &= 8.463 \frac{41.36 + C_p}{350 + C_p} \end{aligned}$$

with C_p in picofarads. Solving for C_p gives

$$C_p = 260.14 \text{ pF}$$

Now the maximum and minimum oscillator capacitances are

$$C_{o\ max} = \frac{C_p \cdot C_{s\ max}}{C_p + C_{s\ max}} = \frac{260.14 \cdot 350}{260.14 + 350} = 149.2 \text{ pF}$$

$$C_{o\ min} = \frac{260.14 \cdot 41.36}{260.14 + 41.36} = 35.69 \text{ pF}$$

EXAMPLE 7.4.2

For the conditions in the Example 7.4.1, find the tuning error in the RF tuning when the oscillator is tuned to receive a 1-MHz signal.

SOLUTION With the oscillator tuned to receive the mid-frequency of 1000 kHz,

$$f_{o\text{ mid}} = \text{IF} + f_{s\text{ mid}} = 455 + 1000 = 1455 \text{ kHz}$$

Using the max/mid ranges,

$$R_{fo} = \frac{f_{o\text{ max}}}{f_{o\text{ mid}}} = \frac{2055}{1455} = 1.4123$$

$$R_{co} = R_{fo}^2 = 1.4123^2 = 1.995$$

$$C_{o\text{ mid}} = \frac{C_{o\text{ max}}}{R_{co}} = \frac{149.2}{1.995} = 74.80 \text{ pF}$$

$$C_{s\text{ mid}} = \frac{\frac{1}{C_{o\text{ mid}}}}{\frac{1}{C_p}} = \frac{\frac{1}{74.80}}{\frac{1}{260.14}} = 105.00 \text{ pF}$$

$$R_{Cs} = \frac{C_{s\text{ max}}}{C_{s\text{ mid}}} = \frac{350}{105.00} = 3.334$$

$$R_{fs} = \sqrt{R_{Cs}} = \sqrt{3.334} = 1.826$$

$$f'_{s\text{ mid}} = \frac{f_{s\text{ max}}}{R_{fs}} = \frac{1600}{1.826} = 876.7 \text{ kHz}$$

Now the RF circuit is mistuned from the 1000 kHz desired frequency by

$$\text{Err} = f'_{s\text{ mid}} - f_{s\text{ mid}} = 876.7 - 1000 = -123.7 \text{ kHz}$$

With two tuned circuits with a Q of 20 in the RF sections, this causes about a 16-dB decrease in the RF signal strength below that for correct tuning. Adjusting the zero-error crossover points to fall nearer to the center of the tuning range would reduce the maximum error.

Modern scanning receivers use varactor diode tuning, in which the tuning capacitors are replaced by the voltage-variable capacity of the varactor diodes. In this case one variable dc voltage is generated by a potentiometer on the front panel or by a D/A converter from a digital circuit containing a number representing the desired signal. Again, the oscillator varactor can be chosen to give the proper range of tuning, or the same type may be used for both RF and oscillator and the oscillator adjusted with a padder. In the latter case, the trimming and padding can be done in the bias circuit with resistive components. Since the source of the tuning voltage does not have to be located in the RF section of the receiver, this system is ideal for remote tuning applications.

Figure 7.4.2 shows one arrangement for an electronically tuned receiver. It has an RF amplifier (not shown) with two tuned circuits on it. The variable capacitances of the varactors parallel-tune with the inductances of the transformer windings. The resulting impedance is coupled to the signal circuit through the secondary winding of the transformer, providing any impedance matching required. All three varactors have

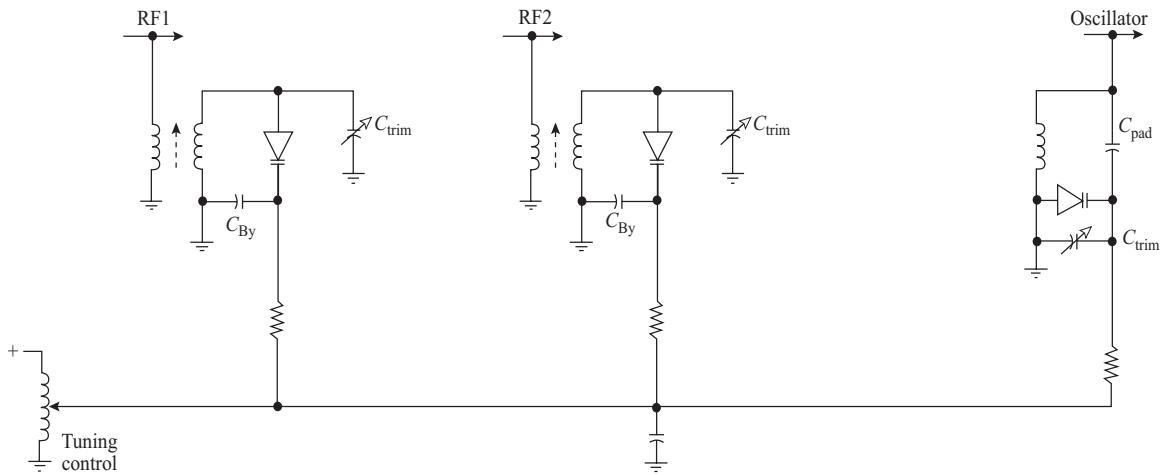


Figure 7.4.2 Electronic tuning for a broadcast receiver.

parallel trimmer capacitors to allow adjustment of their ranges to track each other. The oscillator circuit has a padder capacitor in series with the varactor to shift its tuning range to the required value.

7.5 Sensitivity and Gain

The *sensitivity* of a receiver is its ability to receive weak signals. This sensitivity may be defined in several ways. First, it may be stated in terms of the signal field strength of a signal that will produce a desired demodulated output level under a certain modulation level. The sensitivity is usually stated in terms of the voltage developed by the antenna across the receiver antenna terminals in microvolts. This level ranges from a few microvolts to a few hundred microvolts for typical receivers.

Another way of stating the sensitivity is to state the antenna terminal signal voltage required to produce a specified signal-to-noise ratio (as, for example, 10 μ V to produce a 30-dB signal-to-noise ratio). In the case of receivers for digital signals, the sensitivity is usually stated as the input signal level required to produce a desired *bit-error rate* or BER, which is related to the signal-to-noise ratio. Typically, the desired BER may be 1 bit in 100,000 or less.

Frequency-modulated receivers are designed with a limiting amplifier stage just before the detector, which serves to keep any amplitude variations on the signal from reaching the detector. In this case, the sensitivity is stated as the input voltage level required to just bring the limiting amplifier to the saturation level. Further increase in input signal level does not increase the signal at the detector significantly. Operation of FM receivers below the limiting threshold is usually not recommended since any amplitude variations increase the noise output.

The gain required in the RF and IF amplifier chain of a receiver depends on the required input and output. The input is the minimum usable signal level to be presented at the antenna terminals. The output is the minimum signal level at the input of the detector required to make the detector perform satisfactorily. A typical antenna carrier voltage is 10 μ V presented on a 50- Ω input (that is, about 2 pW). Typically, a peak-rectifying AM detector will require a minimum carrier signal level (with 90% modulation) of about 2 V peak (1.414 V rms) on a 1000- Ω input resistance, or about 2 mW. This represents a total signal power level increase from antenna terminals to detector of 10^9 , which is +90 dB.

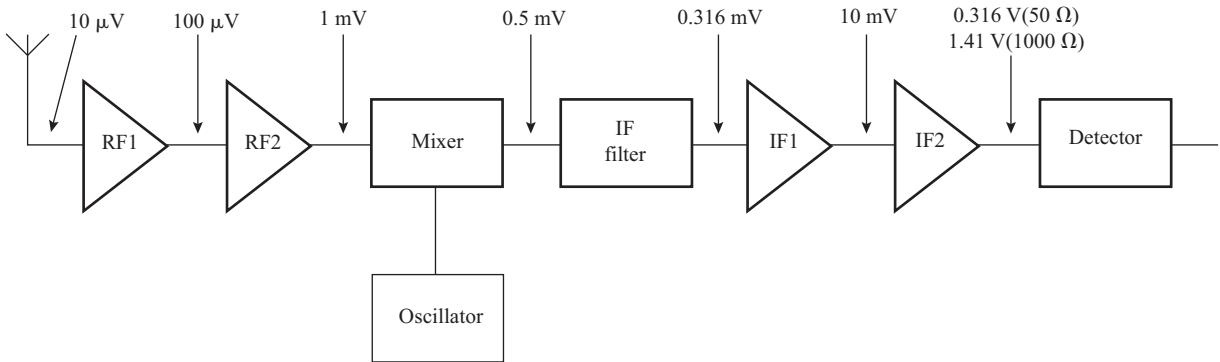


Figure 7.5.1 Voltage levels at various points in a receiver.

To improve the signal-to-noise ratio, the first RF amplifier stage is made with a low noise figure and as much gain as possible. Thus noise generated in later stages does not contribute significantly to the overall noise produced in the receiver. A typical tuned RF amplifier will have a gain of about 20 to 30 dB. Mixer circuits will have conversion gains of about -10 to $+10$ dB. A good IF amplifier stage may have 20- to 30-dB gain. A passive IF filter may have an insertion loss of 1 to 6 dB. The overall gain budget for a possible receiver configuration follows:

First RF amplifier gain	+20 dB
Second RF amplifier gain	+20 dB
Mixer conversion gain	-6 dB
IF filter insertion loss	-4 dB
First IF amplifier gain	+30 dB
Second IF amplifier gain	<u>+30 dB</u>
Total gain	+90 dB

Figure 7.5.1 shows the voltage signal levels occurring at various points in the receiver illustrated by the preceding table. These voltages are all based on a $50\text{-}\Omega$ impedance level, with the final voltage shown for both $50\text{-}\Omega$ and $1000\text{-}\Omega$ impedances.

7.6 Image Rejection

The superheterodyne mixer circuit produces a signal component at the IF frequency that is the difference between the oscillator frequency and the signal frequency. The signal frequency may be either below or above the oscillator frequency and still produce an IF signal. If the desired signal is located at $(f_o - \text{IF})$, a strong signal at $(f_o + \text{IF})$ will interfere with it. This second signal at $(f_o + \text{IF})$ is called the *image*.

Normally, the band-pass of the RF circuits will be narrow enough that the undesired image signal will be prevented from reaching the mixer. However, if the Q of the RF circuits is low, its band-pass will be wide, and if the IF is small, then the image will fall within the band-pass of the tuned circuits and not be rejected. Once the image signal has reached the mixer and is converted, it cannot be separated from the desired signal. The image signal must be prevented from passing the RF amplifiers. Figure 7.6.1 illustrates this relation.

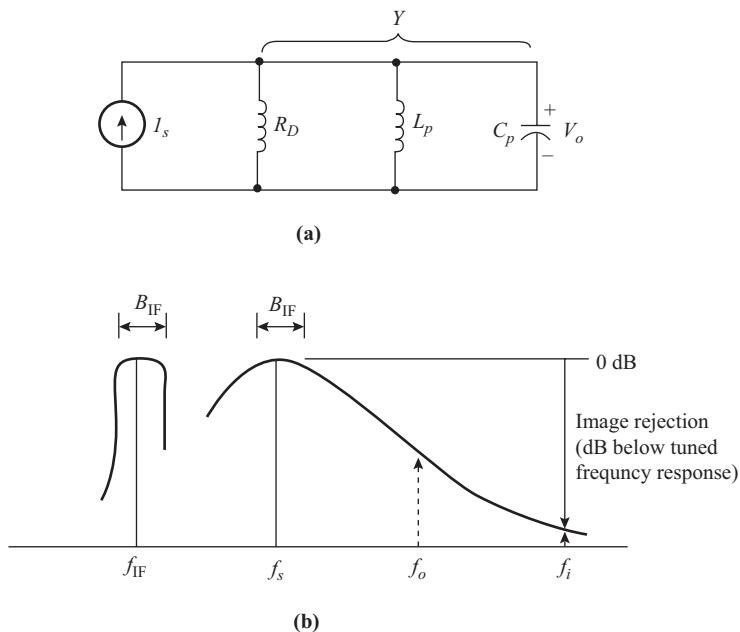


Figure 7.6.1 (a) Parallel tuned circuit. (b) Image frequency rejection.

The image rejection capability of a parallel tuned circuit can be calculated from the relative response of the circuit as given by Eq. (1.4.5), the magnitude being given by

$$|A_r| \cong \frac{1}{\sqrt{1 + (yQ)^2}} \quad (7.6.1)$$

If two or more tuned circuits that are not coupled to each other are used in the RF amplifier chain to improve the selectivity, then the overall response is given by

$$A_r = A_{r1} \cdot A_{r2} \cdots \quad (7.6.2)$$

These ratios are usually stated in decibels (dB), so that

$$A_r(\text{dB}) = A_{r1}(\text{dB}) + A_{r2}(\text{dB}) + \cdots$$

Inexpensive receivers such as those for broadcast reception typically only have one tuned circuit in the RF amplifier, while more expensive communications receivers will have two or three in order to obtain the required image rejection. Using an IF near the signal frequency also helps by placing the image frequency farther out on the image response curve.

EXAMPLE 7.6.1

An AM broadcast receiver has an IF of 465 kHz and is tuned to 1000 kHz, and the RF stage has one tuned circuit with a Q of 50. (a) Find the image frequency. (b) Find the image rejection in decibels.

SOLUTION (a) For $f_s = 1000$ kHz,

$$f_o = f_s + \text{IF} = 1000 + 465 = 1465 \text{ kHz}$$

and

$$f_i = f_o + \text{IF} = 1465 + 465 = 1930 \text{ kHz}$$

(b)

$$y = \frac{f}{f_o} - \frac{f_o}{f} = \frac{f_i}{f_s} - \frac{f_s}{f_i} = \frac{1930}{1000} - \frac{1000}{1930} = 1.412$$

$$A_r = \frac{1}{\sqrt{1 + (yQ)^2}} = \frac{1}{\sqrt{1 + (1.412 \cdot 50)^2}} = 0.0142$$

$$A_r(\text{dB}) = 20 \log A_r = 20 \log 0.0142 = -37.0 \text{ dB}$$

Adding a second tuned circuit with a Q of 50 would double this to -74 dB.

Another phenomenon that is related to the image problem is that of *double spotting*. This is usually more of a problem for receivers with a low value of IF. This means that the image frequency is near to the signal frequency, and image rejection is not as good as it could be. When the receiver is tuned across the band, a strong signal appears to be at two different frequencies, once at the desired frequency and again when the receiver is tuned to 2 times IF below the desired frequency. In this second case, the signal becomes the image, reduced in strength by the image rejection, thus making it appear that the signal is located at two frequencies in the band.

7.7 Spurious Responses

Spurious responses occur when a signal with a frequency near that of the desired frequency passes through the RF amplifier to the mixer with an appreciable amplitude and either it or one of its harmonics mixes with the oscillator or one of its harmonics to produce a frequency within the band-pass of the IF filter. A nonlinear mixer will produce these harmonics, but a linear multiplier circuit used as a mixer will not. Increasing the selectivity of the RF amplifiers by increasing the Q (more tuned circuits) will reduce the effect of spurious response by keeping the interfering signals from reaching the mixer.

The output from the mixer in general is

$$\text{IF} = \pm n \cdot f_o \mp m \cdot f_u \quad (7.7.1)$$

where f_o = oscillator frequency

f_u = unwanted signal frequency

m = harmonic number of unwanted signal (positive integer)

n = harmonic number of oscillator (positive integer)

This equation may be rearranged to solve for all the unwanted signals that may occur for a given oscillator setting. This may be repeated for a number of signal frequencies across the band and the unwanted signals plotted on a linear chart of unwanted signal frequencies versus oscillator frequency. Each pair of harmonic numbers (m, n) generates two lines, one each for the sum and the difference. The fundamental pair (1, 1) generates the desired frequency line and the image frequency line. All lines for which $m = n$ lie parallel to the signal line, while those for m not equal to n intersect the signal line. Any signals that fall within the band-pass of the RF stages about the signal line will produce spurious responses if signals are received at the unwanted frequencies. Figure 7.7.1 illustrates the major spurious responses produced in the receiver of

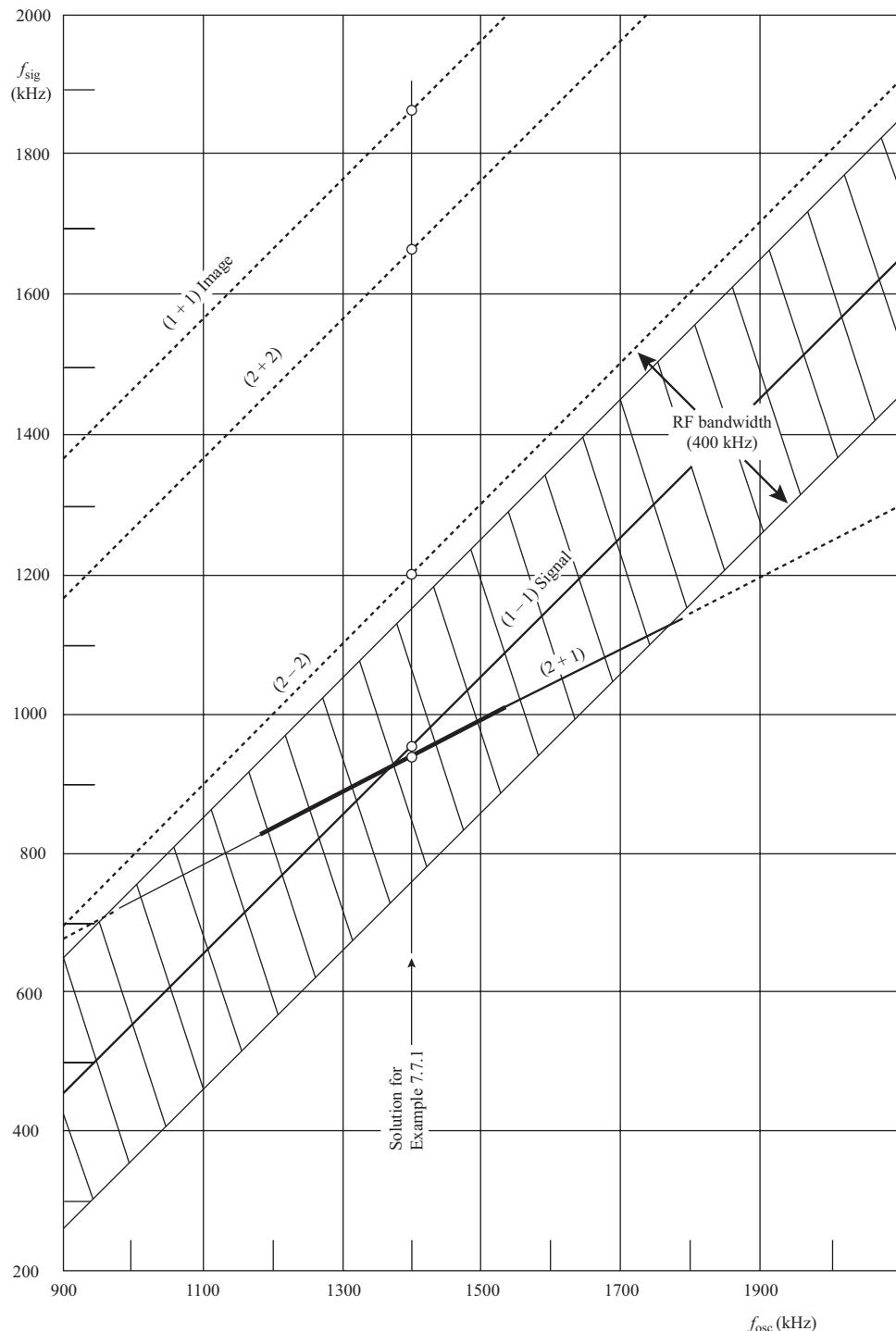


Figure 7.7.1 Spurious response chart for a receiver with a 455-kHz IF.

Example 7.7.1. The line produced by the oscillator second harmonic ($2 - 1$) crosses the signal line. The heavy portion of that line represents the spurious signals, which are most troublesome.

EXAMPLE 7.7.1

The broadcast receiver in Example 7.3.1 is tuned to a signal at 950 kHz. Find all the unwanted signal frequencies for harmonics up to the second. Which of these fall within 200 kHz of the desired frequency?

SOLUTION

$$m = 1, 2, \quad n = 1, 2, \quad \text{IF} = 455 \text{ kHz}$$

$$f_o = f_s + \text{IF} = 950 + 455 = 1405 \text{ kHz}$$

$$f_u = \frac{n}{m} f_o \pm \frac{1}{m} \text{IF}$$

<i>m</i>	<i>n</i>	<i>Sum</i>	<i>Difference</i>
1	1	1860 (image)	950 (signal)
1	2	3265	2355
2	1	930	475
2	2	1632.5	1177.5

Only one unwanted signal falls within the band-pass of the RF stages, that is when the second harmonic of the signal mixes with the oscillator fundamental to give 930 kHz. A strong signal near 930 kHz will produce an interference in the receiver. Preventing the generation of the second harmonic of the oscillator frequency will eliminate the major component of the spurious response ($2 + 1$).

7.8 Adjacent Channel Selectivity

The selectivity of the RF stages of a receiver are a function of frequency, being narrowest at low frequencies and increasing with frequency. It is also difficult to get several high-*Q* tuned circuits to track properly when tuned over a wide range. For this reason the selectivity of the RF stages of most receivers is purposely left much wider than is necessary for single-channel operation, and the final selectivity is obtained in the IF amplifier.

Channel assignments in the crowded spectrum are made as close together as possible, with 10-kHz spacing for AM signals in the MF and HF bands. Spacing for FM broadcast in the VHF band is 200 kHz, while that for TV in the VHF and UHF bands is 6 MHz. It should be possible for a receiver to separate two signals occupying adjacent channels without interference between them or without compromising receiver performance.

The ideal IF channel band-pass characteristic is illustrated in Fig. 7.8.1. This has a flat-topped response within the bandwidth centered on the channel frequency f_1 and ideally complete rejection outside the bandwidth, including the adjacent channel at f_2 . Practical filters are far from ideal, and more or less of the adjacent channel signal will pass through to the detector. A good receiver should provide 60 to 80 dB of rejection of the adjacent channels, or more in a high-quality receiver.

In the past, IF selectivity has been obtained by using several cascaded high-*Q* tuned circuits in different combinations. One system uses an IF system consisting of two or more amplifiers connected by undercoupled transformers with both primary and secondary tuned to the IF. Each tuned circuit contributes a single resonance curve, and the overall IF response is the product of all of them, as shown in Fig. 7.8.2. The *Q*'s are chosen so

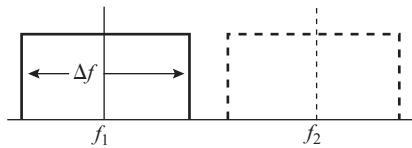


Figure 7.8.1 Ideal IF band-pass characteristic showing a rejected adjacent channel.

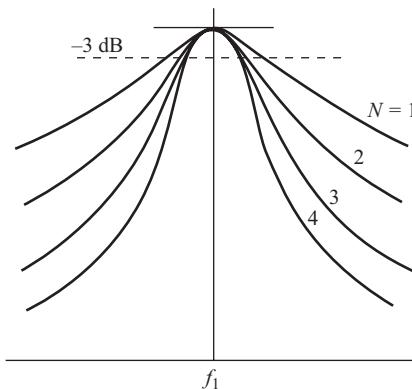


Figure 7.8.2 Band-pass response of cascaded single-tuned circuits.

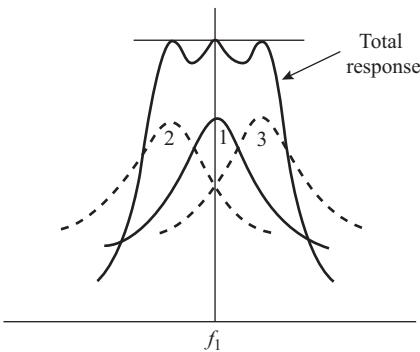


Figure 7.8.3 Band-pass response of stagger tuned circuits.

that the overall response has a -3-dB bandwidth of the required band-pass, and steeper skirts are obtained for each additional tuned circuit. The system suffers somewhat from amplitude and phase distortion over the band pass, but the use of five tuned circuits (three transformers) produces adequate selectivity.

Better in-band distortion characteristics can be obtained with the preceding system by using *stagger tuning*. An odd number of tuned circuits are used, and these are tuned so that one is on the center frequency and each successive pair is tuned to be equidistant and progressively farther from the center frequency, as shown in Fig. 7.8.3. The overall response is again the product of the individual responses, but this time the top has several peaks, forming a ripple. This ripple can be smoothed by adding more tuned circuits with their peaks tuned closer together. The steepness of the skirts of the response is dependent on the total number of

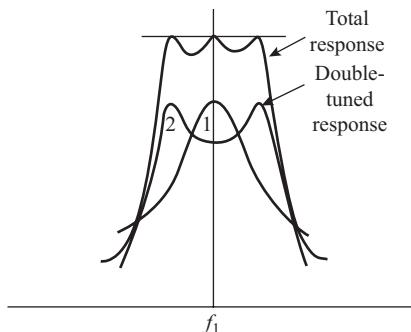


Figure 7.8.4 Band-pass response of overcoupled double-tuned circuits.

tuned circuits used. In both of these systems, the tuned circuits must be isolated from each other so that a five-hump response would require five tuned circuits separated by four amplifiers.

The overcoupled double-tuned transformer described in Section 1.8 offers a unique variation to the stagger tuning problem. When two tuned circuits tuned to the same frequency are tightly coupled to form a transformer, the overall transformer response has a double hump. (The same transformer with looser coupling will only have a single-hump response.) The distance of the hump frequencies from the center frequency is dependent on the degree of coupling, and the steepness of the skirts is dependent on the circuit Q . Two or more of these transformers are cascaded and adjusted to give the same type of response as that for stagger tuning, as shown in Fig. 7.8.4. Typically, three such transformers are used with two amplifiers, and one of the three is undercoupled. The result is a five-hump characteristic with one hump on the center frequency. The ripple amplitude is typically less than 1 dB.

Many of the more expensive communications receivers use a specially designed IF band-pass filter in conjunction with an integrated circuit amplifier to build the IF system. One form of filter that was used extensively until recently is a *mechanical filter* such as that made by the Collins Radio Company, which makes use of the mechanical resonance properties of materials. These filters (described in Section 1.15) worked well for low-frequency IFs but tended to be bulky and expensive. *Crystal lattice filters* composed of quartz piezoelectric crystals of the same type used for oscillator control are also used. These have the advantage of small size and weight, but they require custom manufacture to produce good characteristics. A more recent device that is finding wide acceptance in compact receivers made with integrated circuits is the integrated piezoelectric *ceramic filter*. Thick-film integrated-circuit techniques are used to manufacture these, making them small and inexpensive.

Integrated-circuit IF amplifiers have become available, making it possible to build entire receivers with only one or two chips. These amplifiers are complete broadband amplifiers with no tuning and must be used with external IF filters to obtain the desired band-pass characteristics.

7.9 Automatic Gain Control (AGC)

When a receiver without automatic gain control (AGC) is tuned to a strong station, the signal may overload the later IF and AF stages, causing severe distortion and a disturbing blast of sound. This can be prevented by using a manual gain control on the first RF stage, but usually some form of AGC is provided. The AGC derives a varying bias signal that is proportional to the average received signal strength and uses this bias to vary the gain of one or more IF and/or RF stages. When the average signal level increases, the size of the AGC bias increases, and the gain of the controlled stages decreases. When there is no signal, there is a minimum AGC bias, and the amplifiers produce maximum gain.

Simple AGC is used in most domestic receivers and cheaper communications receivers. In simple AGC receivers, the AGC bias starts to increase as soon as the received signal level exceeds the background noise level, and the receiver immediately becomes less sensitive. The AM detector used in these receivers is a simple half-wave rectifier that produces a dc level that is proportional to the average signal level. This dc level is put through an *RC* low-pass filter to remove the audio signal and then applied to bias the base of the IF and/or RF amplifier transistors. The time constant of the filter must be such that it is at least 10 times longer than the period of the lowest modulation frequency received, which is usually around 50 Hz, or about 0.2 s. If the time constant is made longer, it will give better filtering, but it will cause an annoying delay in the application of the AGC control when tuning from one signal to another.

The circuit of Fig. 7.9.1 uses a time constant of about 0.25 s. The circuit shown uses the main signal detector diode for two purposes, detection and the provision of AGC bias. Some compromise in the

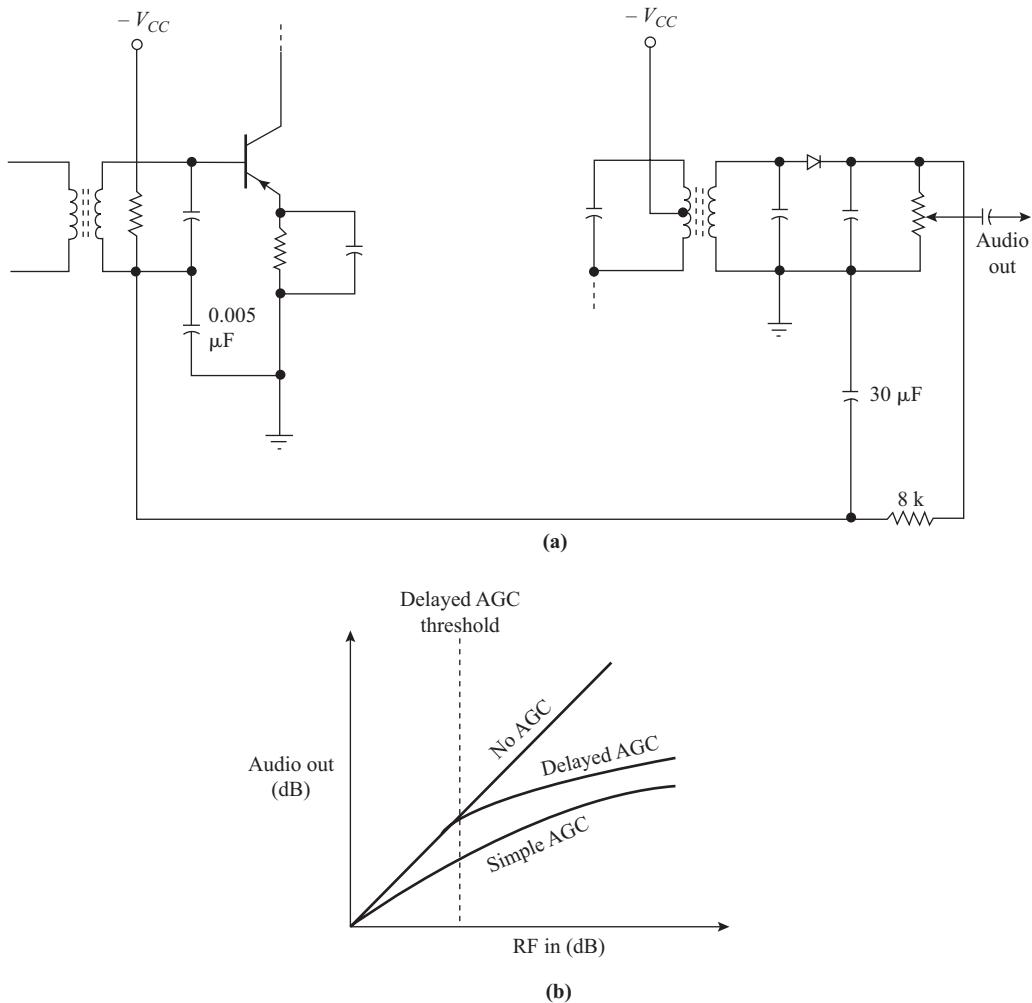


Figure 7.9.1 Automatic gain control. (a) Simple AGC applied to an IF amplifier. (b) Response of a receiver with either simple or delayed AGC compared to one without AGC.

performance of each function is necessary, and in better receivers a separate detector is used for the AGC. Also, signals may be picked off earlier in the IF and fed to a separate IF amplifier to supply the AGC, thus reducing the loading on the IF circuits.

The point in the receiver at which the signal is sampled is important. The best place is after the IF stage filters, but before any limiting amplifiers. If the sample is taken before these filters, then any strong signals in adjacent channels will cause the AGC to engage before the desired signal is up to strength and desensitize the receiver.

Better response can be obtained by including more gain in the feedback loop. This is accomplished by providing a dc amplifier stage after the AGC filter section. This *AGC amplifier* produces a low enough source resistance so that several RF/IF stages may be easily driven from the same AGC line.

Connection of AGC to the RF stage in transformer-coupled circuits is easily accomplished. The lower end of the input transformer secondary is isolated from ground with a bypass capacitor and connected directly to the AGC line. A resistance back to the collector supply provides the bias current to the base necessary to maintain full gain when a very low signal is present. When several stages operating at the same frequency are connected to the same AGC line, decoupling between them must be used to prevent instability. This is done by feeding the AGC to the earlier stage through a second filter section with the same time constant and providing good local bypassing at each stage input point.

Delayed AGC is used in most of the better communications receivers. Delayed AGC is obtained when the generation of the AGC bias is prevented until the signal level exceeds a preset threshold and then increases proportionally after that, so that the maximum sensitivity of the receiver can be realized when receiving low-level signals. The threshold may be fixed by the circuit design or may be adjustable. The threshold is usually adjusted so that AGC starts taking effect when the signal has risen nearly to the level that produces the receiver maximum output under maximum sensitivity conditions (that is, under full gain). A delayed AGC response characteristic is illustrated in Fig. 7.9.1(b), where it is compared with response for no AGC and for simple AGC.

Delayed AGC is easiest to achieve when an AGC amplifier is included in the circuit. In this case the amplifier is biased beyond cut-off by a fixed bias source, from which the detected AGC bias voltage is subtracted. The AGC level then must overcome the fixed bias before any bias is passed on to the controlled amplifiers.

7.10 Double Conversion

The front-end selectivity of any receiver must reject the first image frequency located at twice the IF above the desired signal frequency, and for this to occur the IF must be high, just below the signal band. However, as the IF is made higher, its bandpass becomes larger (because of the changing Q of the tuned circuits). Beyond the HF band (30 MHz) it becomes impossible to obtain the required band-pass and image rejection using ordinary tuned circuits. As a result, single-conversion superheterodyne receivers are seldom used above about 20 MHz.

The *double-conversion receiver* allows the receiver to have good image rejection and also good adjacent channel selectivity. The image rejection is obtained in the first conversion by assuring that the image frequency is well outside the fixed-tuned RF amplifier band-pass. The narrow band pass required for good adjacent channel rejection is obtained in the second IF, which may be as low as 100 kHz. Figure 7.10.1(a) shows the block diagram of a double-conversion receiver that might be used for the 150-MHz FM mobile band. A high first IF of 10.7 MHz and a bandwidth of 150 kHz allow sufficient selectivity in the RF stage tuning to reject the image of the first IF at 171.4 MHz, while at the same time passing a band of adjacent channels. The 150-kHz band pass of the first IF filter allows several 15-kHz-wide adjacent channels to pass to the second mixer. Channel selection is obtained by changing the first local oscillator frequency, either by switching crystals or by reprogramming a digital frequency synthesizer. The oscillator will be near the 10-MHz range, followed by a series

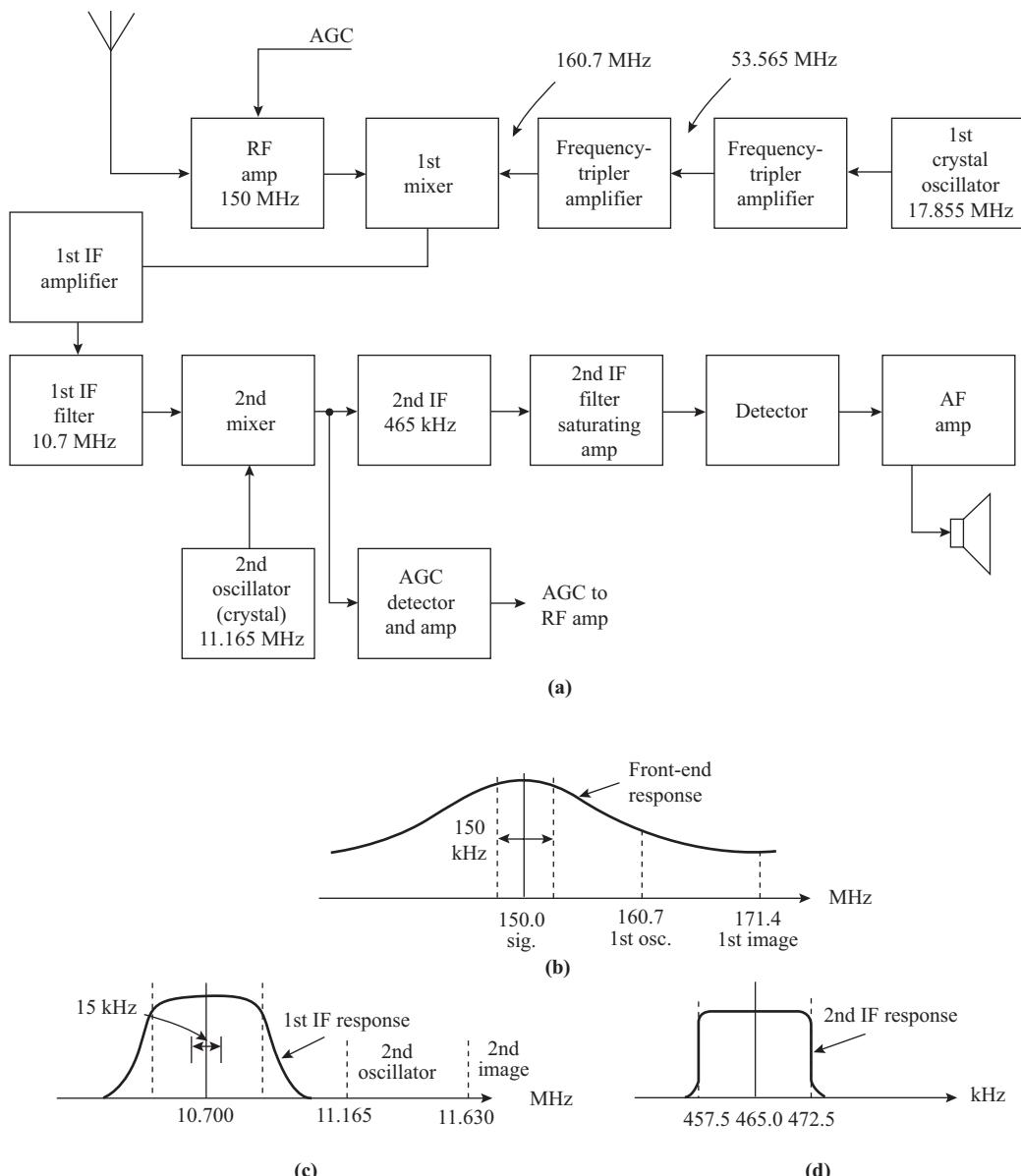


Figure 7.10.1 Double-conversion superheterodyne receiver: (a) block schematic, (b) RF stage response, (c) first IF stage response, and (d) second IF stage response.

of frequency multipliers to obtain the 160-MHz range local oscillator frequency. Figure 7.10.1(b) illustrates this RF stage selectivity.

The first IF filter is a filter block designed with a band pass of about 150 kHz centered on 10.7 MHz, such as would be used in FM receivers. This block may be a crystal lattice filter or an integrated piezoelectric ceramic crystal filter. While this filter will pass a band of several channels, it will not pass the second converter image

frequency located at 11.63 MHz. An amplifier stage in the first IF section compensates for filter and mixer losses in the first converter. Figure 7.10.1(c) shows the band-pass characteristic of the first IF stage.

Figure 7.10.1(d) shows the band pass for the second IF, which rejects the adjacent channels. The second IF filter is again a ceramic filter unit centered on 465 kHz, with a band pass of 15 kHz. The second local oscillator is fixed-tuned at 11.165 kHz, just above the band-pass cutoff of the first IF filter. Many communications receivers used for frequencies above 30 MHz use frequency modulation, so the second IF will be followed by a limiting amplifier. A separate delayed AGC detector picks off the signal before the limiting amplifier to control the RF and first IF stages.

A very useful variation on the double-conversion receiver is to use an upconversion in the first mixer, placing the IF above the received RF band. In this case, Eq. (7.2.1) becomes

$$\text{IF} = f_0 + f_s \quad (7.10.1)$$

The effect of this is to place virtually all the usual image responses above the IF and thus well above the received band. This eliminates the need to tune the RF amplifier within the RF reception band. Only a low-pass filter with its cutoff frequency somewhat above the reception band is required, completely eliminating tracking problems. [Note that the band-pass cutoff must be below the half-IF frequency (IF/2) to prevent direct feedthrough.] Only the local oscillator for the first conversion needs to be tuned, greatly simplifying the receiver circuitry. This process is ideal for receivers that extensively use integrated circuits for compactness, as is the case with mobile receivers or handheld receivers.

7.11 Electronically Tuned Receivers (ETRs)

Many modern receivers have electronic digital tuning systems on them, based on computer control of varactor tuning diodes. This is especially true for automotive receivers, where a microcomputer controller takes over many of the scanning and tuning functions, freeing the driver from distraction.

Figure 7.11.1 shows a representative AM electronically tuned receiver (ETR) tuning arrangement. The local oscillator, mixer, IF amplifiers, and detectors are all contained in a single chip, the National LM1863. The RF stages include an input FET buffer amplifier followed by a tuned input/tuned output RF amplifier. The two RF tuned circuits are varactor tuned, trimmed for tracking, and transformer coupled to the signal circuit for impedance matching. The oscillator tuned circuit is also varactor tuned and trimmed. The same type of varactor diode is used in all three tuned circuits, with the same type of trimmer capacitor, and all three are driven by the same tuning voltage. Proper tracking is accomplished by adjusting the three trimmers to obtain the desired tuning ratios.

The local oscillator in the receiver becomes the voltage-controlled oscillator for a frequency synthesizer system (see Section 6.8) included on a National DS8907 chip. In the IF chip, a sample of the local oscillator signal is amplified by a buffer amplifier and coupled to the feedback input of the phase-locked loop chip to drive the programmable divide-by-($N + 1$) counter. This produces a signal with a frequency of $f_0/(N+1) = 10$ kHz (the channel frequency spacing for the AM band), which becomes one input to the phase detector multiplier. A crystal oscillator on the same chip produces a 4-MHz reference frequency signal, which is then divided by $K = 400$ to produce a 10-kHz reference signal for the other input of the multiplier. For signal frequencies from 550 to 1600 kHz with a 450-kHz IF, the required values of N are from 99 to 204 in steps of 1, giving 106 channels each 10 kHz wide.

The two signals presented to the phase detector multiplier are both at 10 kHz once lock is achieved, but are displaced in phase by 90° plus a phase offset. This phase offset produces a dc bias voltage that is filtered and used to drive the local oscillator varactor, tuning it to the lock frequency at $(N + 1) \times 10$ kHz. The same voltage also drives the RF circuit tuning varactors, which are trimmed to track the local oscillator.

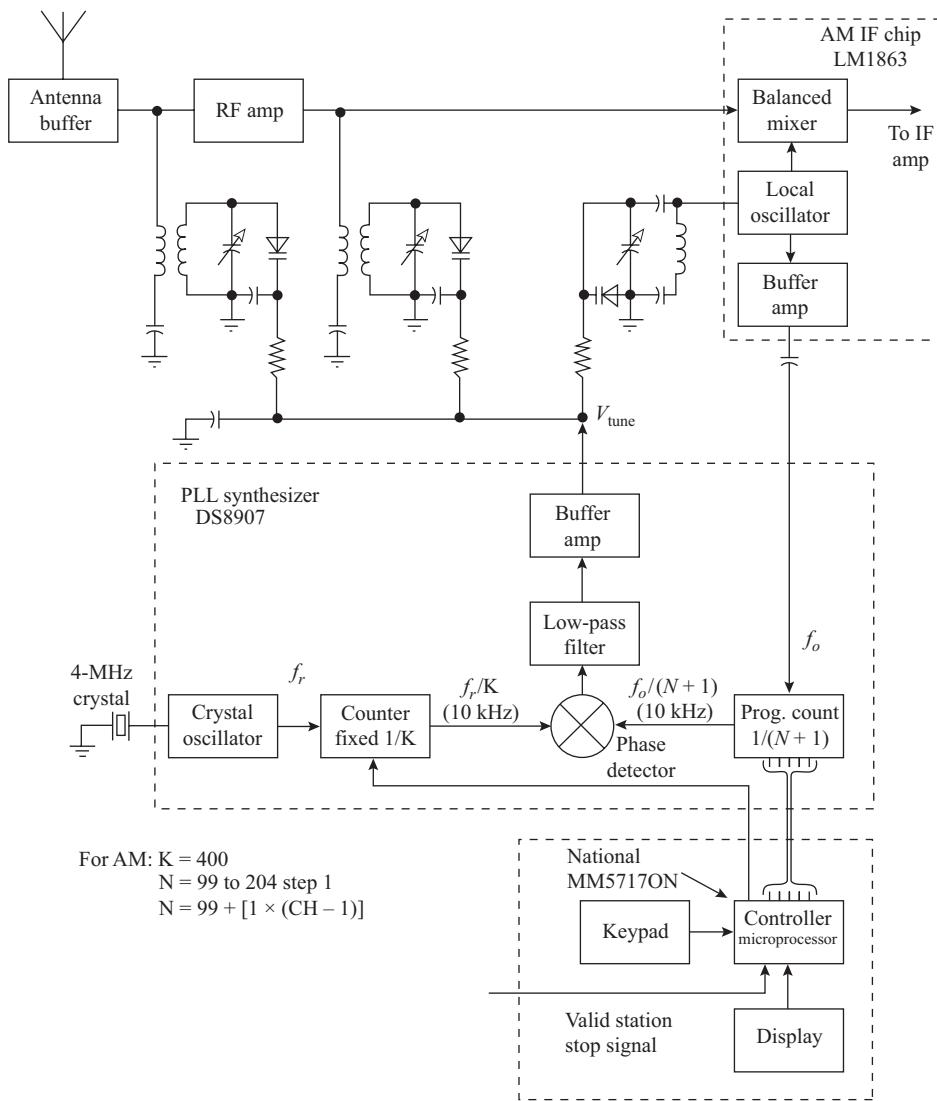


Figure 7.11.1 Electronic tuning system for an AM receiver.

Control of the tuning system is accomplished using a dedicated microprocessor chip. This chip is controlled by a numerical keypad and some control switches and produces a panel LED display of the actual signal frequency to which the receiver is tuned. In a manual mode, the frequency of the desired channel may be entered through the keypad, and the microprocessor will calculate the required value of N and pass it to the programmable counter in the synthesizer chip.

Automatic band scanning is a desirable feature in automotive applications, especially where local station frequency assignments are not known. The signal detector on the IF chip produces a logic high condition (called the *valid station stop*) when a strong signal appears in the IF frequency window while scanning. This signal tells the computer to stop scanning and lock on the current frequency. Automatic scanning is initiated by pressing a

scan start button on the panel. A software channel counter in the control computer starts counting up (or down) from the current channel number and continues until a strong signal produces the valid station stop condition. The value of N at each step of the channel counter is computed from the channel number by the equation

$$N = 99 + 1 \times (\text{CH} - 1) \quad (7.11.1)$$

where CH is the channel number, with values from 1 to 106. When the scanner tunes to a valid channel, indicated by the presence of a carrier to produce the stop signal, scanning stops and the receiver locks to that signal. Because the AGC circuits in the receiver do not respond instantaneously, it is necessary for the scan control to stop on each channel number long enough for the AGC to respond and give a true level for the valid station stop signal. This is typically on the order of more than 100 ms, the same as the AGC time constant.

In an automotive receiver, both AM and FM signals usually are required. Hence the receiver system will include two complete receivers, one for AM and one for FM. The National LM1865 chip is a complete mixer/IF/detector system for an FM receiver, while the LM1863 is the equivalent for an AM receiver. Both the LM1863 and the LM1865 chips are included, with separate tuners for AM and FM. Both receivers are serviced by the same tuning control system and synthesizer. Under control of the microprocessor when FM is desired, the audio output signals are switched, and the value of K is changed to 160. The values of N range from 3951 to 4743 in steps of 8 for the FM band 88 to 108 MHz with a 10.7 MHz IF, embracing 100 channels each 200 kHz wide. The signals at the phase comparator input are at 25 kHz, the eighth submultiple of the channel spacing of 200 kHz. The value of N is computed as

$$N = 3951 + [8 \times (\text{CH} - 1)] \quad (7.11.2)$$

where CH is the channel number, with values from 1 to 100. (Note that the frequency synthesizer divides by $N + 1$.)

7.12 Integrated-circuit Receivers

The previous section describes an electronic tuning system for use with integrated-circuit receiver subsystems such as the LM1863. This section describes such a system in more detail.

Figure 7.12.1 shows a complete AM electronically tuned receiver (not including the audio circuits) built around the National LM1863 chip. As indicated by the block diagram in Fig. 7.12.2, this chip includes the mixer, the IF amplifier, the detector, the local oscillator, the AGC circuits, and the valid station stop detector.

The receiver uses a discrete-component two-stage RF tuner with two tracking tuned circuits to obtain sufficient image rejection. The antenna is coupled through a fixed-tuned broadband pass filter $T1$ to help to reject the image and spurious responses. The first stage is an FET buffer $Q1$ operating in common-source mode with the drain circuit tuned. The tuned circuit is a transformer $T2$ with the secondary tuned by the varactor diode $D1$ and coupled through the primary to the signal circuit to match impedances. The second stage is a common-emitter amplifier $Q2$ with its collector tuned by the transformer $T3$ and varactor $D2$. Both stages share a common bias current source, which is modulated by the RF AGC voltage. The RF amplifier output signal is connected through coupling capacitor $C28$ to the mixer input on pin 18, along with the IF AGC voltage through $R22$.

The mixer is a doubly balanced multiplier-type mixer similar to that on fig. 5.10.5, operated in linear mode so that no second harmonics of the RF and oscillator signals are generated. The mixer output appears on pin 9 and is coupled through singly tuned transformer $T4$ for impedance matching into a 450-kHz ceramic IF filter block with a 20-kHz band pass, to the IF amplifier input on pin 10.

The local oscillator is tuned by inductor $T6$ and varactor $D3$ connected at pin 16. An internal level detector and bias circuit control the output level of the oscillator as its frequency changes to maintain a constant output level. This reduces the harmonic output of the oscillator to reduce spurious responses. The oscillator output

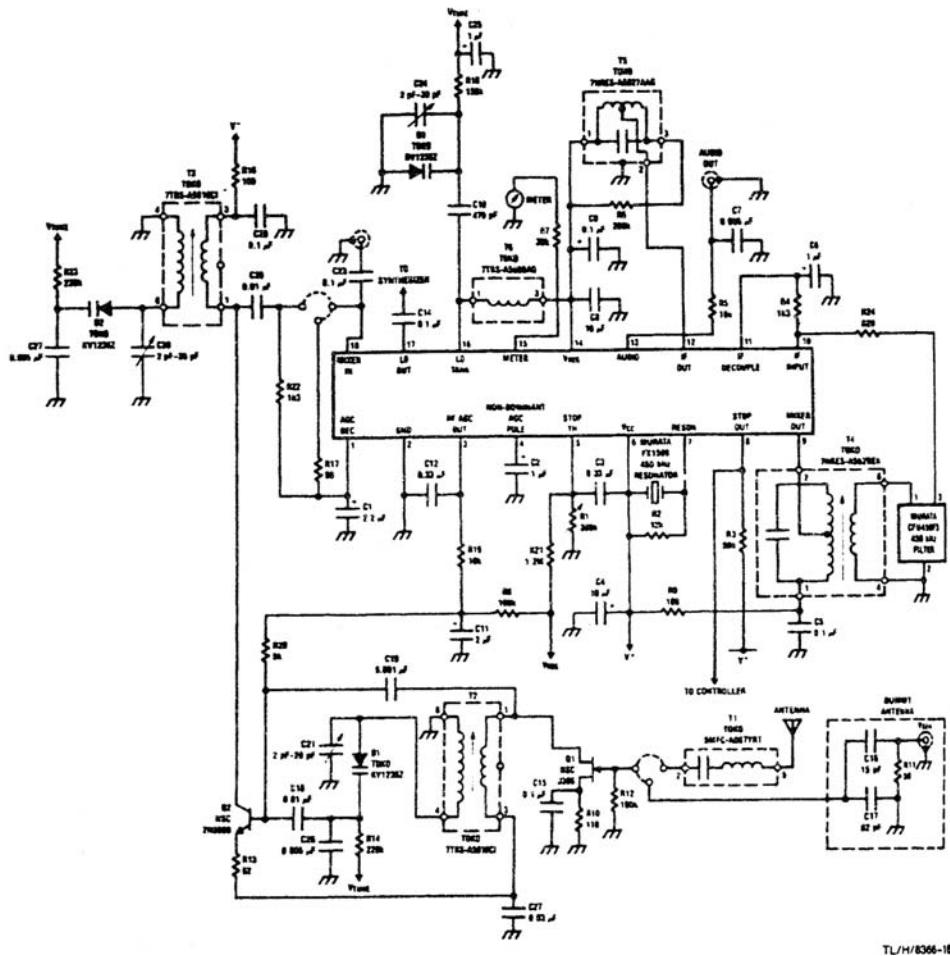


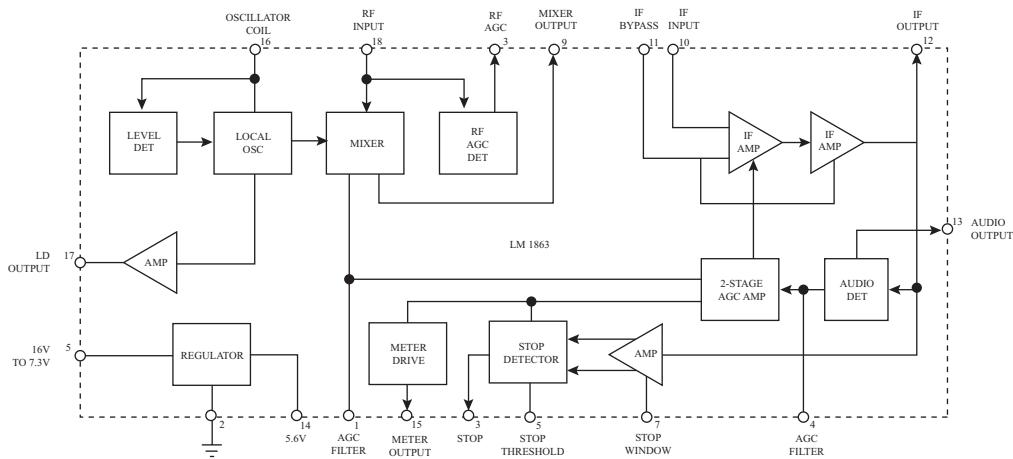
Figure 7.12.1 Electronically tuned AM receiver using National LM1863 mixer/oscillator/IF/detector chip. (By permission of National Semiconductor Corporation.)

is internally coupled to one input of the mixer and also passed through a buffer amplifier to the output pin 17 for feedback to the synthesizer for the phase-locked-loop frequency control.

The IF amplifier is a single-stage differential amplifier operated in single-ended input mode, followed by a common-emitter amplifier whose collector is tuned by a single tuned circuit T_5 connected at pin 12 (the IF amplifier output). The emitter current source of the differential pair is used to apply AGC control.

The IF output is internally coupled through a buffer amplifier to a peak-rectifying envelope detector. The detected audio output is passed through a buffer amplifier to pin 13 and on to successive audio stages.

A sample of the detector output dc level (which is proportional to the carrier level) is separately amplified and passed through a two-stage filter to provide AGC to the mixer stage, which is coupled through pin 1 and R_{22} (mentioned before). The AGC voltage at this point is passed through a second AGC amplifier and applied to the main IF differential stage, all internally. Bypass capacitors C_1 on pin 1 and C_2 on pin 3 complete the two AGC filter blocks. The IF AGC threshold is set so that an input signal level of $30\mu\text{V}$ will give 20-dB



LM1863 Complete Block Diagram

TL/H/8366-6

Performance Summary

Static Characteristics

Supply Current	8.2 mA
Operation Voltage	7.0 – 16V

Dynamic Characteristics

$f_{MOD} = 1\text{kHz}, f_o = 1.0 \text{ MHz } 30\% \text{ MOD}$	
Maximum Sensitivity	2.2 μV
20 dB Quieting	30 μV^*
S/N (10 mV Input)	54 dB
THD	0.2%
Audio Output	125 mV
Stop Signal Threshold	50 μV^{**}
Stop "Window"	4 kHz
Stop Time	<50 ms
RF Bandwidth	28 kHz
Image Rejection	>70 dB
RF AGC Threshold with 16 dB Antenna Pad	3 mV'
External Adjustable	

Figure 7.12.2 Block schematic of the national LM1863 AM ETR chip with its performance characteristics. (By permission of National Semiconductor Corporation.)

suppression of the noise level. A sample of the AGC voltage is also passed to a driver amplifier, producing a voltage to drive an external dc voltmeter to indicate received signal strength at pin 15.

A second AGC detector samples the RF signal at the mixer input, and when this level exceeds 6 mV, it puts out the RF AGC voltage on pin 3 to desensitize the RF amplifiers and prevent overloading the mixer. This voltage is applied to the RF stage after filtering by R19, C11.

A sample of the IF amplifier output at pin 12 is also internally connected to an amplifier tuned to the IF center frequency at 450 kHz. The tuning is accomplished by a ceramic resonator connected to pin 7, in parallel with 12-k Ω resistor R2. This resistor has the effect of spoiling the Q of the resonator so that a 4-kHz-wide "window" is established about the center frequency. When a 450-kHz signal is in the window, it is amplified and passed to a peak detector to produce a bias signal. This is compared with the AGC voltage, and if both are

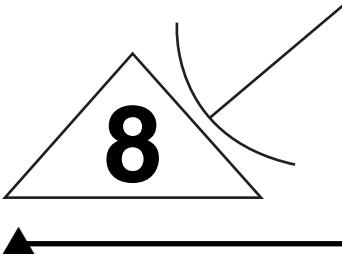
present, a logic low output signal (the valid station stop signal) is passed through pin 8 to the tuning controller to stop automatic scanning. A voltage divider $R21$ and potentiometer $R1$ present a preset voltage to pin 5 to allow adjustment of the signal threshold level, at which scan stop signal occurs.

The companion chip to the LM1863 is the LM1865. This is a complete integrated IF/oscillator/mixer/detector system for an FM receiver with a 10.7-MHz IF. This chip also uses a discrete-component RF tuner using varactors, and the two chips can be installed in the same unit to allow the choice of either AM or FM reception. Both are controlled by the tuning system discussed in Section 7.11. The reader is referred to the National Semiconductor Corporation Linear Applications Data Book for details (AN-382, page 973).

PROBLEMS

- 7.1. A receiver tunes the 3- to 30-MHz HF band in one range, using an IF of 40.525 MHz. Find the range of oscillator frequencies, the range of image frequencies, and the type of filters needed to make the receiver function properly.
- 7.2. A superheterodyne receiver tunes the range from 4 to 10 MHz with an IF of 1.8 MHz. A triple gang tuning capacitor with a maximum capacity of 125 pF per section is used. Find (a) the RF circuit coil inductance, (b) the RF circuit frequency tuning ratio, (c) the RF circuit capacitance tuning ratio, (d) the required minimum capacity per section, (e) the oscillator maximum and minimum frequency and the frequency tuning ratio, and (f) the oscillator capacitance tuning ratio.
- 7.3. The oscillator section of the receiver in Problem 7.2 uses a padder capacitor to set its tuning range. Find the value of the padder capacitor and the oscillator inductor.
- 7.4. Repeat Problem 7.3 using a trimming capacitor.
- 7.5. Find the values of C_t and C_p for three-point tracking for the receiver of Example 7.4.1 using a 350 pF per section tuning capacitor and a crossover frequency at 1000 kHz. (Note: This is a more difficult problem.)
- 7.6. A double-conversion receiver uses a first IF = 10.7 MHz (up conversion) and a second IF = 450 kHz. It is to receive a signal at 3.9 MHz. Find the image frequency and all spurious response frequencies to the second harmonic. Will any of them fall within ± 100 kHz of the first IF? What effect would the second harmonic of an unfiltered signal at 5.35 MHz have on the receiver?
- 7.7. (a) Find the image frequency range for the receiver in Problem 7.2. Do any of the image frequencies fall in the receiver passband? (b) If the RF circuits have combined effective Q of 50 at the top end of the band, find the image rejection ratio in decibels at that frequency.
- 7.8. For the receiver of Fig. 7.10.1, assume that the RF stage contains two parallel tuned circuits, each with a $Q = 20$, tuned to the center channel at 151.5 MHz. (a) Find the relative response to the low-end channel located at 150 MHz. (b) Find the image frequency and image rejection ratio for the low-end channel in decibels.
- 7.9. (a) Find all the spurious response frequencies for the receiver in Problem 7.2 when it is tuned to a signal of 7.3 MHz, assuming that harmonic components above the third are insignificant. (b) Which ones fall within ± 2 IF of the signal frequency?
- 7.10. For the frequency synthesizer of Fig. 7.11.1, operating in the AM band from 550 to 1600 kHz with a 450-kHz IF, find the value of K if the PLL is to lock at 10 kHz. A 1-MHz crystal oscillator is to be used. Also find the minimum and maximum values of N for 106 channels each 10 kHz wide. What is the spacing between adjacent channel values of N ?

- 7.11. Repeat Problem 7.10 for the FM band 88 to 108 MHz, with IF = 10.7 MHz and locking on 25 kHz. The band has 99 channels each 200 kHz wide.
- 7.12. The tuning system in Fig. 7.11.1 uses a DS8907 synthesizer chip. If this chip is replaced by a DS8906, which is designed to lock at 0.5 kHz for AM and 12.5 kHz for FM, repeat Problem 7.10.
- 7.13. Repeat Problem 7.11 using a DS8906 synthesizer.
- 7.14. A receiver is tuned to the 4–30MHz HF band in one range using an IF of 35.525MHz. Find the range of oscillator frequencies, the range of image frequencies, and the type of filters needed to make the receiver function properly.
- 7.15. When a superheterodyne receiver is tuned to 1060kHz, its local oscillator provides the mixer with an input 1415kHz. (a) What is its image frequency? (b) The antenna of this receiver is connected to the mixer via a tuned circuit whose loaded Q is 60. What will be the rejection ratio for the calculated image frequency?
- 7.16. Calculate the image rejection of a receiver having an RF amplifier and an IF of 455kHz, if the Qs of the relevant coils are 60, at an incoming frequency of (a) 12000kHz and (b) 40MHz.
- 7.17. A superheterodyne receiver having an RF and an IF of 455kHz is tuned to 10MHz. Calculate the Qs of the RF and Mixer stages, if the receiver's image rejection is to be 120.
- 7.18. A superheterodyne receiver tunes to the range from 6 to 10MHz with an IF of 2.8MHz. (a) Find the image frequency range. (b) If the RF circuits have combined effective Q of 50 at the top of this band, find the image rejection ratio in dBs at that frequency.
- 7.19. Calculate the image frequency rejection of a double-conversion receiver which has a first IF of 5MHz and a second IF of 600kHz and an RF amplifier whose tuned circuit has a Q of 70 (the same as that of the mixer) and which is tuned to 30MHz signal. Give the answers in decibels.



Amplitude Modulation

8.1 Introduction

To modulate means to regulate or adjust, and in the present context it means to regulate some parameter of a high-frequency *carrier* wave with a lower-frequency information signal. The need for modulation first arose in connection with radio transmission of relatively low frequency information signals such as audio signals. For efficient transmission it was found that the antenna dimensions had to be of the same order of magnitude as the wavelength of the signal being transmitted. The relationship between frequency f and wavelength λ for radio transmission is $f\lambda = c$, where $c = 3 \times 10^8$ m/s is the velocity of light in free space. For a typical low-frequency signal of frequency 1 kHz, the wavelength would therefore be on the order of 300 km (or 188 miles), which is obviously impractical.

The problem was overcome by using the low-frequency signal to modulate a much higher frequency signal termed the *carrier* wave, because it effectively carried the information signal. The relatively short wavelength of the high-frequency carrier wave meant that efficient antennas could be constructed.

For the practical implementation of modulation, the carrier frequency has to be very much greater than the highest frequency in the modulating signal, as will be shown later when specific circuits are examined. In practice, the carrier is always sinusoidal and can be described by

$$e_c(t) = E_{c \text{ max}} \sin(2\pi f_c t + \phi_c) \quad (8.1.1)$$

The parameters that can be modulated are the amplitude $E_{c \text{ max}}$, the frequency f_c , and the phase ϕ_c . The latter two come under the general heading of *angle modulation* and are the subject of Chapter 10.

A number of different forms of amplitude modulation are in use, and it becomes necessary to distinguish between these. The original concept, which is still widely in use, for example in the medium-wave broadcast band, is referred to simply as amplitude modulation or AM (or sometimes as *standard* AM) and is described in the following sections. As mentioned in Section 2.6, where simple harmonic functions are used,

the cosine function is normally taken as reference [see Fig. 2.6.1(c)]. This practice will be followed here, although the term *sinusoidal* will be used when referring to either sine or cosine functions.

8.2 Amplitude Modulation

In amplitude modulation a voltage proportional to the modulating signal is added to the carrier amplitude. Let the added component of voltage be represented in functional notation as $e_m(t)$; then the modulated carrier wave is given by

$$e(t) = [E_{c \max} + e_m(t)]\cos(2\pi f_c t + \phi_c) \quad (8.2.1)$$

The term $[E_{c \max} + e_m(t)]$ describes the *envelope* of the modulated wave. Figure 8.2.1 shows (a) an arbitrary modulating signal, (b) a carrier wave, and (c) the resulting AM wave, where the envelope is seen to follow the modulating signal waveform. This also illustrates graphically why the term *carrier* is used.

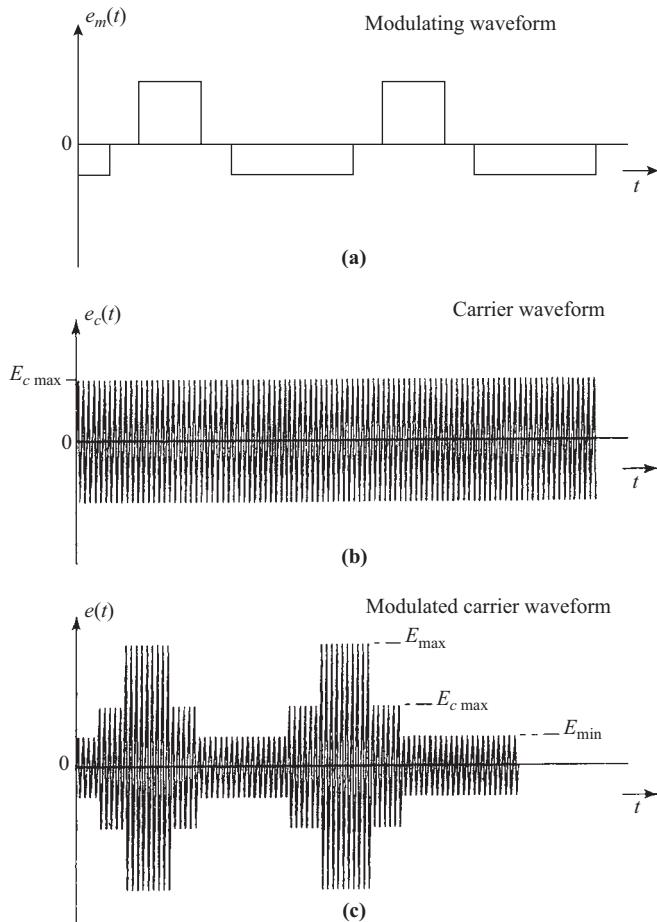


Figure 8.2.1 (a) Modulating voltage waveform. (b) Carrier wave. (c) Modulated waveform, showing the envelope that follows the modulating waveform.

8.3 Amplitude Modulation Index

Referring to Fig. 8.2.1(c), the *modulation index* is defined as

$$m = \frac{E_{\max} - E_{\min}}{E_{\max} + E_{\min}} \quad (8.3.1)$$

In the case of a periodic modulating signal, such as shown in Fig. 8.2.1, it is easy to identify the maximum and minimum voltages of the modulated wave. With a nonperiodic signal, such as a speech waveform, these quantities will vary, and hence the modulation index will also vary. What is important is that the modulation index must not be allowed to exceed unity. If the modulation index exceeds unity the negative peak of the modulating waveform is clipped, as shown in Fig. 8.3.1. This is bad enough in itself, but, in addition, such clipping is a potential source of interference, as will be shown shortly.

It will be noticed that overmodulation ($m > 1$) occurs when the magnitude of the peak negative voltage of the modulating wave exceeds the peak carrier voltage. Under these conditions in practice, E_{\min} is clamped at zero, as shown in Fig. 8.3.1(b). The mathematical expression for the modulated wave, Eq. (8.2.1), is not valid under these conditions. The envelope [$E_c \max + e_m(t)$] goes negative, which mathematically appears as a phase-reversal rather than a clamped level.

Viewing the modulated waveform directly on an oscilloscope is difficult when the modulating waveform is other than periodic because of the problem of synchronizing the sweep to obtain a stationary pattern. The problem can be overcome using the *trapezoidal method* of monitoring the modulation. The trapezoidal

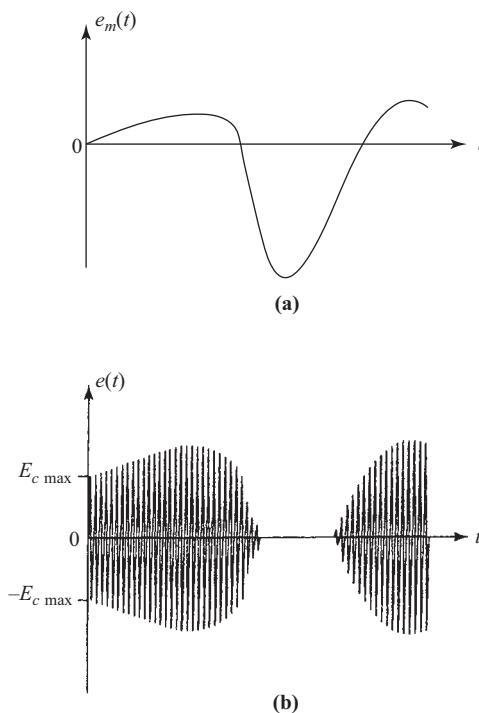


Figure 8.3.1 AM wave for which $m > 1$. (a) Modulating waveform and (b) the modulated waveform, showing clipping on the negative modulation peaks.

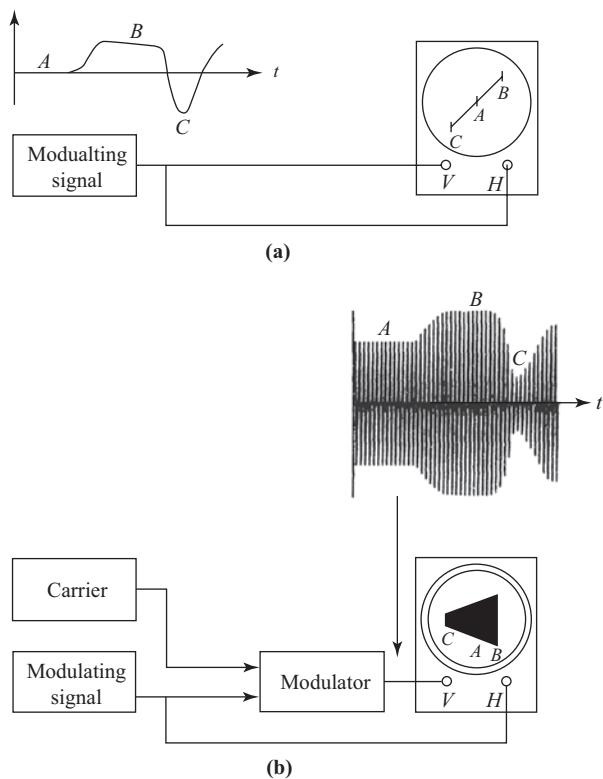


Figure 8.3.2 (a) Basic Lissajous method. (b) Method of obtaining the trapezoidal display.

method is similar to that used to produce Lissajous patterns. Figure 8.3.2(a) shows the basic Lissajous method, in which the same waveform is applied to both horizontal and vertical plates of the oscilloscope. Assuming that the horizontal and vertical gains are equal and that the spot is initially centered on the screen, the spot will be at the screen center whenever the voltage is zero, as at A.

Whenever the voltage goes positive, as at B, the spot is deflected vertically upward and horizontally to the right by equal amounts, irrespective of the waveshape, and therefore the spot traces out the upper part of the diagonal line. Likewise, whenever the voltage goes negative, as at C, the downward deflection is equal to that to the left, producing the lower part of the diagonal line.

When the modulated wave is applied to the vertical plates, the spot is deflected vertically by the carrier voltage. For example, at A in Fig. 8.3.2(b), where the modulating voltage is zero, the spot traces out a vertical line centered on the screen, proportional to the peak-to-peak carrier voltage. As the modulating voltage goes positive, as at B, the peak-to-peak voltage of the modulated wave increases while the spot is deflected to the right. The trace is therefore trapezoidal, rather than just a single diagonal line. Likewise, when the modulating voltage goes negative, as at C, the peak-to-peak voltage decreases, while the horizontal deflection is to the left, resulting in the trapezoidal pattern continuing to the left.

Figure 8.3.3 shows a number of the patterns that can be obtained. Figure 8.3.3(a) shows the normal pattern from which the modulation index is easily obtained. Denoting the peak-to-peak voltage by E_{pp} , the longest vertical displacement is $L_1 = E_{pp \max}$ and the shortest is $L_2 = E_{pp \min}$. But since $E_{pp \max} = 2E_{\max}$ and $E_{pp \min} = 2E_{\min}$, the trapezoidal display gives, on canceling out the common factor 2,

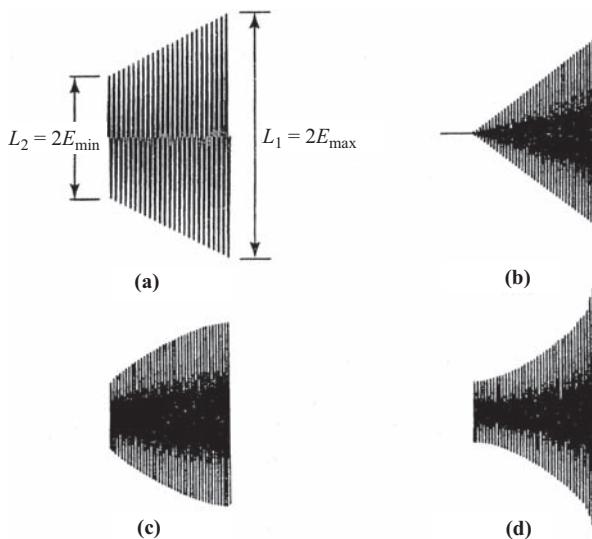


Figure 8.3.3 (a) Normal trapezoidal pattern. (b) Trapezoidal pattern for $m > 1$. (c) Envelope distortion resulting from insufficient RF drive to the modulator. (d) Envelope distortion resulting from non-linearity in the modulator.

$$\begin{aligned}
 m &= \frac{E_{\max} - E_{\min}}{E_{\max} + E_{\min}} \\
 &= \frac{L_1 - L_2}{L_1 + L_2}
 \end{aligned} \tag{8.3.2}$$

Figure 8.3.3(b) shows the pattern obtained when overmodulation occurs. When E_{\min} is zero, the length L_2 becomes zero and the trapezoid reduces to a triangle. Overmodulation results in a spike being produced at the L_2 point of the triangle, since the carrier voltage is cut off completely. (In some practical situations, leakage of the carrier may occur through the circuit, resulting in a blurring of the spike.)

It will be seen that the modulation index is zero when $E_{\max} = E_{\min} E_c \text{ max}$, and it is unity when $E_{\min} = 0$. Thus, in practice, the modulation index should be in the range

$$0 \leq m \leq 1 \tag{8.3.3}$$

Figures 8.3.3(c) and (d) show two of the patterns obtained when envelope distortion is present. In Figure 8.3.3(c), the modulator output flattens off at high modulation levels, which could be a result of insufficient carrier input (or drive) to the modulator at these levels. Figure 8.3.3(d) shows the effect of a nonlinearity in the modulator, which indicates an accentuation of high levels of modulation relative to low levels.

EXAMPLE 8.3.1

A modulating signal consists of a symmetrical triangular wave having zero dc component and peak-to-peak voltage of 11 V. It is used to amplitude modulate a carrier of peak voltage 10 V. Calculate the modulation index and the ratio of the side lengths L_1/L_2 of the corresponding trapezoidal pattern.

SOLUTION

$$E_{\max} = 10 + \frac{11}{2} = 15.5 \text{ V}$$

$$E_{\min} = 10 - \frac{11}{2} = 4.5 \text{ V}$$

$$\therefore m = \frac{15.5 - 4.5}{15.5 + 4.5} = 0.55$$

L_1 is proportional to 15.5 V, L_2 to 4.5 V, and therefore $L_1/L_2 = 15.5/4.5 = 3.44$.

Fortunately, many of the characteristics of AM can be examined using sinusoidal modulation as described in the following sections.

8.4 Modulation Index for Sinusoidal AM

For sinusoidal AM, the modulating waveform is of the form

$$e_m(t) = E_{m \max} \cos(2\pi f_m t + \phi_m) \quad (8.4.1)$$

In general the fixed phase angle ϕ_m is unrelated to the fixed phase angle ϕ_c for the carrier, showing that these two signals are independent of each other in time. However, the amplitude modulation results are independent of these phase angles, which may therefore be set equal to zero to simplify the algebra and trigonometry used in the analysis. The equation for the sinusoidally modulated wave is therefore

$$e(t) = (E_{c \max} + E_{m \max} \cos 2\pi f_m t) \cos 2\pi f_c t \quad (8.4.2)$$

Since in this particular case $E_{\max} = E_{c \max} + E_{m \max}$ and $E_{\min} = E_{c \max} - E_{m \max}$ the modulation index is given by

$$\begin{aligned} m &= \frac{E_{\max} - E_{\min}}{E_{\max} + E_{\min}} \\ &= \frac{E_{m \max}}{E_{c \max}} \end{aligned} \quad (8.4.3)$$

The equation for the sinusoidally amplitude modulated wave may therefore be written as

$$e(t) = E_{c \max} (1 + m \cos 2\pi f_m t) \cos 2\pi f_c t \quad (8.4.4)$$

Figure 8.4.1 shows the sinusoidally modulated waveforms for three different values of m .

8.5 Frequency Spectrum for Sinusoidal AM

Although the modulated waveform contains two frequencies f_c and f_m , the modulation process generates new frequencies that are the sum and difference of these. The spectrum is found by expanding the equation for the sinusoidally modulated AM as follows:

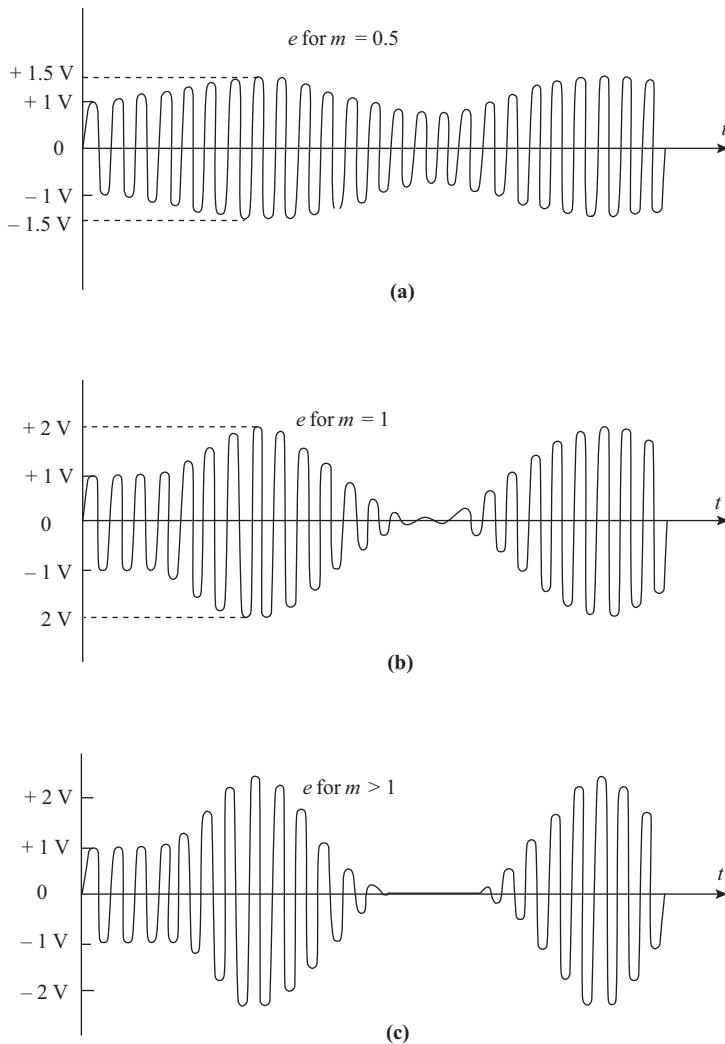


Figure 8.4.1 Sinusoidally amplitude modulated waveforms for (a) $m = 0.5$ (undermodulated), (b) $m = 1$ (fully modulated), and (c) $m > 1$ (overmodulated).

$$\begin{aligned}
 e(t) &= E_c \max (1 + m \cos 2\pi f_m t) \cos 2\pi f_c t \\
 &= E_c \max \cos 2\pi f_c t + m E_c \max \cos 2\pi f_m t \times \cos 2\pi f_c t \\
 &= E_c \max \cos 2\pi f_c t + \frac{m}{2} E_c \max \cos 2\pi(f_c - f_m)t + \frac{m}{2} E_c \max \cos 2\pi(f_c + f_m)t
 \end{aligned} \tag{8.5.1}$$

It is left as an exercise for the student to derive this result making use of the trigonometric identity

$$\cos(A \pm B) = \cos A \cos B \mp \sin A \sin B \tag{8.5.2}$$

Equation (8.5.1) shows that the sinusoidally modulated wave consists of three components: a carrier wave of amplitude $E_c \text{ max}$ and frequency f_c , a *lower side frequency* of amplitude $mE_c \text{ max}/2$ and frequency $f_c - f_m$, and an *upper side frequency* of amplitude $mE_c \text{ max}/2$ and frequency $f_c + f_m$. The amplitude spectrum is shown in Fig. 8.5.1.

EXAMPLE 8.5.1

A carrier wave of frequency 10 MHz and peak value 10 V is amplitude modulated by a 5-kHz sine wave of amplitude 6 V. Determine the modulation index and draw the amplitude spectrum.

SOLUTION $m = \frac{6}{10} = 0.6$

The side frequencies are $10 \pm 0.005 = 10.005$ and 9.995 MHz. The amplitude of each side frequency is $0.6 \times 10/2 = 3$ V. The spectrum is shown in Fig. 8.5.1(b).

This result is of more than mathematical interest. The three components are present physically, and, for example, they can be separated out by filtering. Use is made of this in single-sideband transmission, discussed in Chapter 9.

The modulated wave could be considered to be generated by three separate generators, as shown in Fig. 8.5.2(a). Although such an arrangement would be difficult to set up in practice (because of the difficulty

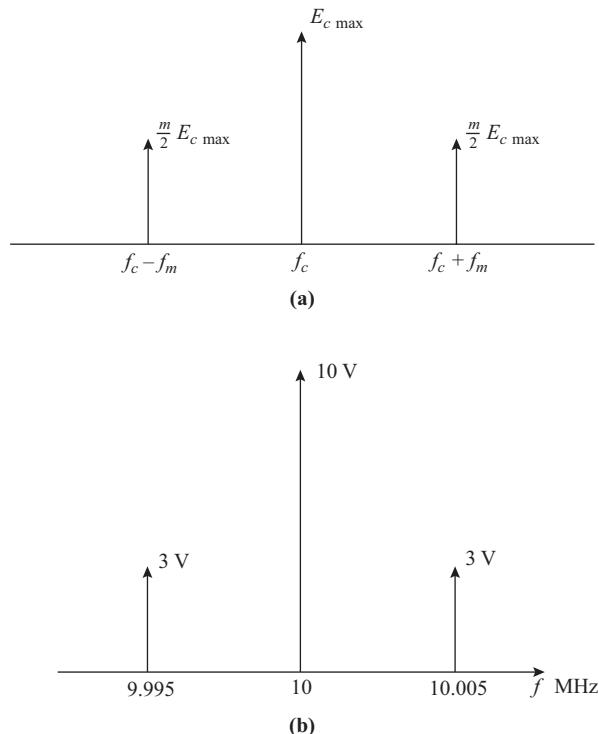


Figure 8.5.1 (a) Amplitude spectrum for a sinusoidally amplitude modulated wave. (b) The amplitude spectrum for a 10-MHz carrier of amplitude 10 V, sinusoidally modulated by a 5-kHz sine wave of amplitude 6 V (Example 8.5.1).

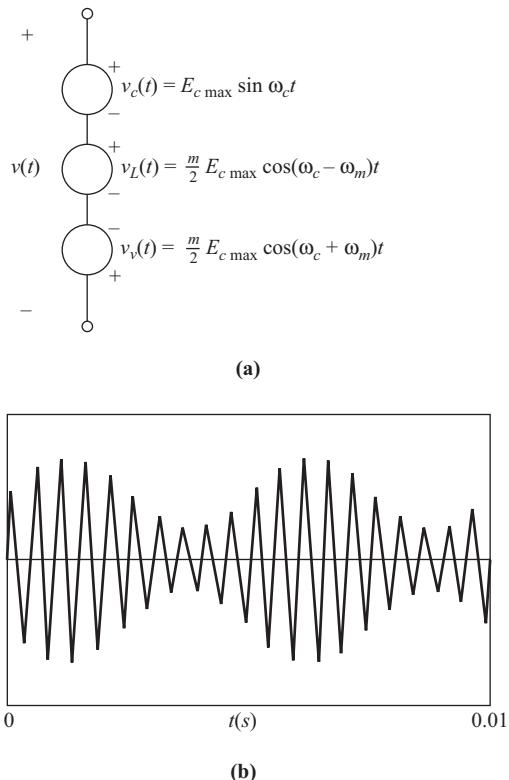


Figure 8.5.2 (a) Generator representation of a sinusoidally amplitude modulated wave. (b) Result of a computer simulation of the three-generator arrangement for $E_{c \max} = 5$ V, $m = 0.5$, $f_c = 100$ kHz, and $f_m = 10$ kHz.

in maintaining exactly the right frequencies), it is easily simulated on a computer, and the result of such a simulation obtained using Mathcad is shown in Fig. 8.5.2(b).

8.6 Average Power for Sinusoidal AM

Figure 8.5.1(a) shows that the sinusoidally modulated wave can be represented by three sinusoidal sources connected in series. A general result of ac circuit theory is that the average power delivered to a load R by series-connected sinusoidal sources of different frequencies is the sum of the average powers from each source. The average power in a sine (or cosine) voltage wave of peak value E_{\max} developed across a resistor R is $P = E_{\max}^2/2R$. Applying these results to the spectrum components of the sinusoidally modulated wave gives, for the average carrier power,

$$P_C = \frac{E_{c \max}^2}{2R} \quad (8.6.1)$$

and for each side frequency

$$\begin{aligned} P_{SF} &= \frac{(mE_{c \max}/2)^2}{2R} \\ &= \frac{m^2}{4} P_C \end{aligned} \quad (8.6.2)$$

Hence the total average power is

$$\begin{aligned} P_T &= P_C + 2 \times P_{SF} \\ &= P_C \left(1 + \frac{m^2}{2} \right) \end{aligned} \quad (8.6.3)$$

At 100% modulation ($m = 1$), the power in any one side frequency component is $P_{SF} = P_C/4$ and the total power is $P_T = 1.5P_C$. The ratio of power in any one side frequency to the total power transmitted is therefore 1/6. The significance of this result is that all the original modulating information is contained in the one side frequency, and therefore a considerable savings in power can be achieved by transmitting just the side frequency rather than the total modulated wave. In practice, the modulating signal generally contains a band of frequencies that results in *sidebands* rather than single side frequencies, but, again, single-sideband (SSB) transmission results in more efficient use of available power and spectrum space. Single-sideband transmission is considered in Chapter 9.

8.7 Effective Voltage and Current for Sinusoidal AM

The effective or rms voltage E of the modulated wave is defined by the equation

$$\frac{E^2}{R} = P_T \quad (8.7.1)$$

Likewise, the effective or rms voltage E_c of the carrier component is defined by

$$\frac{E_c^2}{R} = P_C \quad (8.7.2)$$

It follows from Eq. (8.6.3) that

$$\begin{aligned} \frac{E^2}{R} &= P_C \left(1 + \frac{m^2}{2} \right) \\ &= \frac{E_c^2}{R} \left(1 + \frac{m^2}{2} \right) \end{aligned} \quad (8.7.3)$$

from which

$$E = E_c \sqrt{1 + \frac{m^2}{2}} \quad (8.7.4)$$

A similar argument applied to currents yields

$$I = I_c \sqrt{1 + \frac{m^2}{2}} \quad (8.7.5)$$

where I is the rms current of the modulated wave and I_c the rms current of the unmodulated carrier. The current equation provides one method of monitoring modulation index, by measuring the antenna current with and without modulation applied. From Eq. (8.7.5),

$$m = \sqrt{2\left[\left(\frac{I}{I_c}\right)^2 - 1\right]} \quad (8.7.6)$$

The method is not as sensitive or useful as the trapezoidal method described earlier, but it provides a convenient way of monitoring modulation where an ammeter can be inserted in series with the antenna, for example. A true rms reading ammeter must be used, and care must be taken to avoid current overload because such instruments are easily damaged by overload.

EXAMPLE 8.7.1

The rms antenna current of an AM radio transmitter is 10 A when unmodulated and 12 A when sinusoidally modulated. Calculate the modulation index.

SOLUTION $m = \sqrt{2\left[\left(\frac{12}{10}\right)^2 - 1\right]} = 0.94$

8.8 Nonsinusoidal Modulation

Nonsinusoidal modulation has already been illustrated in Fig. 8.2.1 and the modulation index determined as shown in Section 8.3. Sometimes the *modulation depth*, rather than modulation index, is used as a measure of modulation. The modulation depth is the ratio of the downward modulation peak to the peak carrier level, usually expressed as a percentage. As shown in Fig. 8.3.1, overmodulation occurs if the modulation depth exceeds 100%, irrespective of the modulating waveshape. (For sinusoidal modulation, modulation depth is equal to the modulation index. Signal generators generally employ sinusoidal modulation but have meters calibrated in modulation depth.)

Nonsinusoidal modulation produces upper and lower *sidebands*, corresponding to the upper and lower side frequencies produced with sinusoidal modulation. Suppose, for example, that the modulating signal has a line spectrum as shown in Chapter 2 so that it can be represented by

$$e_m(t) = E_{1\max} \cos 2\pi f_1 t + E_{2\max} \cos 2\pi f_2 t + E_{3\max} \cos 2\pi f_3 t + \dots \quad (8.8.1)$$

As before, the AM wave is

$$e(t) = [E_c \max + e_m(t)] \cos 2\pi f_c t \quad (8.8.2)$$

If in general the i th component is denoted by subscript i , then individual modulation indexes may be defined as $m_i = E_i \max / E_c \max$ and the trigonometric expansion for Eq. (8.8.2) yields a spectrum with side frequencies at $f_c \pm f_i$ and amplitudes $m_i E_c \max / 2$. This is sketched in Fig. 8.8.1(a). Thus, taken together, the side frequencies form sidebands either side of the carrier component. Again, the practicalities of AM demand that the carrier frequency be much greater than the highest frequency in the modulating wave, so the sidebands are bandlimited about the carrier frequency as shown.

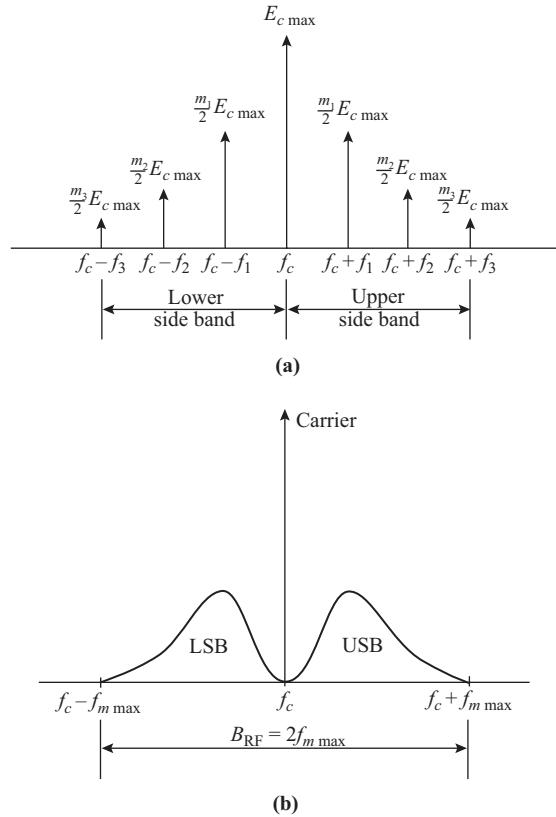


Figure 8.8.1 (a) Amplitude spectrum resulting from line spectra modulation. (b) Amplitude spectrum for a power density modulating spectra.

The total average power can be obtained by adding the average power for each component (just as was done for single-tone modulation), which results in

$$P_T = P_c \left(1 + \frac{m_1^2}{2} + \frac{m_2^2}{2} + \frac{m_3^2}{2} + \dots \right) \quad (8.8.3)$$

Hence an effective modulation index can be defined in this case as

$$m_{\text{eff}} = \sqrt{m_1^2 + m_2^2 + m_3^2 + \dots} \quad (8.8.4)$$

It follows that the effective voltage and current in this case are

$$E = E_c \sqrt{1 + \frac{m_{\text{eff}}^2}{2}} \quad (8.8.5)$$

$$I = I_c \sqrt{1 + \frac{m_{\text{eff}}^2}{2}} \quad (8.8.6)$$

When the modulating signal is a random power signal such as speech or music, then the concept of power spectral density must be used, as shown in Chapter 2. Thus, if the power spectral density curve is as sketched in Fig. 2.17.1, when used to amplitude modulate the carrier, double sidebands are generated as shown in Fig. 8.8.1(b). Again it is assumed that the modulating signal is bandlimited such that the highest frequency in its spectrum is much less than the carrier frequency.

It will be seen therefore that standard AM produces upper and lower sidebands about the carrier, and hence the RF bandwidth required is double that for the modulating waveform. From Fig. 8.8.1,

$$\begin{aligned} B_{RF} &= (f_c + f_{m \max}) - (f_c - f_{m \max}) \\ &= 2f_{m \max} \end{aligned} \quad (8.8.7)$$

where $f_{m \max}$ is the highest frequency in the modulating spectrum. As with the sinusoidal modulation, either sideband contains all the modulating signal information, and therefore considerable savings in power and bandwidth can be achieved by transmitting only one sideband. Single sideband (SSB) transmission is the subject of Chapter 9.

Previously, overmodulation was shown to result in distortion of the modulation envelope (see Fig. 8.3.1). Such distortion also results in sideband frequencies being generated that lie outside the normal sidebands and that may overlap with the sidebands of adjacent channels. This form of interference is referred to as *sideband splatter* and must be avoided. Hence the necessity of ensuring that the modulation index does not exceed unity.

8.9 Double-sideband Suppressed Carrier (DSBSC) Modulation

Certain types of amplitude modulators make use of a multiplying action in which the modulating signal multiplies the carrier wave. The balanced mixer described in Section 5.10 is one such circuit, and in fact these are generally classified as *balanced modulators*. As shown by Eq. (5.10.17), the output current contains a product term of the two input voltages. When used as a modulator, the oscillator input becomes the carrier input, and the signal input becomes the modulating signal input. The output voltage can then be written as

$$e(t) = ke_m(t) \cos 2\pi f_c t \quad (8.9.1)$$

where k is a constant of the multiplier circuit. The expression for standard AM with carrier is

$$\begin{aligned} e(t) &= (E_{c \max} + e_m(t)) \cos 2\pi f_c t \\ &= E_{c \max} \cos 2\pi f_c t + e_m(t) \cos 2\pi f_c t \end{aligned} \quad (8.9.2)$$

The major difference between the multiplier result and this is that carrier term $E_{c \max} \cos 2\pi f_c t$ is absent from the multiplier result. This means that the carrier component will be absent from the spectra, which otherwise will be the same as for AM with carrier. The constant multiplier k can be regarded simply as a scaling factor, and it will not materially affect the results. This type of amplitude modulation is therefore known as double-sideband suppressed carrier (DSBSC). The spectra are sketched in Fig. 8.9.1 for sinusoidal modulation and for the general case.

The absence of a carrier component means that DSBSC utilizes the transmitted power more efficiently than standard AM; however, it still requires twice the bandwidth compared to single sideband (SSB). It should be noted that, although the bandwidth is double that required for SSB, the received power is also double that obtained with SSB, and therefore the signal-to-noise ratio is the same. However, conserving bandwidth is an important aim in communications systems, and usually DSBSC represents one step in generating SSB, as described in Chapter 9.

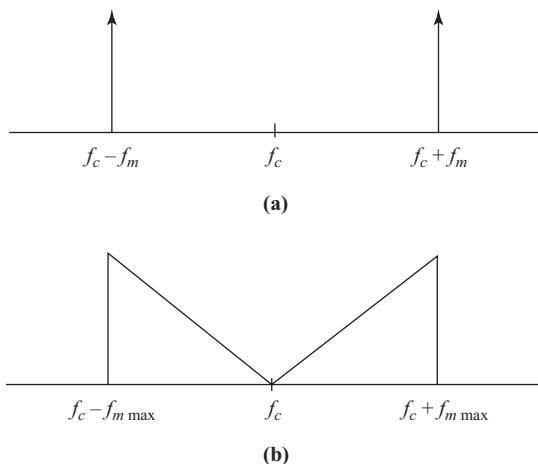


Figure 8.9.1 DSBSC spectrum for (a) sinusoidal modulation and (b) the general case.

8.10 Amplitude Modulator Circuits

Amplitude modulation may be achieved in a number of ways, the most common method being to use the modulating signal to vary the otherwise steady voltage on the output electrode of an amplifier. Vacuum tubes are used for very high power outputs (in the kilowatt and higher ranges), and transistors are used for lower powers. (Pentode vacuum tubes and transistors have similarly shaped output characteristics, although they operate at greatly different voltage and current levels.) The basic circuit for a BJT modulator is shown in Fig. 8.10.1. The transistor is normally operated in the class C mode in which it is biased well beyond cutoff. The carrier input to the base must be sufficient to drive the transistor into conduction over part of the RF cycle, during which the collector current flows in the form of pulses. These pulses are periodic at the carrier frequency and can therefore be analyzed into a trigonometric Fourier series, as shown in Chapter 2. The tuned circuit in the collector is tuned to resonate at the fundamental component, and thus, to a close approximation, the RF voltage at the collector is sinusoidal.

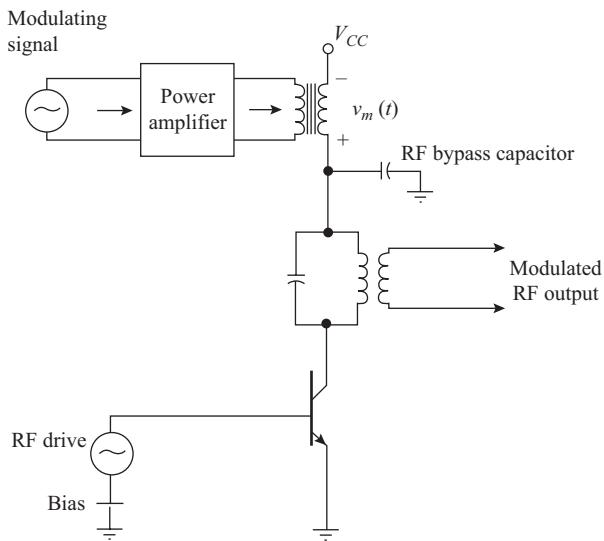


Figure 8.10.1 Basic circuit for a BJT collector modulator.

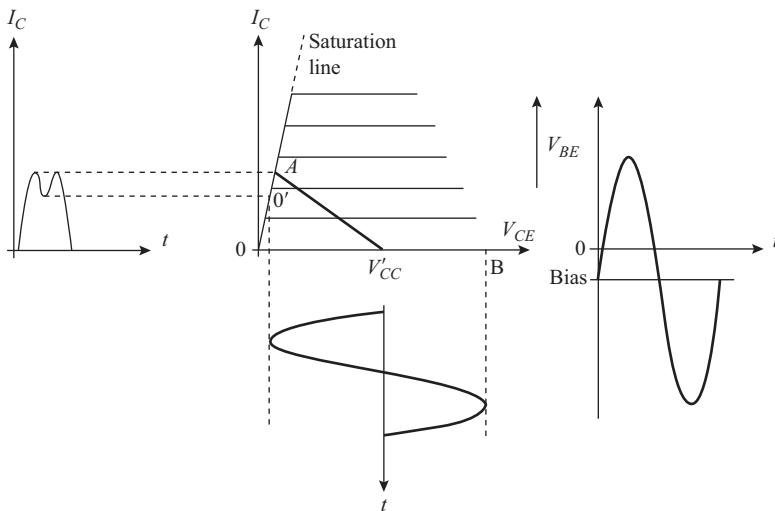


Figure 8.10.2 The BJT output characteristics, also showing one RF cycle of operation.

When a modulating voltage is applied, the steady collector voltage changes to a slowly varying voltage (slow compared to the RF cycle) given by $V'_CC = VCC + v_m(t)$. The modulating voltage $v_m(t)$ is applied in series with VCC through the low-frequency transformer. The RF bypass capacitor provides a low-impedance path for the RF to ground so that negligible RF voltage is developed across the LF transformer secondary.

For the modulating voltage to have a controlling effect on the peak value of the current pulse, the RF voltage on the collector must swing the transistor into its saturation region over part of each RF cycle. The saturation line is shown on the output characteristics of Fig. 8.10.2. The region to the right of the saturation line is known as the forward active region, and it will be seen that the collector current is virtually independent of collector voltage in this region. Thus the RF load line must include the section $O'A$, the complete RF load line for a fixed value of V'_CC being shown as $O'AV'C'CB$.

One cycle of the base voltage is shown to the right of the output characteristics, and the corresponding cycle of collector voltage is under the V_{CE} axis. To the left is shown the current pulse that occurs during the cycle.

Figure 8.10.3(a) shows the output characteristics with three different load lines corresponding to three different values of modulating voltage. The collector voltage is shown in Fig. 8.10.3(b) and the collector current pulses in Fig. 8.10.3(c). When the class C modulator is properly adjusted, the RF voltage from collector to ground has a peak-to-peak value almost equal to $2V'_CC$, as shown in Fig. 8.10.3.

The modulated output is obtained through mutual inductive coupling, as shown in the circuit diagram. The coupling prevents the “steady” voltage from being transferred to the output so that the RF varies about a mean value of zero. This is shown in Fig. 8.10.3(d). It will be seen that the peak-to-peak voltages are transformed in a linear fashion so that, denoting voltages on the collector side of the output transformer by V and on the output side by E , $E_{c \max} = KV_{c \max}$, $E_{\max pp} = KV_{\max pp}$, and $E_{\min pp} = KV_{\min pp}$, where K is a constant. It follows therefore that the modulation index is given by

$$\begin{aligned}
 m &= \frac{E_{\max} - E_{\min}}{E_{\max} + E_{\min}} \\
 &= \frac{E_{\max pp} - E_{\min pp}}{E_{\max pp} + E_{\min pp}} \\
 &= \frac{V_{\max pp} - V_{\min pp}}{V_{\max pp} + V_{\min pp}}
 \end{aligned} \tag{8.10.1}$$

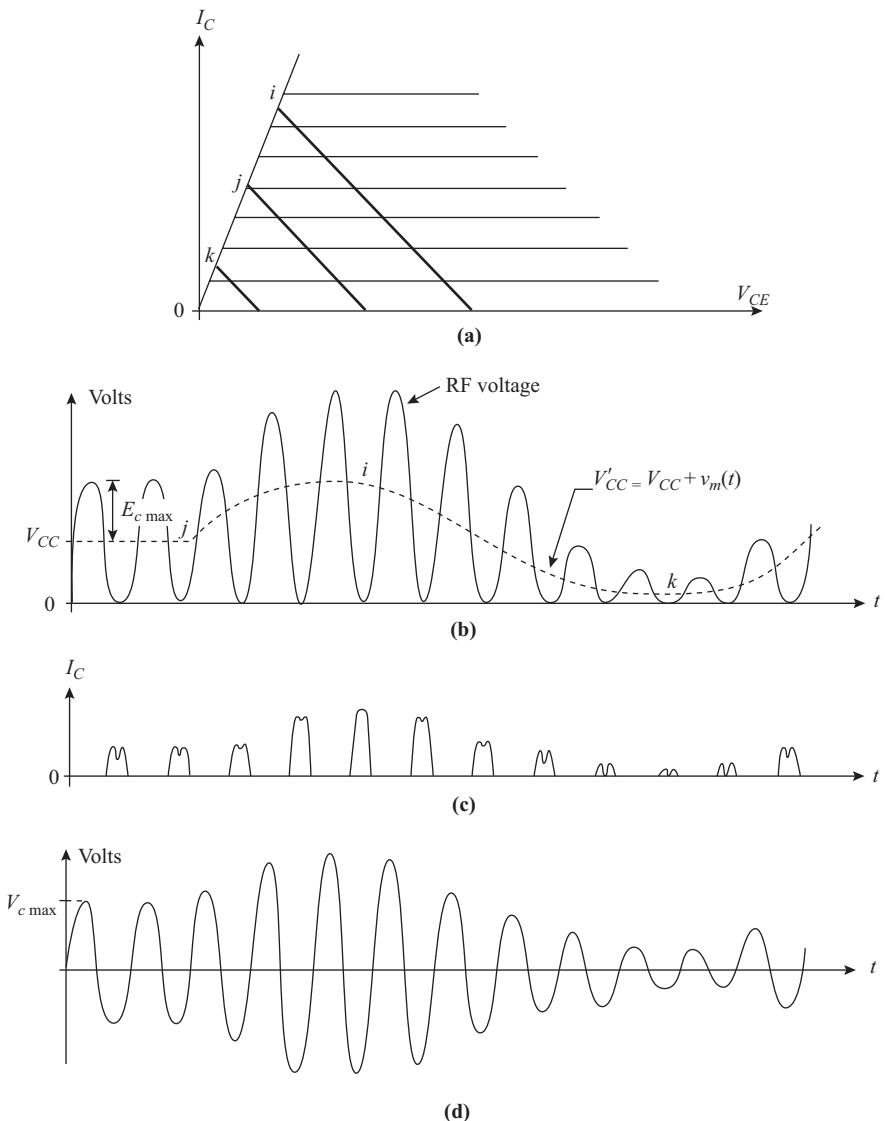


Figure 8.10.3 (a) Output characteristics showing three load lines. (b) Collector voltage. (c) Current pulses. (d) Modulated output voltage.

It is left as an exercise for the student to show that for sinusoidal modulation in particular

$$m = \frac{V_{m \text{ max}}}{V_{CC}} \quad (8.10.2)$$

where $V_{m \text{ max}}$ is the peak value of $v_m(t)$.

The class C amplifier can be considered as a power converter. When no modulation is applied, it converts the dc input power to the collector to the unmodulated RF output power, or $P_C = \eta P_{CC}$, where P_C is the unmodulated carrier power, P_{CC} the dc input power to the collector, and η is the conversion efficiency

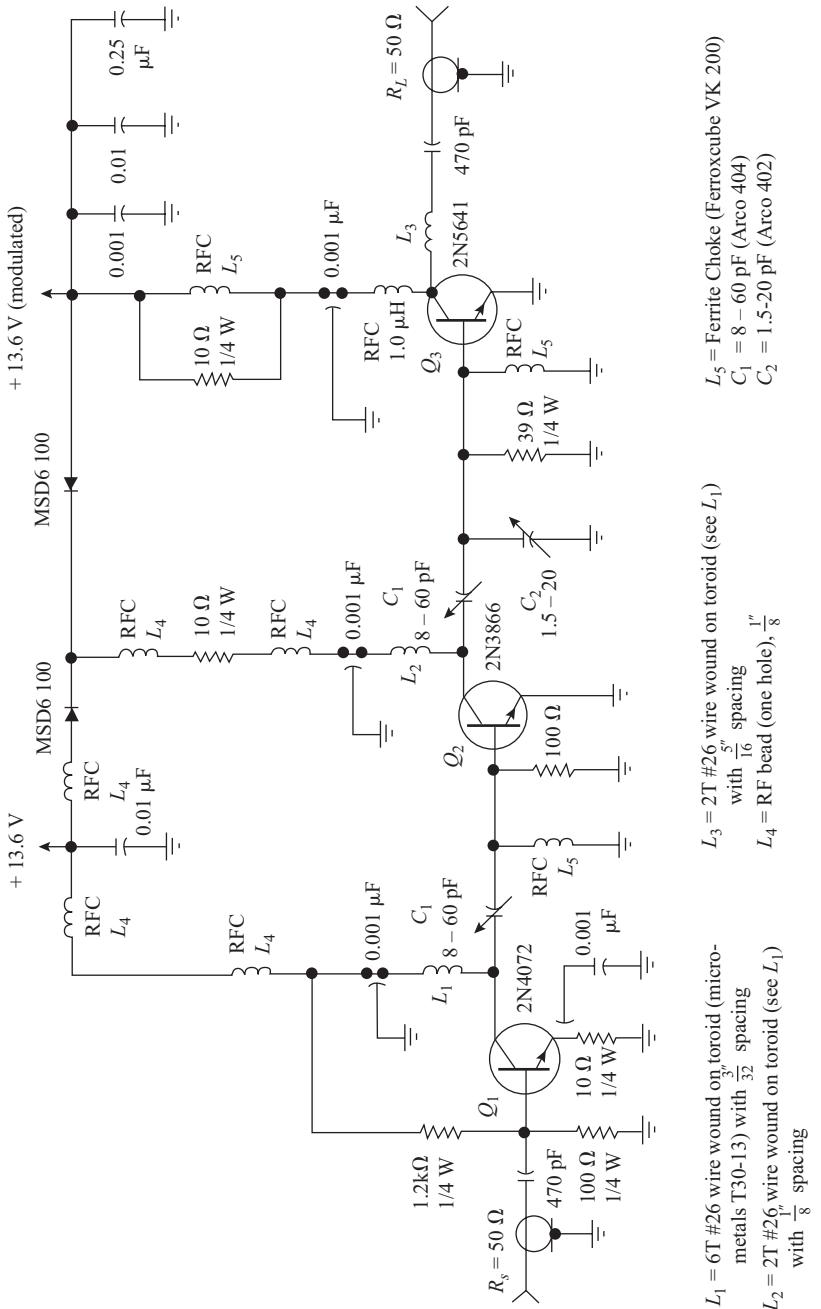


Figure 8.10.4 Transistorized collector modulator circuit. (Courtesy of Motorola Inc, Applicable Note AN507.)

(conversion efficiencies can be quite high, typically 70% to 75%). When modulation is applied, the “dc input” becomes a slowly varying input and assuming that the conversion efficiency remains the same, the additional power supplied by the modulator goes into the creation of the sidebands or $P_{SB} = \eta P_{\text{mod}}$, where P_{SB} is the total average sideband power and P_{mod} the average power supplied by the modulator. Looking at this in another way, the modulator has to supply power equal to

$$P_{\text{mod}} = \frac{P_{\text{SF}}}{\eta} \quad (8.10.3)$$

Thus, for large power transmitters, the modulator must be capable of delivering large amounts of power. For example, a 50-kW (unmodulated) transmitter would generate 25 kW of side-frequency power at 100% sinusoidal modulation (see Section 8.6). Assuming a class C modulator efficiency of 75%, the modulator would have to provide 33.3 kW of audio power (assuming an audio broadcast transmission). To avoid the need for such large modulator powers, modulation would take place at a lower power level, and class B power amplification used to increase the modulated signal up to the required level, as described in Section 8.12.

Another disadvantage with AM is the high voltage rating required of the modulator stage. As shown above, the peak-to-peak collector voltage is $2(V_{CC} + v_m(t))$. Again, with 100% sinusoidal modulation the voltage can reach a peak value of $2(V_{CC} + V_{m \text{ max}}) = 4V_{CC}$ and the rating of the transistor must take this into account.

If the RF input to the base is sufficiently high to drive the modulator to full RF output at the peaks of modulation, overdrive occurs at the lower levels of modulation, which can result in the production of excessive harmonics. One way of overcoming this is to apply partial modulation to the driver stage to take effect for positive peaks of modulating signal. Figure 8.10.4 shows one such circuit.

In this circuit, the transistor Q_3 stage is fully modulated, while diodes MSD6100 (dual-package diodes) allow stage Q_2 to be modulated on the upward modulation swing; when the modulating voltage swings below 13.6 V, the diode connected to the modulated supply ceases to conduct, thus cutting Q_2 off from the modulation while the other diode conducts, connecting Q_2 to the unmodulated 13.6-V supply. This means that the RF drive to Q_3 is increased at the same time as the collector voltage increases because of modulation, thus increasing the drive to the Q_3 output stage and preventing clipping.

The tuned output stage for Q_3 is the series circuit of 470 pF and L_3 , while the various radio-frequency chokes and capacitors in the collector line are for the purposes of filtering. The modulating amplifier is not shown in the figure.

8.11 Amplitude Demodulator Circuits

At the receiver, a circuit must be provided that recovers the information signal from the modulated carrier. The most common circuit in use is the *diode envelope detector*, which produces an output voltage proportional to the envelope, which is the modulating or information signal. The basic circuit is shown in Fig. 8.11.1(a). The diode acts as a rectifier and can be considered an ON switch when the input voltage is positive, allowing the capacitor C to charge up to the peak of the RF input. During the negative half of the RF cycle, the diode is off, but the capacitor holds the positive charge previously received, so the output voltage remains at the peak positive value of RF. There will, in fact, be some discharge of C , producing an RF ripple on the output waveform, which must be filtered out.

As the input voltage rises with the modulating cycle, the capacitor voltage has no difficulty in following this, but during the downward swing in modulation the capacitor may not discharge fast enough unless an additional discharge path is provided by the resistor R . The time constant of the CR load has to be short enough to allow the output voltage to follow the modulating cycle and yet long enough to maintain a relatively high output voltage. The constraints on the time constant are determined more precisely in the next section.

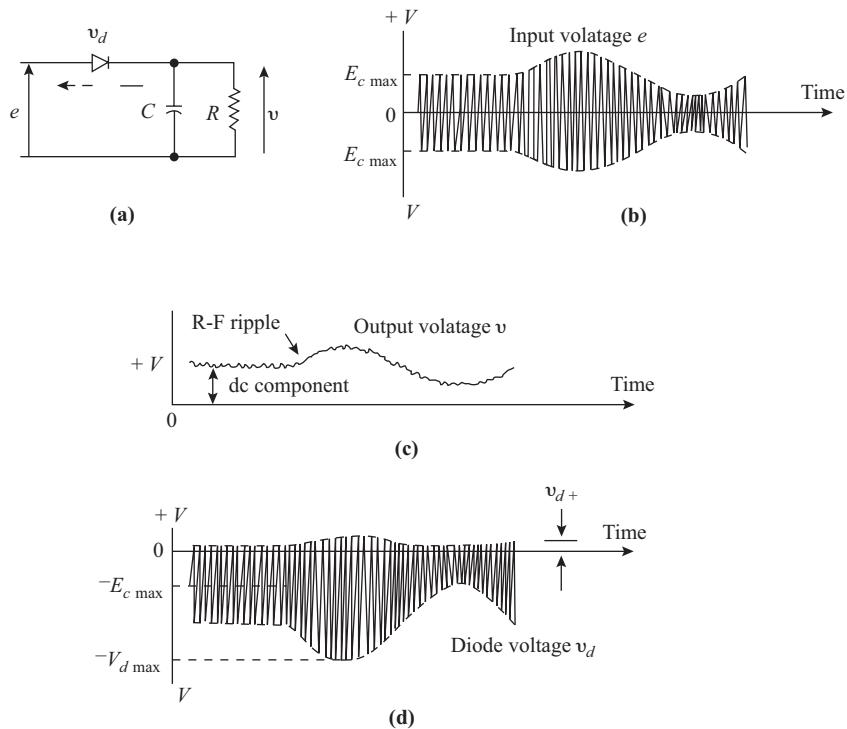


Figure 8.11.1 (a) Basic diode envelope detector. (b) Voltage input waveform. (c) Voltage output waveform. (d) Voltage across the diode.

Applying Kirchhoff's voltage law to the circuit, the diode voltage v_d is found to be

$$v_d = e - v \quad (8.11.1)$$

where e is the input voltage and v the output voltage. Figure 8.11.1(b) shows e , and Fig. 8.11.1(c) shows v , both for sinusoidal modulation. By graphically subtracting v from e the graph of v_d in Fig. 8.11.1(d) is obtained. It is interesting to see that v_d is positive only for very short periods, as indicated by the peaks v_d^+ , and it is during these peaks that the capacitor is charged to make up for discharge losses. The peak voltage across the diode can rise to $4E_{c \max}$ at 100% sinusoidal modulation, a condition that should be compared with the condition at the collector of the modulator described in the previous section.

Diagonal Peak Clipping. This is a form of distortion that occurs when the time constant of the RC load is too long, thus preventing the output voltage from following the modulation envelope. The output voltage is labeled V_{AV} in Fig. 8.11.2(a), to show that it is the average voltage that follows the modulation envelope (that is, the RF ripple is averaged out). The curve of V_{AV} for sinusoidal modulation is shown in Fig. 8.11.2(b). At some time t_A the modulation envelope starts to decrease more rapidly than the capacitor discharges. The output voltage then follows the discharge curve of the RC network until time t_B , when it meets up with the modulation envelope as it once again increases.

For sinusoidal modulation the condition necessary for the avoidance of diagonal peak clipping is found as follows. Because of the capacitive nature of the RC load, the current leads the voltage as shown in

Fig. 8.11.2(c). The average current consists of two components, a dc component I_{DC} and an ac component that has a peak value I_p , as shown in Fig. 8.11.2(c). The dc component of voltage is approximately equal to the maximum unmodulated carrier voltage or $V_{DC} \approx E_{c\ max}$ and the direct current is $I_{DC} = V_{DC}/R$. The peak value of the average output voltage is $V_p \approx m_{ec\ max}$, and the corresponding value of the peak current is $I_p = V_p/Z_p$, where Z_p is the impedance of the RC load at the modulating frequency.

If the envelope falls faster than the capacitor discharges, the diode ceases to conduct (since the capacitor voltage biases it off), and the current I_{AV} supplied by the diode goes to zero. This is shown in Fig. 8.11.2(c). During the period the current is zero, the load voltage follows the discharge law of the RC network, resulting in the diagonally clipped peak shown in Fig. 8.11.2(b). From Fig. 8.11.2(c), it is seen that for the avoidance of diagonal peak clipping the direct current has to be greater than the peak current, or $I_{DC} \geq |I_p|$. Hence

$$\frac{V_{DC}}{R} \geq \frac{mV_{DC}}{|Z_p|}$$

$$\therefore m \leq \frac{|Z_p|}{R} \quad (8.11.2)$$

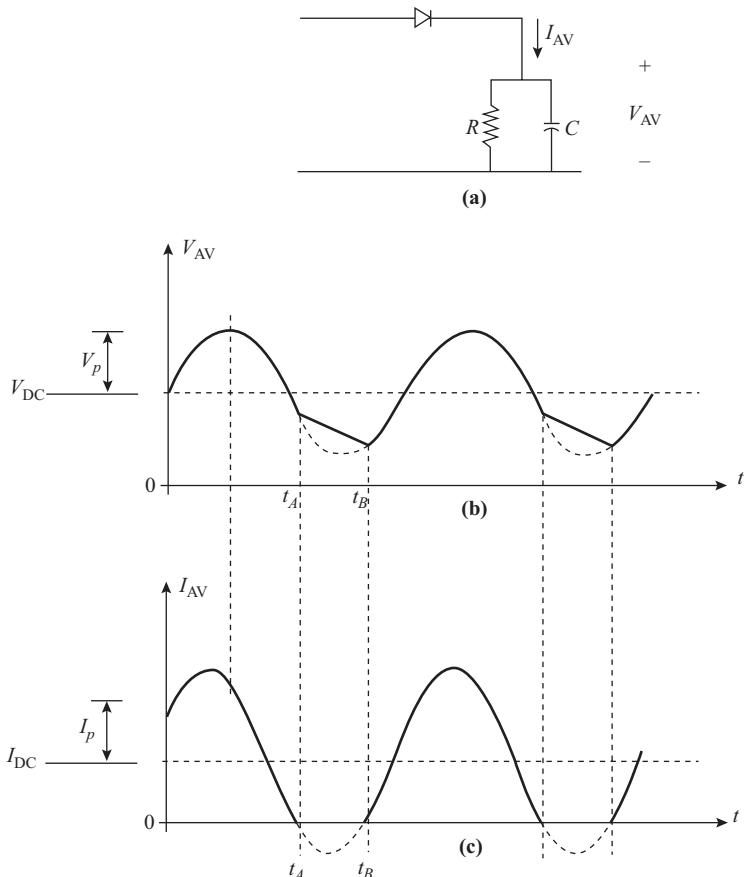


Figure 8.11.2 (a) Diode circuit supplying an average current I_{AV} to the RC load. (b) Voltage waveform, illustrating diagonal peak clipping. (c) Current waveform.

EXAMPLE 8.11.1

The RC load for a diode detector consists of a 1000-pF capacitor in parallel with a 10-k Ω resistor. Calculate the maximum modulation depth that can be handled for sinusoidal modulation at a frequency of 10 kHz if diagonal peak clipping is to be avoided.

SOLUTION The admittance of the RC load is

$$\begin{aligned} Y_p &= \frac{1}{R} + j2\pi f_m C \\ &= 10^{-4} + j6.2810^{-5} \text{ S} \end{aligned}$$

Hence

$$|Z_p| = \frac{1}{|Y_p|} = 8467 \Omega$$

The maximum modulation index that can be handled without distortion is therefore

$$m = \frac{|Z_p|}{R} \cong 0.85$$

Since it is always the case that $R > |Z_p|$, the diode demodulator must introduce diagonal peak clipping as the modulation index approaches unity. In practice, however, the modulation index at the transmitter is prevented from approaching unity to avoid overmodulation, with its consequent distortion and sideband splatter.

Negative Peak Clipping. This is similar in appearance to diagonal peak clipping, but results from the loading effect of the network R_1C_1 following the RC load [Fig. 8.11.3(a)]. Capacitor C_1 is a dc blocking capacitor, and resistor R_1 represents the input resistance of the following stage.

Considering the normal situation where the reactance of C_1 is very small, and that of C is very large at the modulating frequency (assumed sinusoidal), the ac impedance is simply R in parallel with R_1 or $|Z_{pl}| = R_p = RR_1/(R + R_1)$. The modulation index must now meet the condition

$$m \leq \frac{R_p}{R} \quad (8.11.3)$$

In this situation, the current I_{AV} is in phase with V_{AV} , and over the period when I_{AV} is zero [Fig. 8.11.3(b)], the C_1 capacitor voltage remains approximately constant at V_{DC} , which in turn develops a voltage equal to $V_{MIN} = V_{DC} R / (R + R_1)$ across R . It is this voltage that keeps the diode biased off. The voltages across R and R_1 are shown in Fig. 8.11.3(c). The shape of the output voltage curve shows why the term *negative peak clipping* is used.

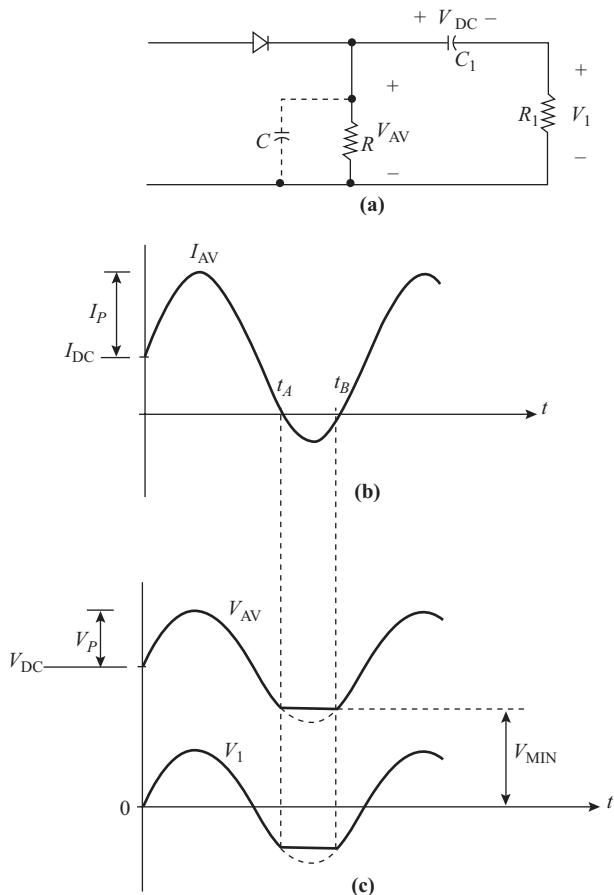


Figure 8.11.3 (a) Diode circuit including a dc blocking capacitor C_1 and the input resistor R_1 of the following stage. (b) Average diode current. (c) The voltages across R and R_1 .

8.12 Amplitude-modulated Transmitters

Figure 8.12.1(a) shows the block diagram of a typical AM transmitter. The carrier source is a crystal-controlled oscillator at the carrier frequency or a submultiple of it. This is followed by a tuned buffer amplifier and a tuned driver, and if necessary frequency multiplication is provided in one or more of these stages.

The modulator circuit used is generally a class C power amplifier that is collector modulated as described in Section 8.10. The audio signal is amplified by a chain of low-level audio amplifiers and a power amplifier. Since this amplifier is controlling the power being delivered to the final RF amplifier, it must have a power driving capability that is one-half the maximum power the collector supply must deliver to the RF amplifier under 100% modulation conditions. A transformer-coupled class B push-pull amplifier is usually used for this purpose.

Low-power transmitters with output powers up to 1 kW or so may be transistorized, but as a rule the higher-power transmitters use vacuum tubes in the final amplifier stage, even though the low-level stages may be transistorized. In some cases where the reliability and high overall efficiency of the transistor are

mandatory, higher powers can be obtained by using several lower-power transistorized amplifiers in parallel. The system is complicated, and usually the vacuum-tube version will do the same job at lower capital cost.

Sometimes the modulation function is done in one of the low-level stages. This allows low-power modulation and audio amplifiers, but it complicates the RF final amplifier. Class C amplifiers cannot be used to amplify an already modulated (AM) carrier, because the transfer function of the class C amplifier is not linear. The result of using a class C amplifier would be an unacceptable distortion of the modulation envelope. A linear power amplifier, such as the push-pull class B amplifier, must be used to overcome this problem [Fig. 8.12.1(b)].

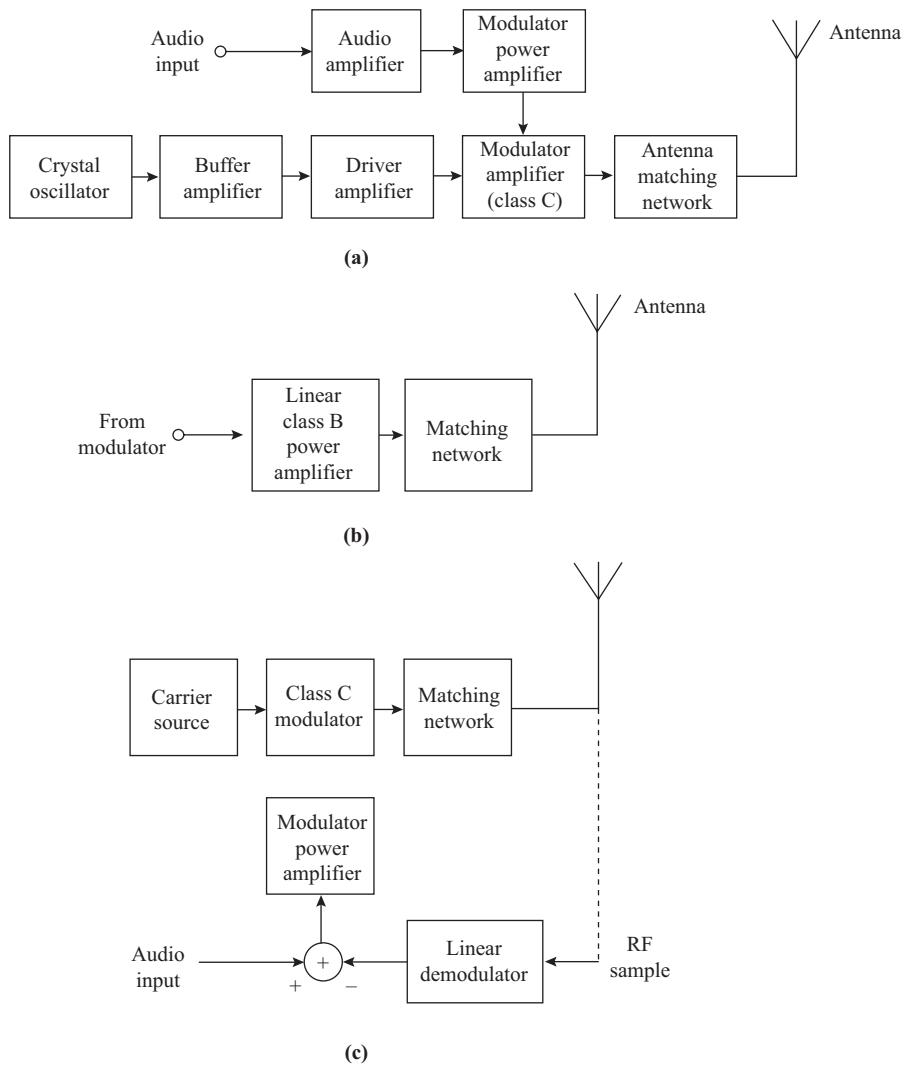


Figure 8.12.1 Amplitude modulated transmitters: (a) transmitter with a modulated class C final power amplifier; (b) linear class B push-pull power amplifier used when modulation takes place in a low-level stage; (c) negative feedback applied to linearize a class C modulator.

Unfortunately, the efficiency of this type of amplifier is lower than that of the comparable class C amplifier, resulting in more costly equipment. Larger tubes or transistors must be used that are capable of dissipating the additional heat generated.

The output of the final amplifier is passed through an impedance-matching network that includes the tank circuit of the final amplifier. The Q of this circuit must be low enough so that all the sidebands of the signal are passed without amplitude/frequency distortion, but at the same time must present an appreciable attenuation at the second harmonic of the carrier frequency. The bandwidth required in most cases is a standard 3 dB at ± 5 kHz around the carrier. For amplitude-modulation broadcast transmitters, this response may be broadened so that the sidebands will be down less than 1 dB at 5 kHz where music programs are being broadcast and very low distortion levels are desired, or special sharp-cutoff filters may be used. Because of the high power levels present in the output, this is not usually an attractive solution.

Negative feedback is quite often used to reduce distortion in a class C modulator system. The feedback is accomplished in the manner shown in Fig. 8.12.1(c), where a sample of the RF signal sent to the antenna is extracted and demodulated to produce the feedback signal. The demodulator is designed to be as linear in its response as possible and to feed back an audio signal that is proportional to the modulation envelope. The negative feedback loop functions to reduce the distortion in the modulation.

AM Broadcast Transmitters

Most domestic AM broadcast services use the medium-wave band from 550 to 1600 kHz. International AM broadcasts take place in several of the HF bands scattered from 1600 kHz up to about 15 MHz. The mode of transmission in all cases is double-sideband full carrier, with an audio baseband range of 5 kHz. Station frequency assignments are spaced at 10 kHz intervals, and power outputs range from a few hundred watts for small local stations to as much as 100 kW in the MW band and even higher for international HF transmitters.

A main requirement of an AM broadcast transmitter is to produce, within the limits of the 5-kHz audio bandwidth available, the highest possible fidelity. The modulator circuits in the transmitter must produce a linear modulation function, and every trick available is used to accomplish this. A typical AM broadcast transmitter is shown in Fig. 8.12.2. The crystal oscillator is temperature-controlled to provide frequency stability. It is followed by a buffer amplifier and then by tuned class C amplifiers that provide the necessary power gain to drive the final power amplifier. For high power output, vacuum tubes would be used as described next. The modulator system is the *triple equilibrium* system, in which the main part of the modulation is performed by plate-modulating the final class C power amplifier. Secondary modulation of both the final grid and the plate of the driver stage is also included to compensate for bias shift in the final amplifier that results from the nonlinear characteristic of the amplifier.

The final power amplifier is a push-pull parallel stage in which each side of the push-pull stage is composed of several vacuum tubes operating in parallel, to obtain the power required. A further advantage of this system is that, if one or more of the tubes in the system should fail, the remaining tubes will provide partial output until repairs can be made, thus making a more secure system. Power dissipation in these final tubes can be as high as 50 kW, in addition to several kilowatts of heater power. Water cooling systems are used to dissipate the large quantities of heat produced.

The modulator amplifier is an audio-frequency push-pull parallel amplifier, which is transformer-coupled to the modulator. The audio preamplifier stage includes a difference amplifier and an envelope detector that demodulates a sample of the transmitter output and uses the signal to provide negative feedback. This feedback further linearizes the modulation characteristic of the system.

Antenna systems for AM transmitters are large and usually must be located at some point remote from the studio operations. All the studio signal operations are performed at relatively low levels and transmitted to the main transmitter location, either over telephone wire lines or a radio link such as a microwave system.

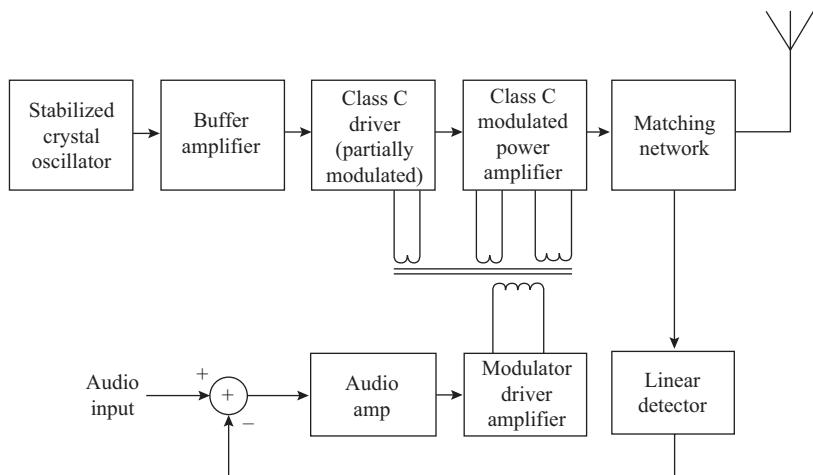


Figure 8.12.2 AM broadcast transmitter.

Often the transmitters are unattended and are remotely controlled from the studio location. Service personnel make periodic visits to do routine maintenance.

8.13 AM Receivers

The general principles of the superheterodyne receiver are described in Chapter 7, and specific operating details of the AM envelope detector are discussed in Section 8.11. Most receivers in use today are assembled from discrete components, although there is a trend toward the use of integrated circuits for subsections in the receiver. Therefore, in this section, a very commonly encountered transistorized receiver will be described, followed by the description of two integrated circuit-type receivers.

Discrete Component AM Receiver

The circuit for a standard broadcast receiver using discrete components is shown in Fig. 8.13.1. This is a superheterodyne receiver, transistor Q_1 functioning as both a mixer and an oscillator in what is known as an autodyne mixer. The oscillator feedback is through mutual inductive coupling from collector to emitter, the base of Q_1 being effectively grounded at the oscillator frequency.

The AM signal is coupled into the base of Q_1 via coil L_1 . Thus it is seen that Q_1 operates in grounded base mode for the oscillator while simultaneously operating in grounded emitter mode for the signal input.

Tuned IF transformer T_1 couples the IF output from Q_1 to the first IF amplifier Q_2 . The output from Q_2 is also tuned-transformer-coupled through T_2 to the second IF amplifier Q_3 . The output from Q_3 , at IF, is tuned-transformer-coupled to the envelope detector D_2 , which has an RC load consisting of a $0.01\text{-}\mu\text{F}$ capacitor in parallel with a $25\text{-k}\Omega$ potentiometer. This potentiometer is the manual gain control, the output from which is fed to the audio preamplifier Q_4 . The audio power output stage consists of the push-pull pair Q_5, Q_6 .

Automatic gain control (AGC) is also obtained from the diode detector D_2 , the AGC filter network being the $15\text{-k}\Omega$ resistor and the $10\text{-}\mu\text{F}$ capacitor (Fig. 8.13.1). The AGC bias is fed to the Q_2 base.

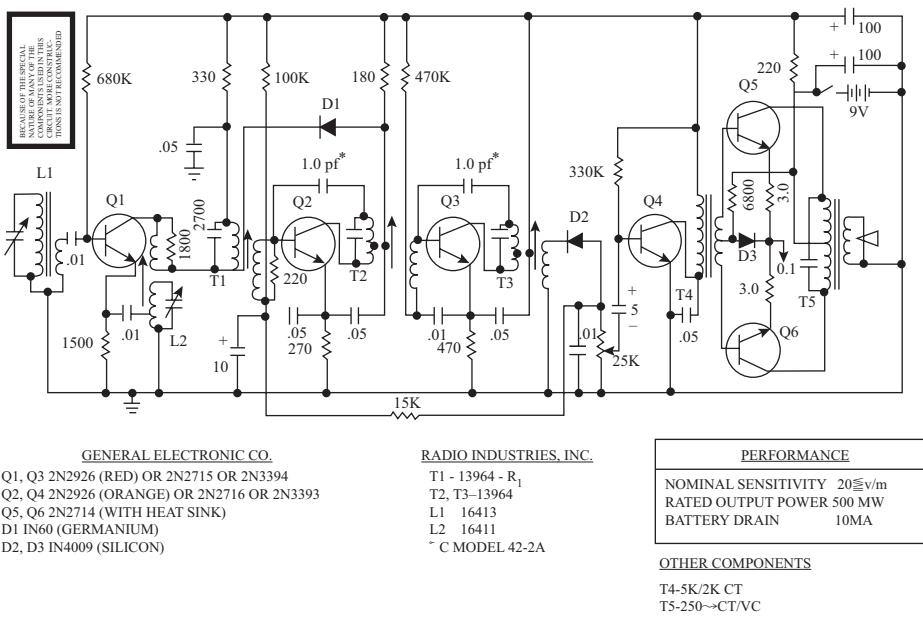


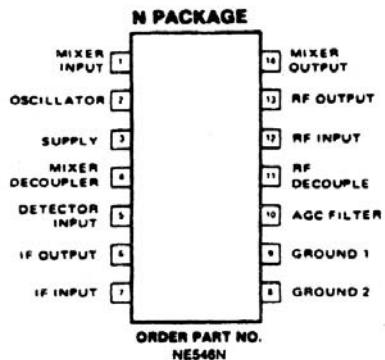
Figure 8.13.1 Six-transistor 9-V receiver. (Courtesy General Electric Company.)

Diode D_1 provides auxiliary AGC action. At low signal levels, D_1 is reverse biased, the circuit being arranged such that the collector of Q_1 is more positive than the collector of Q_2 . As the signal level increases, the normal AGC bias to Q_2 reduces Q_2 collector current, resulting in an increase in Q_2 collector voltage. A point is reached where this forward biases D_1 , the conduction of D_1 , then damping the T_1 primary and so reducing the mixer gain.

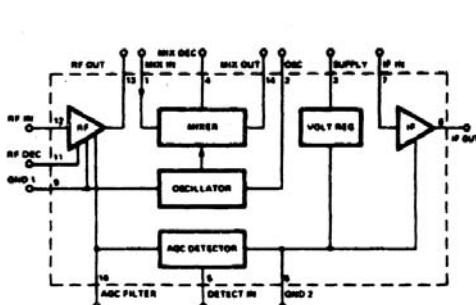
AM Receiver Using Integrated Circuit Subsystem

Figure 8.13.2 shows the details of the integrated subsystem, and Figure 8.13.3 shows how this is connected as an AM receiver. This package itself, illustrated in Fig. 8.13.2(a), is a 14-pin dual-in-line package, nominally measuring $19 \times 6.3 \times 3.2$ mm, and yet it contains six major circuit blocks, as shown in Figure 8.13.2(b). These blocks are integrated on one chip, the complete subsystem circuit being shown in Fig. 8.13.2(c). The circuit action is better understood in relation to the AM receiver application shown in Fig. 8.13.3. It will be immediately seen that certain bulky components, notably the tuned circuits, have to be added externally.

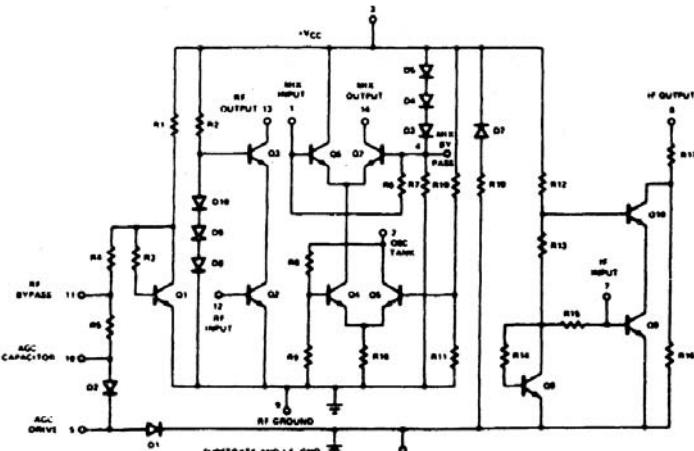
The AM signal input is fed via a tuned coupled circuit to pins 11 and 12, which are the RF inputs. Pin 11 is grounded at RF, while pin 12 is connected to the base of transistor Q_2 (Fig. 8.13.2(c)). Transistors Q_2 and Q_3 form a cascode amplifier (see Section 5.8). The RF output at pin 13 is tuned-circuit-coupled to the mixer input, pin 1. The mixer transistors are Q_6 , Q_7 , connected as an emitter-coupled pair with the base of Q_7 being grounded for RF and IF signals. The local oscillator signal is generated by another emitter-coupled pair Q_4 , Q_5 (requiring an external tuned circuit at pin 2). The oscillator is seen to control the current to the mixer pair [Fig. 8.13.2(c)], and hence multiplicative mixing occurs between the AM input and the oscillator signals. An IF output is obtained from the mixer output, pin 14, which is tuned-transformer-coupled back into the integrated circuit at pin 7. This goes to the input of a cascode IF amplifier pair Q_9 , Q_{10} , and the amplified IF output, at pin 6, is fed via a tuned transformer to an envelope detector.

PIN CONFIGURATION

(a)

BLOCK DIAGRAM

(b)



(c)

Figure 8.13.2 AM radio receiver subsystem available in integrated circuit form: (a) the package details, nominally measuring $19 \times 6.3 \times 3.2$ mm; (b) block diagram of the subsystem; (c) equivalent schematic of the subsystem. (Permission to reprint granted by Signetics Corporation, a subsidiary of U.S. Philips Corp., 811 E. Arques Avenue, Sunnyvale, CA 94086.)

The AGC in this case employs a diode system separate from the detector diode. The IF signal is coupled via the external 5-pF capacitor connected between pins 5 and 6 to the AGC diode D_2 [see Fig. 8.13.2(c)]. The rectifying action of D_2 , along with the filtering action of the externally connected capacitor at pin 10 (Fig. 8.13.3), produces an AGC bias that is fed through the coil between pins 11 and

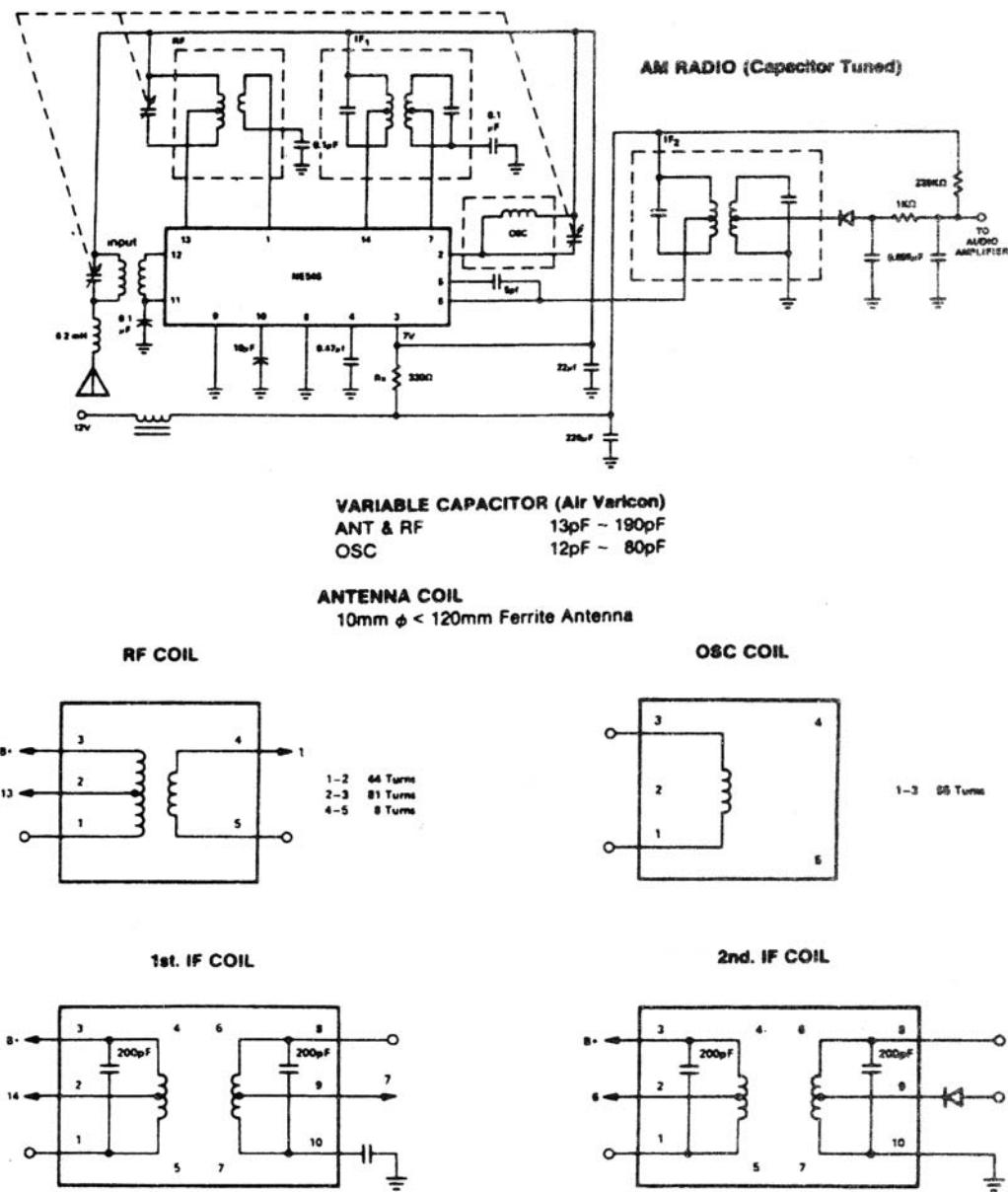


Figure 8.13.3 Details of an AM radio receiver utilizing the subsystem of Fig. 8.13.2. (Permission to reprint granted by Signetics Corporation, a subsidiary of U.S. Philips Corp., 811 E. Arques Avenue, Sunnyvale, CA 94086.)

12 (Fig. 8.13.3) to the base of the RF amplifier Q_2 . Transistor Q_1 , along with the associated circuitry, provides the reference bias for the AGC operation, and diode D_1 provides auxiliary AGC action for large signals through its damping action on the IF_2 transformer in a manner similar to that described for diode D_1 in Fig. 8.13.1.

Diode D_7 and resistor R_{19} form the voltage regulator, and the diode chains D_{10}, D_9, D_8 , and D_3, D_4, D_5 , provide bias levels for Q_3 and Q_7 , respectively. Transistor Q_8 sets the bias level for Q_9 .

The active circuitry is easily integrated on a single chip, and because transistors are much more economical to fabricate than resistors and capacitors in integrated circuits, the circuit design philosophy is to use transistors and diodes wherever possible for bias control.

AM Receiver Using a Phase-locked Loop (PLL)

Figure 8.13.4(a) shows the basic circuit blocks in a phase-locked loop (PLL), yet another approach to the use of integrated circuits in receivers. When the PLL locks onto the AM signal, the VCO frequency adjusts automatically to be equal to the AM carrier frequency. If the free-running value of the VCO frequency (i.e., its value before lock-in) is close to the AM carrier frequency, the VCO output voltage is approximately 90° out of phase with the AM carrier voltage.

To compensate for the 90° phase difference that occurs within the PLL, the carrier is externally shifted by a further 90° , as shown in Fig. 8.13.4(b). It does not matter whether the total phase difference is zero or 180° ; the fundamental component of the VCO output will be proportional to $\sin \omega_c t$, where ω_c is the angular frequency of the carrier.

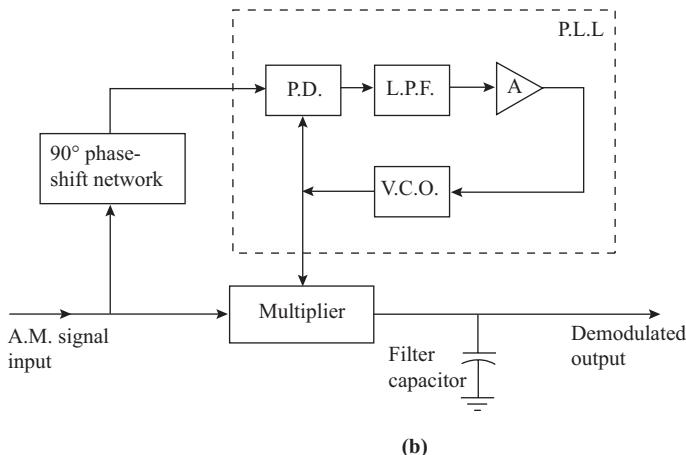
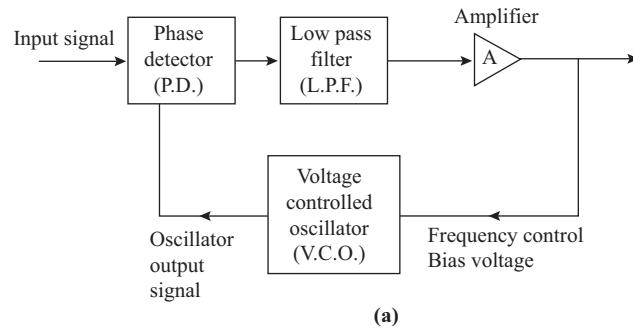


Figure 8.13.4 (a) Basic phase-locked loop (PLL). (b) AM receiver utilizing the PLL.

Let the modulating signal be represented by $m(t)$ and the AM signal, therefore, by $(1 + m(t))\cos \omega_c t$. The output from the multiplier [Fig. 8.13.4(b)] is proportional to

$$(1 + m(t))\cos \omega_c t \cos \omega_{oc} t = \frac{(1 + m(t))}{2} (1 + \cos 2\omega_c t) \quad (8.13.1)$$

The $\cos 2\omega_c t$ term is filtered out and the modulating signal recovered from the term $(1/2)m(t)$. This type of detection is known as *synchronous detection*.

The circuit for a receiver using a commercially available PLL is shown in Fig. 8.13.5(a), and the circuit for the PLL unit is shown in Fig. 8.13.5(b). For a complete description of the circuit, the reader is referred to the Signetics Analog Data Manual. Briefly, $Q11, 12, 13, 14$ form the multivibrator for the VCO. $Q10, 25, 6, 22$ amplify the feedback control signal from the PLL phase detector, acting through $Q21, 23$ to control the emitter currents of $Q12, 13$ and thus the charging current to the multivibrator timing capacitor C_o (connected to pins 2 and 3) to change the VCO frequency.

The PLL phase detector consists of $Q6, 7, 8, 9$ and $Q17, 18$. $Q6, 9$ and $Q7, 8$ are coupled through a resistor network to the VCO output $Q12, 13$. The AM signal (which may be the output of a conventional AM IF amplifier) is shifted by 90° by the network R_y, C_y and fed to the phase detector input on pin 13. Typical values are $R_y = 3000 \Omega$, $C_y = 135 \text{ pF}$ for standard broadcast reception. Bypass capacitor C_B from pin 12 completes the shifted signal return path. The output from the phase detector from $Q7$ drives the feedback amplifier $Q10, 25, 6, 22$. Capacitor C_L connected between pins 14, 15 provides the low-pass filtering of the control signal.

Transistors $Q1, 2, 3, 4$ and $Q15, 16$ form the multiplier circuit for the AM detector. This circuit operates in the same manner as the phase detector for the phase-locked loop. $Q1, 4$ and $Q2, 3$ are coupled to the VCO output (providing the synchronous carrier signal), and $Q15, 16$ are differentially driven from the unshifted AM input signal at pin 4. The demodulated output from $Q1, 3$ appears at pin 1, with C_x acting as a low-pass filter to remove the RF components before the signal is passed to an audio amplifier.

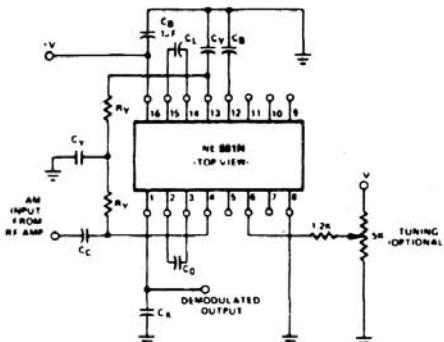
It is worth noting that the action of the two multiplier circuits in this chip is identical to that of the balanced mixer circuit described in Section 5.10, which is also a multiplier. Also, with minor connection changes, the same circuit can be used to demodulate an FM signal, operating in the manner described in Section 10.14.

It should also be noted that no external tuned circuits are required with the PLL detector, since, once the VCO locks onto the incoming carrier, selectivity is automatically achieved. In practice, some RF selectivity will be provided ahead of the PLL detector to prevent the VCO locking onto large unwanted carriers.

8.14 Noise in AM Systems

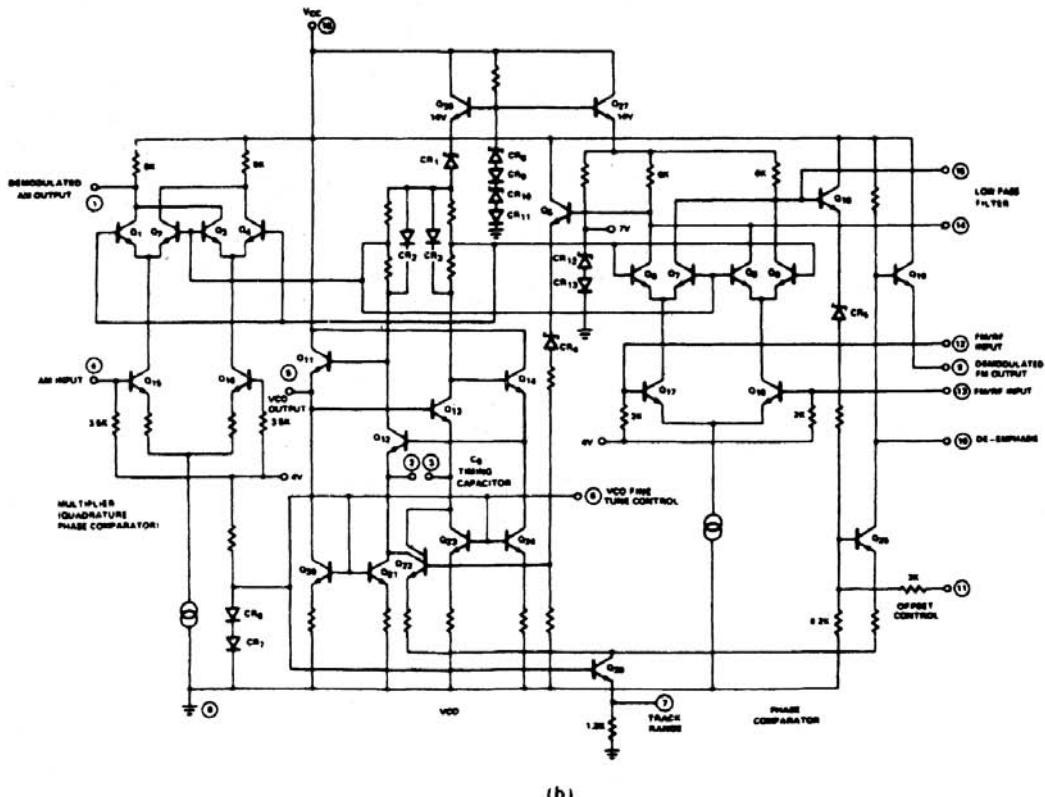
For AM systems that carry analog-type message signals, the signal-to-noise ratio is the most commonly used measure of performance. As discussed in Chapter 4, all the noise generated within the receiver can be referred to the receiver input, which makes it easy to compare receiver noise, antenna noise, and received signal. The antenna and receiver noise powers can be added, so the receiving system can be modeled as shown in Fig. 8.14.1.

From the point of view of determining the signal-to-noise ratio, the additional information needed is the bandwidth of the system. From Fig. 8.14.1, it is seen that between the antenna and the detector stage there is the bandwidth of the RF stages, followed by the bandwidth of the IF stages. Normally, the bandwidth of the IF stages is very much smaller than the RF bandwidth, and this will be the bandwidth that determines the noise reaching the detector. Following the detector, the bandwidth is that of the *baseband*, which is that required by the modulating signal. To distinguish these clearly, the baseband bandwidth will be denoted by W

PHASE LOCKED AM RECEIVER

C_B = Bypass Capacitor
 C_C = Coupling Capacitor

(a)

SCHEMATIC DIAGRAM OF 561N

(b)

Figure 8.13.5 (a) Phase-locked loop AM receiver. (b) Schematic diagram of the integrated circuit block used in the receiver. (Permission to reprint granted by Signetics Corporation, a subsidiary of U.S. Philips Corp., 811 E. Arques Avenue, Sunnyvale, CA 94086.)

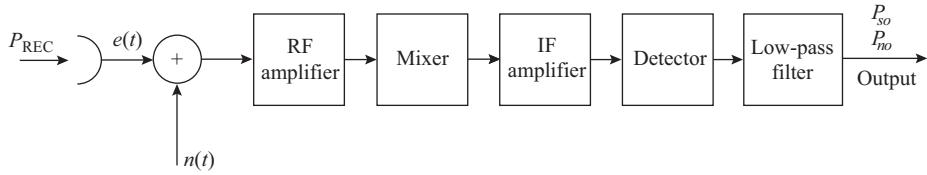


Figure 8.14.1 Model of an AM receiver showing noise added.

and the IF bandwidth by B_{IF} . Also, for AM systems it may be assumed that $B_{IF} \cong 2 W$. In Section 4.2, the concept of equivalent noise bandwidth was explained, and it will be further assumed that B_{IF} and W refer to the equivalent noise bandwidths.

Equation (4.20.3), which gives the noise output from a bandpass system, is rewritten here as

$$n_{IF}(t) = n_I(t) \cos 2\pi f_{IF}t - n_Q(t) \sin \omega_{IF}t \quad (8.14.1)$$

When the noise waveform is passed through the detector, the resulting noise output is very much dependent on whether or not a carrier is present and on the size of the carrier. When evaluating the signal-to-noise ratio, a carrier must be present. Let the modulated carrier be represented by

$$\begin{aligned} e(t) &= E_c \max(1 + m \cos 2\pi f_m t) \cos 2\pi f_{IF}t \\ &= A_c(t) \cos 2\pi f_{IF}t \end{aligned} \quad (8.14.2)$$

The input to the detector is therefore

$$\begin{aligned} e_{det}(t) &= e(t) + n_{IF}(t) \\ &= (A_c(t) + n_I(t)) \cos 2\pi f_{IF}t - n_Q(t) \sin 2\pi f_{IF}t \end{aligned} \quad (8.14.3)$$

The AM envelope detector recovers the envelope of this waveform, and therefore the waveform needs to be expressed as a cosine carrier wave of the form $R(t) \cos(\omega_{IF}t + \psi(t))$. Derivation of the amplitude and phase angle terms is left as an exercise for the student (see Problem 8.51). For the present application, the phase angle $\psi(t)$ can be ignored, and the amplitude term is given by

$$R(t) = \sqrt{[(A_c(t) + n_I(t))^2 + [n_Q(t)]^2]} \quad (8.14.4)$$

Although complete, this expression for $R(t)$ needs to be simplified to get a clearer picture of how the noise adds to the output. For many situations, it can be assumed that the AM carrier is much greater than the noise voltage for most of the time (remembering that the noise is random and that occasional large spikes will be encountered that are greater than the carrier). With this assumption, $R(t)$ simplifies to

$$\begin{aligned} R(t) &= \sqrt{[(A_c(t))^2 + 2A_c(t)n_I(t) + [n_I(t)]^2 + [n_Q(t)]^2]} \\ &\approx \sqrt{[(A_c(t))^2 + 2A_c(t)n_I(t)]} \end{aligned} \quad (8.14.5)$$

A further expansion and simplification of the square-root term can be made using the binomial theorem.

$$\begin{aligned} R(t) &= \sqrt{[(A_c(t)]^2 + 2A_c(t)n_I(t)} \\ &= A_c(t) \left(1 + 2\frac{n_I(t)}{A_c(t)}\right)^{1/2} \\ &= A_c(t) + n_I(t) \end{aligned} \quad (8.14.6)$$

For sinusoidal modulation, this becomes

$$R(t) = E_{c \text{ max}} + mE_{c \text{ max}} \cos 2\pi f_m t + n_I(t) \quad (8.14.7)$$

The envelope is seen to consist of a dc term, the modulating signal voltage, and the noise voltage $n_I(t)$. As shown in Section 8.11, the dc output from the detector is blocked so that only the ac components contribute to the final output. For the noise, the available power spectral density is given by Eq. (4.20.4) as $2kT_s$, and hence the available noise power output is

$$P_{no} = 2kT_s W \quad (8.14.8)$$

The peak signal voltage at the output is $mE_{c \text{ max}}$, and hence the rms voltage is mE_c , where E_c is the rms voltage of the unmodulated carrier at the receiver. The available signal power output is therefore

$$P_{so} = \frac{m^2 E_c^2}{4R_{\text{out}}} \quad (8.14.9)$$

The output signal-to-noise ratio is

$$\begin{aligned} \left(\frac{S}{N}\right)_o &= \frac{P_{so}}{P_{no}} \\ &= \frac{m^2 E_c^2}{8R_{\text{out}} k T_s W} \end{aligned} \quad (8.14.10)$$

Standard practice is to compare the output signal-to-noise ratio to a reference ratio, which is the signal-to-noise ratio at the detector input *but with the noise calculated for the baseband bandwidth* (W in this case). The noise power spectral density at the detector input is kT_s , and so the reference noise power

$$P_{n \text{ REF}} = kT_s W \quad (8.14.11)$$

The available signal power from a source with internal resistance R_s is [see Eq. (8.6.3)]

$$P_R = \frac{E_c^2}{4R_s} \left(1 + \frac{m^2}{2}\right) \quad (8.14.12)$$

Hence the reference signal-to-noise ratio is

$$\begin{aligned} \left(\frac{S}{N}\right)_{\text{REF}} &= \frac{P_R}{P_{n \text{ REF}}} \\ &= \frac{E_c^2 (1 + m^2/2)}{4R_s k T_s W} \end{aligned} \quad (8.14.13)$$

A figure of merit that is used is the ratio of these two signal-to-noise ratios. Denoting this by R_{AM} , then

$$\begin{aligned} R_{AM} &= \frac{(S/N)_o}{(S/N)_{REF}} \\ &= \frac{m^2}{(2 + m^2)} \frac{R_s}{R_{out}} \end{aligned} \quad (8.14.14)$$

The higher the figure of merit, the better the system. Normally, $R_{out} \cong R_s$ and hence the highest value is $\frac{1}{3}$, achieved at 100% modulation.

In the case of sinusoidal DSBSC, the received signal is of the form

$$e(t) = E_{max} \cos \omega_m t \cos 2\pi f_{IF} t \quad (8.14.15)$$

where E_{max} is the peak value of the received signal. The input to the detector is therefore

$$\begin{aligned} e_{det}(t) &= e(t) + n(t) \\ &= (E_{max} \cos \omega_m t + n_I(t)) \cos 2\pi f_{IF} t - n_Q(t) \sin \omega_{IF} t \\ &= A(t) \cos \omega_{IF} t - n_Q(t) \sin \omega_{IF} t \end{aligned} \quad (8.14.16)$$

where $A(t) = (E_{max} \cos \omega_m t + n_I(t))$. With DSBSC a different type of detection is used, known as *coherent detection*. Demodulation of the DSBSC signal utilizes a balanced mixer (or similar circuit) as described in Section 5.10. A locally generated carrier is required that is exactly locked onto the incoming carrier $\cos \omega_{IF} t$, and the two signals are fed into the balanced mixer. One complication with DSBSC detection (which also applies to SSB detection as described in Chapter 9) is generating the local carrier. However, methods are available for achieving this, and as a result the output of the balanced demodulator is

$$e_{out}(t) = k e_{det}(t) \cos \omega_{IF} t \quad (8.14.17)$$

where k is a constant of the multiplier circuit. Multiplying this out in full, which is left as an exercise for the student, results in

$$e_{out}(t) = \frac{k}{2} (E_{max} \cos \omega_m t + n_I(t)) + \text{high frequency terms} \quad (8.14.18)$$

Low-pass filtering following the balanced demodulator removes the high-frequency terms, leaving, as the baseband output,

$$e_{BB}(t) = \frac{k}{2} (E_{max} \cos \omega_m t + n_I(t)) \quad (8.14.19)$$

Thus, apart from the multiplying constant $k/2$ and the absence of a dc term, this output is the same as that given in Eq. (8.14.7) for the standard AM case. The $k/2$ factor is common to signal and noise and can be ignored. It should be noted, however, that, whereas the standard AM result requires that the carrier be much greater than the noise, this approximation is not required for the DSBSC case. Equation (8.14.10) applies for

the output signal-to-noise ratio, but with E_{\max} replacing $mE_c \max$ as seen by comparing eqs. (8.14.19) and (8.14.7). The result is

$$\begin{aligned} \left(\frac{S}{N}\right)_o &= \frac{(E_{\max}/\sqrt{2})^2}{8R_{\text{out}}kT_sW} \\ &= \frac{E_{\max}^2}{16R_{\text{out}}kT_sW} \end{aligned} \quad (8.14.20)$$

For the reference signal-to-noise ratio, the reference noise is $P_n \text{REF} = kT_sW$ as given by Eq. (8.14.11). The rms voltage of the received (DSBSC) signal $E_{\max} \cos \omega_m t \cos \omega_{\text{IFT}} t$ is $E_{\max}/2$ (note not $E_{\max}/\sqrt{2}$) as is readily checked from ac circuit theory. The available signal power at the input is therefore

$$\begin{aligned} P_R &= \frac{(E_{\max}/2)^2}{4R_s} \\ &= \frac{E_{\max}^2}{16R_s} \end{aligned} \quad (8.14.21)$$

The reference signal-to-noise is therefore

$$\left(\frac{S}{N}\right)_{\text{REF}} = \frac{E_{\max}^2}{16R_s kT_s W} \quad (8.14.22)$$

The figure of merit is therefore

$$\begin{aligned} R_{\text{DSBSC}} &= \frac{(\text{S/N})_o}{(\text{S/N})_{\text{REF}}} \\ &= \frac{R_s}{R_{\text{out}}} \end{aligned} \quad (8.14.23)$$

For $R_{\text{out}} \cong R_s$, the figure of merit is unity, which is three times better than the best that can be achieved for R_{AM} . It will be shown in Chapter 9 that single-sideband (SSB) transmission also has a unity figure of merit, and since this requires half the bandwidth of a DSBSC signal, it is the preferred method of AM carrier transmission (apart from AM broadcast applications, where simplicity of receiver design has established standard AM as the preferred method).

PROBLEMS

- 8.1.** A sinusoidal carrier is amplitude modulated by a square wave that has zero dc component and a peak-to-peak value of 2 V. The periodic time of the square wave is 0.5 ms. The carrier amplitude is 2.5 V, and its frequency is 10 kHz. Write out the equations for the modulating signal, the carrier, and the modulated wave, and plot these functions over a time base equal to twice the periodic time of the square wave.
- 8.2.** A sinusoidal carrier is amplitude modulated by a triangular wave. The triangular wave has zero mean value and is an even function, the first quarter-cycle being described by $-1 + 8t$ volts, with t in milliseconds. The

periodic time of the triangular wave is 0.5 ms. The carrier amplitude is 2.5 V, and its frequency is 100 kHz. Write out the equations for the modulating signal, the carrier, and the modulated wave, and plot these functions over a time base equal to twice the periodic time of the triangular wave.

- 8.3. Calculate the modulation index for each of the modulated waves in Problems 8.1 and 8.3.
- 8.4. A sinusoidal carrier has an amplitude of 10 V and frequency 30 kHz. It is amplitude modulated by a sinusoidal voltage of amplitude 3 V and frequency 1000 Hz. Plot accurately to scale the modulated waveform, showing two complete cycles of the modulating wave, and determine the modulation index.
- 8.5. For a standard AM transmission, the maximum peak-to-peak voltage is 150 V and the minimum peak-to-peak voltage is 50 V. Calculate the modulation index.
- 8.6. Why is it important in AM broadcast transmissions to prevent 100% modulation depth being exceeded? Describe the trapezoidal method for monitoring such transmissions.
- 8.7. By using a suitable computer routine (such as Mathcad), or otherwise, plot the trapezoidal pattern for the modulated wave of Problem 8.1.
- 8.8. By using a suitable computer routine (such as Mathcad), or otherwise, plot the trapezoidal pattern for the modulated wave of Problem 8.2.
- 8.9. By using a suitable computer routine (such as Mathcad), or otherwise, plot the trapezoidal pattern for the modulated wave of Problem 8.4.
- 8.10. A carrier of 10 V peak and frequency 100 kHz is amplitude modulated by a sine wave of 4 V peak and frequency 1000 Hz. Determine the modulation index for the modulated wave and draw the amplitude spectrum.
- 8.11. A 45-V (rms) carrier is amplitude modulated by a 30-V (rms) sine wave. Determine (a) the maximum and minimum values of the peak-to-peak voltage of the modulated wave, (b) the amplitude of the side frequencies, and (c) the modulation index.
- 8.12. The carrier for a standard AM transmission is given by $e_c = 10 \sin 2\pi 10^5 t$, and one side frequency for sinusoidal modulation is given by $e_L = 3 \cos 1.96\pi 10^5 t$. Determine the expression for the other side frequency. Using a suitable computer routine (such as Mathcad) reconstruct the modulated wave by summing the three spectrum components. Plot the result for at least two cycles of the modulating waveform.
- 8.13. Determine for the Problem 8.12 the equation for the modulating waveform, stating clearly the amplitude and frequency of the wave.
- 8.14. A standard AM transmission, sinusoidally modulated to a depth of 30%, produces side frequencies of 4.928 and 4.914 MHz. The amplitude of each side frequency is 75 V. Determine the amplitude and frequency of the carrier.
- 8.15. A modulating signal given by $e_m = 2 \sin 2\pi 10^4 t$ volts is used to amplitude modulate a carrier given by $e_c = 10 \sin 2\pi 10^6 t$, where t is the time in seconds. The modulated voltage wave is developed across a $50\text{-}\Omega$ load resistor. (a) Write down the expression for the modulated wave. (b) Draw accurately to scale the spectrum for the modulated wave. (c) Calculate the rms current in the load. (d) Calculate the total average power. (e) Calculate the power at one side frequency. (f) Calculate the power in the carrier component.
- 8.16. An AM transmitter has an unmodulated power output of 100 W. When sinusoidally modulated, the power increases to 132 W. The transmitter feeds a resistive load of $100\text{ }\Omega$. (a) Calculate the modulation index. (b) For the trapezoidal method of measurement of modulation index, determine the ratio of the lengths of the parallel sides of the trapezoid. (c) Neatly sketch the spectrum for a carrier frequency of 1 MHz and a modulating frequency of 10 kHz.

- 8.17.** A sinusoidal carrier has a peak value of 100 V and a frequency of 100 kHz. Standard amplitude modulation is employed, the modulating signal being a sine wave of amplitude 75 V and frequency 5 kHz. Determine (a) the modulation index. (b) the amplitude and frequencies of the side frequencies, and (c) the power in each spectral component. The load resistance is 300 Ω .
- 8.18.** The output power of an AM transmitter is 1 kW when sinusoidally modulated to a depth of 100%. Calculate the power at each side frequency when the modulation depth is reduced to 50%.
- 8.19.** Calculate (a) the total power and (b) the power in each side frequency for a standard AM transmission that is sinusoidally modulated to a depth of 80%, if the unmodulated carrier power is 50 kW.
- 8.20.** Determine the rms voltage for the modulated wave in Problems 8.4 and 8.10.
- 8.21.** Determine the rms voltage for the modulated wave in Problems 8.11 and 8.12.
- 8.22.** Determine the rms voltage and the rms load current for the modulated wave of Problem 8.15.
- 8.23.** Define the term *modulation index* as applied to an amplitude modulated wave. A modulating signal is given by $e_m = \sin \omega_{mt} + 3 \sin 3\omega_{mt}$ and the carrier by $e_c = 10 \sin \omega_ct$, where all amplitudes are in volts, and $f_c = 500$ kHz and $f_m = 4$ kHz. Neatly draw the spectrum for the modulated wave and determine the average powers for (a) the carrier, (b) each side frequency, and (c) the complete modulated wave, given that the load resistance is 50 Ω .
- 8.24.** Using a computer routine such as Mathcad, plot the modulated waveform for Problem 8.23 over two cycles of the modulating waveform.
- 8.25.** Two sinusoidal signals simultaneously amplitude modulate a carrier to produce a standard AM wave. Signal 1 has an amplitude of 3 V and frequency of 300 Hz, and signal 2 an amplitude of 4 V and frequency of 800 Hz. The carrier has an amplitude of 10 V and frequency of 20 kHz. Using a computer routine such as Mathcad, plot the modulating waveform and the modulated waveform over two cycles of the lowest-frequency component in the combined modulating waveform.
- 8.26.** Is the combined modulating waveform in Problem 8.25 periodic? Give reasons for your answer.
- 8.27.** Plot the amplitude spectrum for the modulated waveform of Problem 8.25, and determine the total power in the modulated wave. The load resistance is 300 Ω .
- 8.28.** Determine the rms voltage and the rms load current for the modulated wave of Problem 6.23.
- 8.29.** Determine the rms voltage for the modulated wave of Problem 8.25.
- 8.30.** For Problem 8.16, determine the rms load voltage and current for the modulated and unmodulated conditions.
- 8.31.** Determine the spectrum bandwidths for the modulated signals of Problems 8.23 and 8.25.
- 8.32.** A modulating signal $e_m = \sin \omega_{mt} + 3 \sin 3\omega_{mt}$ is used to multiply a carrier given by $e_c = 10 \sin \omega_ct$, where all amplitudes are in volts, and $f_c = 500$ kHz and $f_m = 4$ kHz. The multiplier constant is $k = 1\text{V}^{-1}$. Neatly draw the spectrum for the modulated wave and determine the average power in each side frequency, given that the load resistance is 50 Ω . Compare the spectrum with that obtained in Problem 8.23.
- 8.33.** Repeat Problem 8.25 for the situation where the modulating signals are used to multiply the carrier, producing a DSBSC signal. Assume a multiplier constant $k = 0.1\text{V}^{-1}$.
- 8.34.** Describe the operation of a class C modulator in which the modulating signal is applied to the collector. If the dc power to such a modulator is 500 W, determine the power the modulator must supply for 100% sinusoidal modulation.
- 8.35.** Explain why in a collector modulator the operating range must include the saturation line of the output characteristics.
- 8.36.** For a class C collector modulator, the dc power to the collector is 300 W unmodulated, for which the carrier output power to an antenna is 210 W. Calculate the overall conversion efficiency of the stage.

- Assuming the efficiency remains constant, calculate the additional power that must be supplied by the modulating amplifier to achieve 70% modulation depth for sinusoidal modulation.
- 8.37.** The voltage transfer ratio for the output transformer of a class C amplifier is $K = 0.7$. When unmodulated, the voltage across the output load has a peak-to-peak value of 120 V. Calculate the peak-to-peak voltage at the collector. Assuming that the minimum collector voltage (reached at the minimum of each RF cycle) is zero, calculate the dc voltage and the peak voltage at the collector.
- 8.38.** For a class C collector modulator, the direct collector voltage is 12 V and the direct collector current is 12 A. The unmodulated rms current to a $50\text{-}\Omega$ load connected to the modulator is 1.5 A. When sinusoidal modulation is applied and monitored using the trapezoidal method, the parallel sides of the trapezoid are 1 and 5 cm. Find (a) the modulation index, (b) the output carrier power (unmodulated), (c) the output power with modulation, (d) the dissipation in the transistor with modulation applied (assuming the transistor is the only source of losses), and (e) the collector conversion efficiency (assumed constant).
- 8.39.** For a class C, pentode plate modulator, the maximum peak-to-peak voltage is 1000 V and the minimum peak-to-peak voltage is 50 V. Calculate the modulation index.
- 8.40.** For a pentode class C plate modulator, the steady plate potential is 1500 V. Assuming that the plate voltage goes to zero at the minimum of each RF cycle, determine the maximum plate voltage reached for 100% modulation. The direct current to the plate is 8 A. Calculate the dc plate power and the RF output power, given that the plate conversion efficiency is 72%.
- 8.41.** For the pentode in Problem 8.40, calculate the maximum peak-to-peak output voltage, given that the output load is $50\text{ }\Omega$.
- 8.42.** For the pentode in Problem 8.40, calculate the power that must be supplied by the modulating amplifier to achieve 100% modulation. State any assumptions made.
- 8.43.** A diode detector load consists of a $0.01\text{-}\mu\text{F}$ capacitor in parallel with a $5\text{-k}\Omega$ resistor. Determine the maximum depth of sinusoidal modulation that the detector can handle without diagonal peak clipping when the modulating frequency is (a) 1000 Hz and (b) 10,000 Hz.
- 8.44.** Distinguish between *negative peak clipping* and *diagonal peak clipping* in an envelope detector. The output of a diode envelope detector is fed through a dc blocking capacitor to an amplifying stage, which has an input resistance of $10\text{ k}\Omega$. If the diode load resistor is $5\text{ k}\Omega$, determine the maximum depth of sinusoidal modulation the detector can handle without negative peak clipping.
- 8.45.** If the detector in Problem 8.43 is coupled through a $0.1\text{-}\mu\text{F}$ capacitor into an amplifier with an input resistance of $50\text{ k}\Omega$, what modulation index can the detector handle without negative peak clipping occurring?
- 8.46.** The dc power to a modulated class C amplifier is 500 W. Determine the power the modulator must supply for 100% sinusoidal modulation applied at the output electrode.
- 8.47.** The rms antenna current from an AM transmitter increases by 15% over the unmodulated value when sinusoidal modulation is applied. Determine the modulation index.
- 8.48.** The modulator circuit of Fig. 8.10.1 is found to produce some envelope distortion even though it is properly adjusted. How could this non-linearity be removed?
- 8.49.** How is synchronous demodulation of an amplitude-modulated signal accomplished? Show a circuit that will do this, and explain how it works.
- 8.50.** What should the minimum voltage rating of the collector of $Q3$ in Fig. 8.10.4 be? What total power with 100% modulation could the circuit be expected to put out? What power output could be expected if the battery voltage dropped to 11.5 V?

- 8.51.** Starting from Eq. (8.14.3), derive Eq. (8.14.4) and show that the phase angle $\psi(t)$ is given by

$$\psi(t) = \tan^{-1} \frac{n_Q(t)}{A_c(t) + n_I(t)}$$

- 8.52.** Given that the received modulating signal power is $P_m = 0.1$ pW for a standard sinusoidal AM wave and that the baseband bandwidth is $W = 4$ kHz, calculate the output signal-to-noise ratio in decibels. The system noise temperature is 500 K. Calculate the reference signal-to-noise ratio for a modulation index of 0.5. Boltzmann's constant $k = 1.38 \times 10^{-23}$ J/K. Assume $R_s = R_{\text{OUT}}$.
- 8.53.** Calculate the figure of merit in dB for the system in Problem 8.52.
- 8.54.** In Eq. (8.14.17), what are the dimensions of the constant k ?
- 8.55.** Suppose that the received carrier as given by Eq. (8.14.2) was of the form $A_c(t) \sin \omega_{IF}t$. Would this make any difference to the final outcome? Explain your answer.
- 8.56.** A DSBSC receiving system has an equivalent noise temperature (including antenna noise) of 1000 K and a baseband bandwidth of 4 kHz. The received signal from a sinusoidally modulated DSBSC wave is 1 μ V rms across 50 Ω . Calculate (a) the output signal-to-noise ratio, (b) the reference signal-to-noise ratio, and (c) the figure of merit. Assume $R_s = R_{\text{OUT}}$.
- 8.57.** Generate *Amplitude Modulation* using the carrier $e_c(t) = 5\sin(2000\pi t)$ and modulating signal $e_m(t) = \sin(200\pi t)$. Let the modulation depth $m_a = 0.5$. Use MATLAB to generate and plot.
- 8.58.** Explore the *modulate(.)* function of MATLAB to generate AM.
- 8.59.** The current of the transmitting antenna of an AM transmitter is 20A when only carrier is sent, but it rises to 24A when modulated by a single sine wave. (a) Find the modulation index. (b) Determine the antenna current when the modulation index is changed to 90%.
- 8.60.** An AM broadcast transmitter radiates 30kW of power when unmodulated and 36.5kW when modulated by a sine wave. (a) Calculate the modulation index/percentage of modulation. (b) If another sine wave simultaneously modulates the carrier to the depth of 60%, determine the total power transmitted.
- 8.61.** A 10MHz carrier is simultaneously modulated with 6000Hz, 7000Hz, and 30KHz audio sine wave signals. What will be the frequencies present at the output?
- 8.62.** For a carrier power of 20kW, plot the total transmitter power of modulation index varying from 0 to 100%. Use MATLAB for coding.
- 8.63.** The output power of a 80% modulated AM generator is 2.8A. (a) To what value will this current rise if it is simultaneously modulated by another signal to the extent of 60%? (b) What will be the percentage saving in power if the carrier and one of the side-bands are suppressed before transmission?
- 8.64.** Three sine waves simultaneously amplitude modulate a carrier. Signal 1 is of amplitude 4V and frequency 400Hz, while the second one is of amplitude 5V and frequency 500Hz. The third signal has an amplitude of 10V at a frequency 1000Hz. The carrier amplitude is 25V and its frequency is 20kHz. Plot the modulating waveform and the A_{Mout} signal using MATLAB.
- 8.65.** A sinusoidal carrier is amplitude modulated by a square wave that has zero DC component and a peak-to-peak value of 20V. The period of the square wave is 5ms. The carrier amplitude is 25V and its frequency is 10kHz. (a) Write the equations for the modulating signal and the modulated signal. (b) Plot the above signals overtime and obtain the modulation index.
- 8.66.** A carrier wave of frequency 100kHz and peak value 10V is amplitude modulated by a 2kHz sine wave of amplitude 5V. Determine the *modulation index* and the *amplitude spectrum*.



Single-sideband Modulation

9.1 Introduction

Communications in the HF bands have become increasingly crowded in recent years, requiring closer spacing of signals in the spectrum. Single-sideband systems requiring only half the bandwidth of normal AM and considerably less power are used extensively in this portion of the spectrum as a result.

It was noted in Chapter 8 that each sideband of a normal AM signal contains all the information necessary for signal transmission and recovery. It was also pointed out that for 100% sinusoidal modulation each sideband contains one-sixth of the total signal power, while the carrier contains two-thirds of the total power. Furthermore, the carrier itself carries no information contributed by the modulating signal. Figure 9.1.1 shows a comparison of the signal spectra of normal AM (DSBFC) in (a), double-sideband suppressed carrier (DSBSC) in (b), upper-sideband SSB (SSBSC) in (c), and lower-sideband SSB in (d). Note that in (c) and (d) only one sideband is present and that each requires only one-half of the bandwidth of either (a) or (b).

9.2 Single-sideband Principles

In Section 8.9 it is shown that the output from a balanced modulator contains the term

$$e(t) = k e_m(t) \cos \omega_c t \quad (9.2.1)$$

where k is the multiplier constant. With sinusoidal input $e_m(t) = E_m \max \cos \omega_m t$, the modulator output becomes a DSBSC signal containing two side frequencies,

$$\begin{aligned} e(t) &= k E_m \max \cos \omega_m t \cos \omega_c t \\ &= E_{\max} [\cos(\omega_c - \omega_m)t + \cos(\omega_c + \omega_m)t] \end{aligned} \quad (9.2.2)$$

where $E_{\max} = k(E_m \max /2)$.

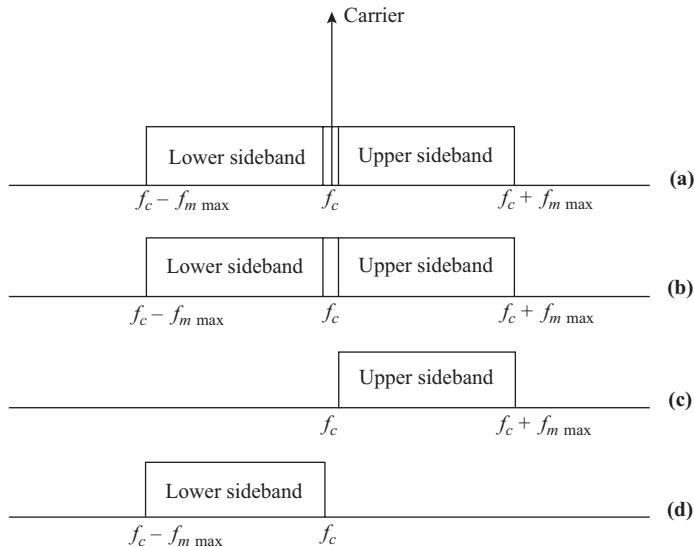


Figure 9.1.1 Amplitude-modulated signal spectra: (a) normal amplitude modulation, or double-sideband full carrier; (b) double-sideband suppressed carrier (DSBSC); (c) single-sideband suppressed carrier (SSBSC) using the upper sideband (USB); (d) single-sideband suppressed carrier (SSBSC) using the lower sideband (LSB).

Now if one of the side frequencies in the DSBSC signal is removed, either by filtering or by cancellation, the other side frequency will remain. For cosinusoidal modulation the *upper side frequency* (USF) signal is described by

$$e_{\text{USF}} = E_{\max} \cos(\omega_c + \omega_m)t \quad (9.2.3)$$

and the *lower side frequency* (LSF) signal is described by

$$e_{\text{LSF}} = E_{\max} \cos(\omega_c - \omega_m)t \quad (9.2.4)$$

Since all the transmitted power goes into the side frequency, then

$$P_T = \frac{E_{\max}^2}{2R} \quad (9.2.5)$$

This should be compared to Eq. (8.6.3) for a standard AM signal.

Where the modulating signal contains a band of frequencies (usually the case in practice), the terms *upper sideband* (USB) and *lower sideband* (LSB) are used.

Demodulation of a single-sideband signal is achieved by multiplying it with a locally generated *synchronous* carrier signal at the receiver. Detectors using this principle are called *product detectors*, and balanced modulator circuits are used for this purpose. It is important that the carrier be as closely synchronized in frequency and phase with the original carrier as possible to avoid distortion of the modulated output.

To demonstrate that the multiplying process does demodulate an SSB signal, consider an LSF signal $E_{\max} \cos(\omega_c - \omega_m)t$ multiplied by a local oscillator signal $E_c \cos \omega_c t$ using a balanced modulator with a gain k . Equation (5.10.11) shows that the mixer operating with a large oscillator input signal contains a term

$$e_{\text{out}} = kE_{\max} \cos(\omega_c - \omega_m)t \cos \omega_c t$$

$$= \frac{kE_{\max}}{2} [\cos \omega_m t + \cos(2\omega_c - \omega_m)t] \quad (9.2.6)$$

The first term on the right of the equation is the required information signal, while the second term is the lower side frequency at the second harmonic of the local carrier frequency. Low-pass filtering easily removes this, leaving only the demodulated information (or baseband) signal as

$$e_{bb}(t) = \frac{kE_{\max}}{2} \cos \omega_m t \quad (9.2.7)$$

9.3 Balanced Modulators

Balanced modulators are the building blocks from which a wide variety of frequency mixers, modulators, and demodulators are built. Any circuit that multiplies two input signals while canceling the feedthrough of one of these is a singly balanced modulator, and one that cancels both is a doubly balanced modulator. The output contains a double-sideband suppressed carrier signal.

An FET Singly Balanced Modulator Circuit

Figure 9.3.1 shows two matched FETs connected in a differential amplifier, which acts as a singly balanced modulator in which the carrier oscillator signal is canceled from the output, but the modulating signal appears in the output. The input (modulating) signal is applied in the differential input mode, and the carrier signal is applied as a common-mode signal. The signal applied to the gate of M_1 is the sum of the two input voltages ($e_c + e_m$), while the signal applied to M_2 is the difference ($e_c - e_m$). These two components are squared by the second-order terms of the transistor transfer functions. The common-mode carrier signal remaining is canceled as the two drain currents are subtracted in the output transformer primary.

$$V_{gs1} = e_c + e_m \quad (9.3.1)$$

$$V_{gs2} = e_c - e_m \quad (9.3.2)$$

$$i_{d1} = I_o + aV_{gs1} + bV_{gs1}^2 \quad (9.3.3)$$

$$i_{d2} = I_o + aV_{gs2} + bV_{gs2}^2 \quad (9.3.4)$$

$$i_p = i_{d1} - i_{d2} = a(V_{gs1} - V_{gs2}) + b(V_{gs1} + V_{gs2})(V_{gs1} - V_{gs2}) \quad (9.3.5)$$

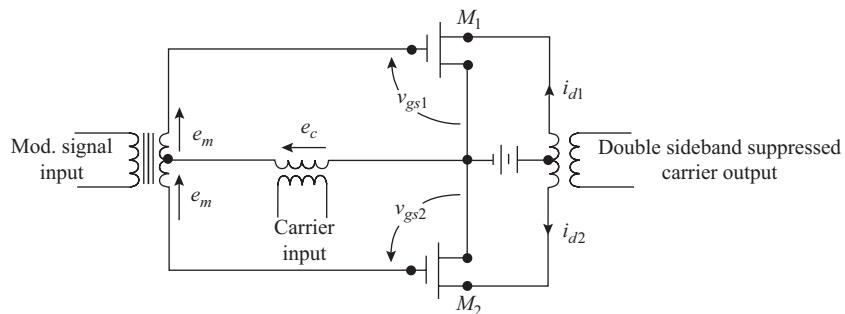


Figure 9.3.1 FET singly balanced modulator circuit.

Substituting Eqs. (9.3.1) and (9.3.2) into (9.3.5) gives

$$i_p = 2a(e_m) + 4b(e_m)(e_c) \quad (9.3.6)$$

Substituting sinusoidal signals in Eq. (9.3.6) yields the output

$$i_p = 2aE_{m \max} \cos \omega_m t + 2bE_{c \max} E_{m \max} [\cos(\omega_c - \omega_m)t + \cos(\omega_c + \omega_m)t] \quad (9.3.7)$$

This output contains the original modulating signal and the two sidebands about the carrier frequency position. The carrier is absent. It should be noted that any imbalance in the circuit so that either the a 's or b 's for the two FETs differ from each other will allow some of the carrier signal to feed through to the output. In practice, the FETs would be a very closely matched pair on a single chip, and the bias currents to the two FETs would be adjusted for minimum carrier feedthrough. Since the output would be fed through a band-pass filter, the low-frequency modulating signal component would be removed at that point.

Integrated-circuit Doubly Balanced Modulators

The disadvantages of the FET circuit described are that the modulating input signal feeds through to the output, the circuit is difficult to balance, and the input and output require specially balanced transformers. Integrated-circuit doubly balanced modulators like the LM1596 described in Section 5.10 operate as multiplier circuits that produce only sideband pairs at the output. Application is simple, requiring only bias and an appropriate band-pass filter to eliminate sideband pairs at harmonics of the carrier. Very little adjustment is required to obtain good balance.

An important advantage of the integrated-circuit balanced modulator is that, when it is operated with a large carrier signal, the output signal amplitude is independent of the carrier amplitude. The result is that the output amplitude depends only on the amplitude of the input signal (which is the modulating signal when it is used as a modulator or the sideband signal when it is used as a demodulator).

Doubly Balanced Diode Ring Modulator

A circuit known as the *double-balanced ring modulator*, which is widely used in carrier telephony, is shown in Fig. 9.3.2(a). The name comes from the fact that the circuit is balanced to reject both the carrier and modulating signals using a ring of diodes. The output contains only sideband pairs about the carrier frequency position and several of its harmonics.

Operation of the circuit is similar to that described for the integrated-circuit balanced modulator in Section 5.10. A large signal carrier acts as a switching signal to alternate the polarity of the modulating signal at the carrier frequency. With a negative carrier voltage V_c applied, diodes AB and CD conduct and diodes AD and BC block to give the effective connection shown in Fig. 9.3.2(b). With a positive carrier voltage, diodes AD and BC conduct and diodes AB and CD block to give the connection of Fig. 9.3.2(c). The effect is to multiply the modulating signal by a fixed-amplitude square wave at the carrier frequency, producing the required DSBSC signal, with harmonics. Band-pass filters remove the unwanted harmonics from the output. Fig. 9.3.3 shows the time response waveform for a single sinusoid of modulation and its spectrum.

These circuits have been extensively used for low-frequency telephone applications, where they require balanced input and output transformers and some adjustment of circuit balance for good performance. Care must be exercised if the ring modulator is used at radio frequencies, since the high-level carrier signal may result in the radiation of interference. The doubly balanced diode ring circuit is widely used as a mixer in microwave applications where shielded enclosures prevent radiation.

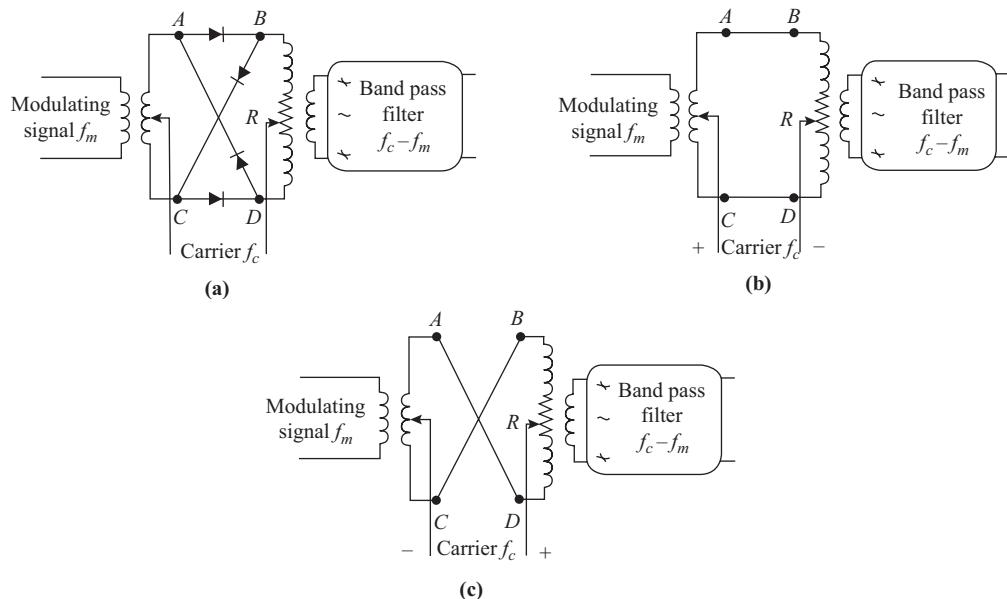


Figure 9.3.2 (a) Double-balanced ring modulator; (b) the conducting paths when diodes AB and CD are forward biased; (c) the conducting paths when diodes BC and DA are forward biased.

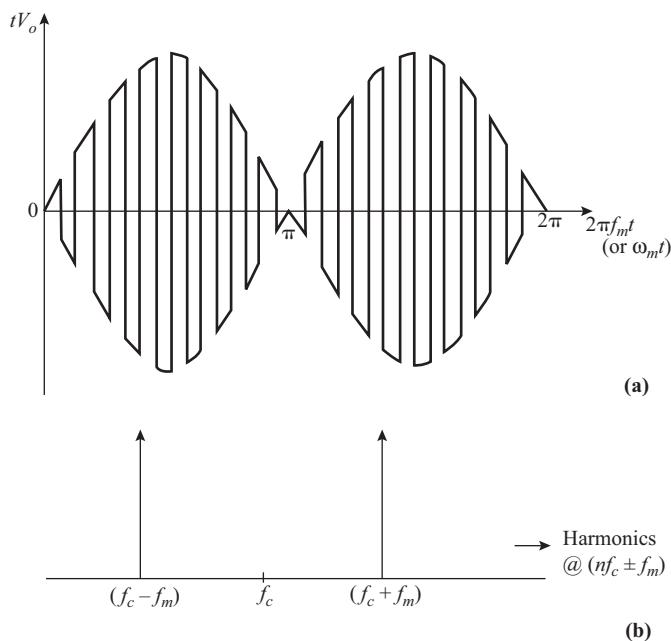


Figure 9.3.3 (a) Time waveform of a DSBSC signal for one cycle of modulation. (b) Spectrum for the signal of (a).

9.4 SSB Generation

Balanced Modulator–Filter Method

Early SSB transmitters used balanced modulator circuits to generate DSBSC signals followed by sideband filters to remove the unwanted sidebands. Such a transmitter is illustrated in Fig. 9.4.1. Initial modulation takes place in the balanced modulator at a low frequency (such as 100 kHz) because of the difficulty of making adequate filters at higher frequencies. The filter is a band-pass filter with a sharp cutoff at each side of the band-pass to obtain satisfactory adjacent sideband rejection. In this case, a single-sideband filter is used, and the carrier oscillator crystal is switched to place the desired sideband in the filter window. Alternatively, two sideband filters (one for each sideband) could be used with a fixed carrier frequency.

The filtered signal is up-converted in a mixer (the second balanced modulator) to the final transmitter frequency and then amplified before being coupled to the antenna. Linear power amplifiers are used to avoid distorting the sideband signal, which might result in regeneration of the second sideband or distortion of the modulated information signal.

The sideband filters are the critical part of this system. Early sideband filters were expensive and did not have the sharp cutoff characteristic required. The integrated ceramic filters available now offer a very inexpensive and effective solution to this problem.

Phasing Method

Figure 9.4.2 shows a different means of obtaining an SSBSC signal. This circuit does not have any sideband filters, and the primary modulation can be done at the transmitting frequency. It relies on phase shifting and cancellation to eliminate the carrier and the unwanted sideband.

Assume sinusoidal signals for both carrier and modulation and that the circuit shown produces the lower side frequency, given by

$$e_{LSF} = E_L \max \cos(\omega_c - \omega_m)t \quad (9.4.1)$$

The standard trigonometric identity for the difference of two angles gives

$$e_{LSF} = E_L \max [\cos \omega_c t \cos \omega_m t + \sin \omega_c t \sin \omega_m t] \quad (9.4.2)$$

but

$$\sin \omega_c t = \cos \left(\omega_c t - \frac{\pi}{2} \right) \quad (9.4.3)$$

$$\sin \omega_m t = \cos \left(\omega_m t - \frac{\pi}{2} \right) \quad (9.4.4)$$

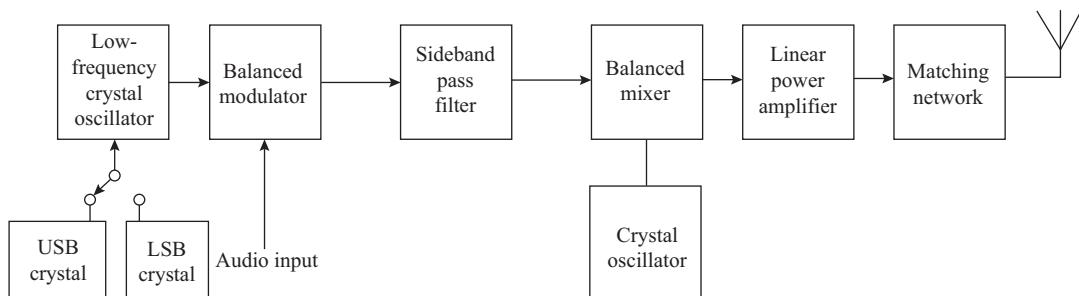


Figure 9.4.1 Single-sideband suppressed carrier transmitter using band-pass filters to eliminate the unwanted sideband.

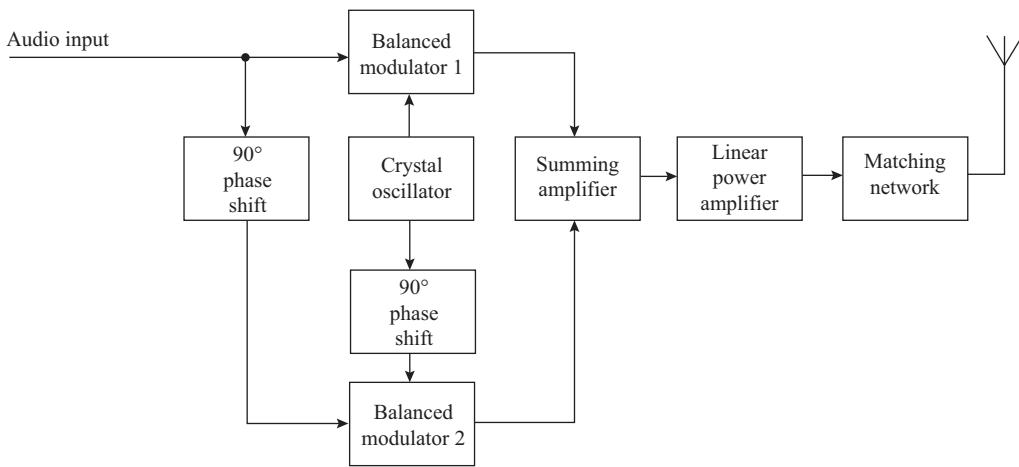


Figure 9.4.2 SSB suppressed carrier transmitter using phase shift to obtain cancellation of sidebands.

Therefore,

$$e_{LSF} = E_{L \max} \left[\cos \omega_c t \cos \omega_m t + \cos \left(\omega_c t - \frac{\pi}{2} \right) \cos \left(\omega_m t - \frac{\pi}{2} \right) \right] \quad (9.4.5)$$

The first term on the right of Eq. (9.4.5) is the result of balanced modulator 1, which multiplies the two unshifted signals. The second term is the result of balanced modulator 2, which multiplies the two signals, each shifted by -90° . The -90° shift for the carrier is easily accomplished by feeding the signal through a controlled current source (transconductance amplifier) into a capacitor. The phase shifting network for the baseband signal must accurately provide a constant 90° phase shift over a wide frequency range. Such circuits are tricky to build.

The carrier signal is canceled out in this circuit by both of the balanced modulators, and the unwanted sidebands cancel at the output of the summing amplifier. It is left as an exercise for the student to expand the outputs of the two balanced modulators into sideband form and show that the cancellation does occur on summing. The two outputs are summed to produce the lower sideband signal.

Examination of the trigonometric identity shows that if the two outputs are subtracted instead of added the upper sideband will result, since

$$\begin{aligned} e_{USF} &= E_{U \max} \cos(\omega_c + \omega_m)t \\ &= E_{U \max} [\cos \omega_c t \cos \omega_m t - \sin \omega_c t \sin \omega_m t] \end{aligned} \quad (9.4.6)$$

While the system is more complex than one using filters, the individual circuits are quite straightforward, and by using integrated-circuit balanced modulators, very little adjustment is required. Only a simple band-pass filter to remove any harmonics is required in the output before application to the final transmitter amplifier.

It should be noted that the modulation signal is usually a broad band of frequencies of varying amplitudes, which the modulator system must not distort. If the capacitor-transconductance amplifier combination causes such distortion, complete cancellation of the unwanted sideband will not occur, and the wanted sideband will have distortions introduced into it. The third method described next eliminates this problem.

Third Method

The third method of generating SSBSC modulation is attributed to D. K. Weaver and was developed during the 1950s. It is similar to the phase shifting method presented previously, but it differs in that the modulating signal is first modulated on a low-frequency subcarrier (including phase shifts), which is then modulated onto the high-frequency carrier.

The circuit connections for generating an LSB signal are shown in Fig. 9.4.3. Modulators *BM1* and *BM2* both have the unshifted modulating signal as inputs. *BM1* also takes the low-frequency subcarrier with a 90° shift introduced in it from the oscillator signal. *BM2* takes the subcarrier signal directly from the oscillator. Assuming unity magnitudes and sinusoidal single-frequency modulation, the output from *BM1* becomes

$$\begin{aligned} e_{BM1} &= \cos\left(\omega_o t + \frac{\pi}{2}\right) \cos \omega_m t \\ &= \frac{1}{2} \left[\cos\left(\omega_o t + \omega_m t + \frac{\pi}{2}\right) + \cos\left(\omega_o t - \omega_m t + \frac{\pi}{2}\right) \right] \end{aligned} \quad (9.4.7)$$

and the output of *BM2* becomes

$$e_{BM2} = \cos \omega_o t \cos \omega_m t + \frac{1}{2} [\cos(\omega_o t + \omega_m t) + \cos(\omega_o t - \omega_m t)] \quad (9.4.8)$$

Low-pass filters with a cutoff frequency set at the subcarrier frequency f_o removes the sum (the first) term from each of the above signals, leaving only the second (difference) terms as inputs to *BM3* and *BM4*. These signals are the lower sidebands on f_o . They are identical except that the signal applied to *BM3* is shifted by $+90^\circ$ from that applied to *BM4*. This process eliminates the need to provide a wideband 90° phase shifting network for the baseband signals, as was the case for the phase shifting method.

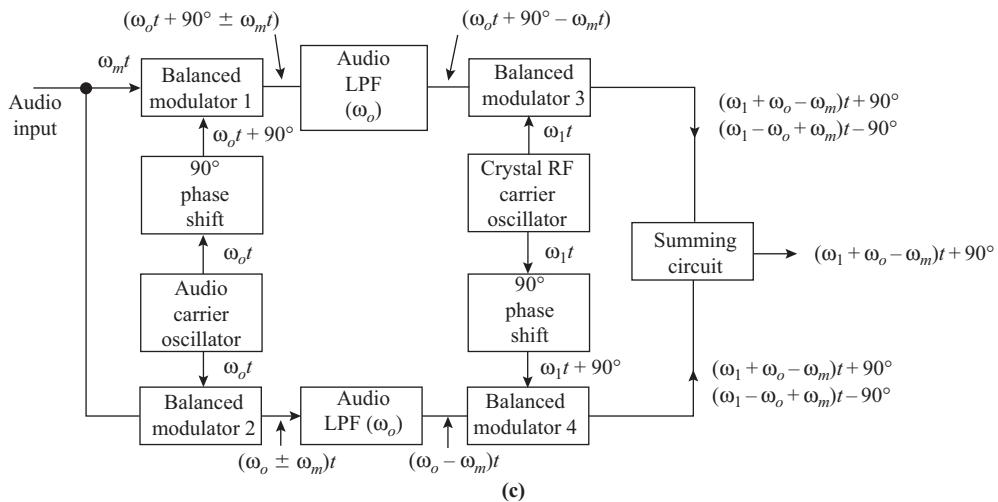


Figure 9.4.3 The “third method” of generating an SSBSC signal.

The high-frequency oscillator signal at f_1 is applied directly to $BM3$, but it is shifted by $+90^\circ$ before being applied to $BM4$. The output from $BM3$ becomes

$$e_{BM3} = \cos \omega_1 t \cos \left((\omega_o - \omega_m)t + \frac{\pi}{2} \right) \quad (9.4.9a)$$

$$e_{BM3} = \frac{1}{2} \left[\cos \left(\omega_1 t + \left((\omega_o - \omega_m)t + \frac{\pi}{2} \right) \right) + \cos \left(\omega_1 t - \left((\omega_o - \omega_m)t + \frac{\pi}{2} \right) \right) \right] \quad (9.4.9b)$$

$$e_{BM3} = \frac{1}{2} \left[\cos \left((\omega_1 + \omega_o)t - \omega_m t + \frac{\pi}{2} \right) + \cos \left((\omega_1 - \omega_o)t + \omega_m t - \frac{\pi}{2} \right) \right] \quad (9.4.9c)$$

and the output of $BM4$ becomes

$$e_{BM4} = \cos \left(\omega_1 t + \frac{\pi}{2} \right) \cos(\omega_o - \omega_m)t \quad (9.4.10a)$$

$$e_{BM4} = \frac{1}{2} \left[\cos \left(\left(\omega_1 t + \frac{\pi}{2} \right) + (\omega_o - \omega_m)t \right) + \cos \left(\left(\omega_1 t + \frac{\pi}{2} \right) - (\omega_o - \omega_m)t \right) \right] \quad (9.4.10b)$$

$$e_{BM4} = \frac{1}{2} \left[\cos \left((\omega_1 + \omega_o)t - \omega_m t + \frac{\pi}{2} \right) + \cos \left((\omega_1 - \omega_o)t + \omega_m t + \frac{\pi}{2} \right) \right] \quad (9.4.10c)$$

The first terms in Eqs. (9.4.9c) and (9.4.10c) are identical lower sidebands on an offset carrier frequency $f_c = f_1 + f_o$. The second terms are the upper sidebands on an offset carrier at $f_c = f_1 - f_o$, but are 180° out of phase with each other. The oscillator frequency f_1 must be adjusted so that the output carrier frequency f_c and the desired sideband fall in the correct position in the output frequency spectrum.

Usually, f_o is chosen to fall at the midpoint of the modulating signal baseband, so that $f_o = W/2$. The result is that both the USB and LSB spectrums are centered on f_1 , with the carrier position f_c for the LSB located at the upper edge of the pass-band and that for the USB at the lower edge. This is illustrated in Fig. 9.4.4.

The outputs from $BM3$ and $BM4$ are added in a summing amplifier to produce the final output. The first two terms add, but the second two cancel, leaving the output as

$$e_{\text{out}} = \cos \left((\omega_1 + \omega_o - \omega_m)t + \frac{\pi}{2} \right) \quad (9.4.11)$$

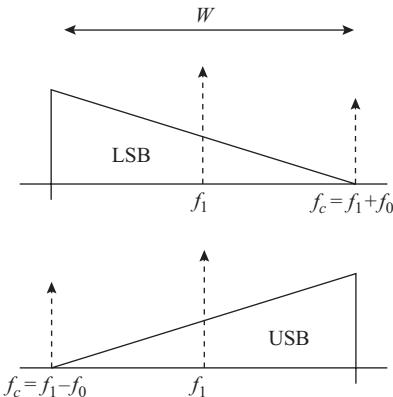


Figure 9.4.4 Output spectra for the “third-method” circuit (a) for LSB and (b) for USB.

This is the lower sideband on the carrier frequency ($f_1 + f_o$). The $+90^\circ$ shift in the output is of no consequence since the original carrier has been eliminated. This signal may be applied to a linear power amplifier and antenna for radiation. Modulation in *BM3* and *BM4* may take place at the final transmission frequency so that no further conversion is needed.

If the output from *BM3* (or *BM4*) is inverted before the input to the adder, the phasing becomes such that the first terms cancel and the second terms add, giving the upper sideband on the carrier frequency ($f_1 - f_o$).

9.5 SSB Reception

Equation 9.2.6 shows that when an SSB signal is multiplied with a synchronous carrier signal the result contains the original modulation signal as one component. Balanced modulator circuits or product demodulator circuits are used for demodulation.

The carrier signal for the demodulator must be locally generated if the signals are true SSB signals with the carrier completely suppressed. This requires extreme stability for the local oscillator signals used for demodulating and for the superheterodyne converters in the receiver front end. Crystal-controlled oscillators are universally used, often in conjunction with frequency synthesizer circuits.

Very good adjacent channel selectivity must be provided since SSB signals are usually packed closely together in the frequency spectrum. Double conversion is often used in SSB receivers. The second mixer oscillator is crystal controlled and may also provide a primary frequency reference for the demodulator oscillator and the first converter oscillator.

Several variations on SSB are used in communications. First, either the upper or lower sideband may be used for a particular channel. In some cases, as for stereo or for telephone multiplexing, the two sidebands of a carrier may be independently modulated with different signals. Next, a reference or *pilot* carrier signal may be transmitted (not necessarily at the same frequency as the actual suppressed SSB carrier). In another method a partial synchronous carrier may be transmitted with the SSB signal (that is, a normal carrier signal, but at much lower amplitude).

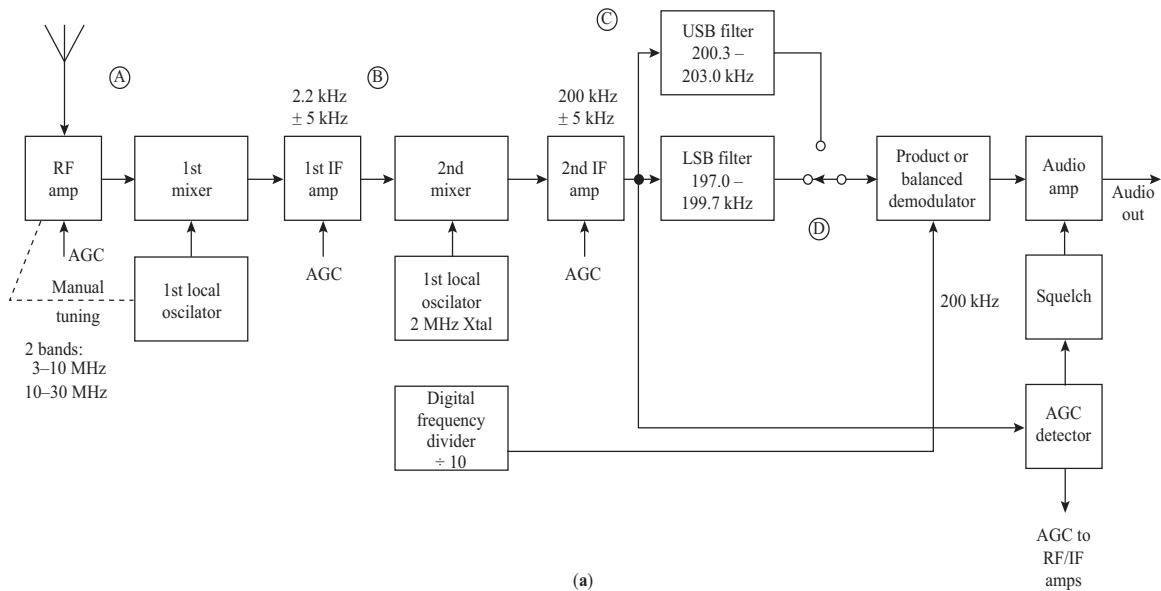
Figure 9.5.1 shows the block diagram of a scanning communications receiver designed for SSB reception in the HF range (3 to 30 MHz). The circuitry is that of a standard double-conversion AM receiver down to the output of the second IF amplifier, except for the local oscillators. The first IF has a bandwidth of 10 kHz centered at 2.2 MHz. The second IF also has a bandwidth of 10 kHz, but centered at 200 kHz, down to the SSB filter inputs. This band-pass is wide enough to pass a normal AM signal (or two adjacent SSB signals), and an envelope detector could be added at this point to allow for AM reception as well as SSB.

The first local oscillator and RF amplifier are manually tuned in two switched bands. The second local oscillator is crystal controlled at 2 MHz. Its output is divided by 10 in a digital counter to provide the 200 kHz carrier signal for the demodulator.

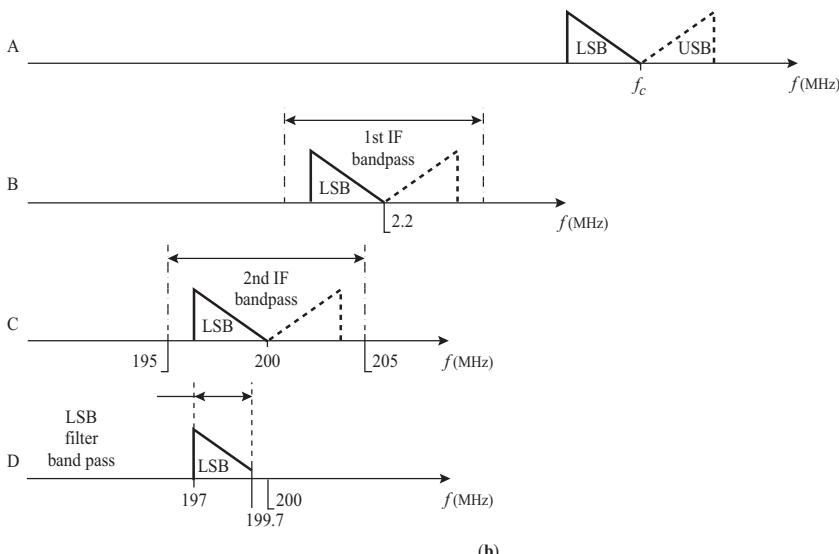
Two SSB filters follow the second IF amplifier. The USB filter passes the IF upper sideband of 200.3 to 203 kHz and rejects the lower sideband. The LSB filter passes the 197- to 199.7-kHz IF lower sideband. The appropriate sideband is selected by a switch that connects the output of the desired filter to the product detector. The RF oscillator must be adjusted to position the incoming SSB IF signal in the IF band-pass so that it exactly falls in the selected SSB filter window and matches the 200-kHz demodulator oscillator to allow distortionless reception.

The output from the detector is passed through a gated audio amplifier that turns off the output to keep the noise down when the signal level drops below a preset threshold. This is called *squelch*. The amplified IF signal (before the detector) is rectified to provide the AGC voltage for the RF and IF amplifiers and for the squelch circuit.

Manually tuned receivers like this one are sometimes difficult to use. The demodulator oscillator is very stable, but the first converter local oscillator must be both stable and tunable. Any variation of this oscillator frequency will shift the SSB signal frequencies relative to the IF band-pass window and to the demodulator carrier frequency. This shift results in the introduction of distortion in the output. Digital frequency synthesizers in integrated circuit form with a crystal-controlled reference provide good stability, easy digital tuning, and low price. One of the largest applications of this technique is in multichannel citizens' band (CB) transceivers.



(a)



(b)

Figure 9.5.1 (a) Single-sideband HF receiver. (b) Spectra in the HF receiver for an LSB signal: (Ⓐ), received RF signal; (Ⓑ), output of first IF amp; (Ⓒ), output of second IF amp; (Ⓓ), output of LSB filter.

9.6 Modified SSB Systems

Pilot Carrier SSB

A pilot carrier SSB system is arranged so that a low-level carrier signal is transmitted with the single sideband in its proper place in the spectrum, but at a much lower level than would be the case for normal amplitude modulation. This *pilot carrier* is used at the receiver to synchronize the local oscillator used for the demodulator, thus eliminating any modulation distortion due to incorrect carrier frequency.

Figure 9.6.1 shows a pilot carrier transmitter and receiver, with the spectra of signals at various points within the system. The audio modulating signal, such as a 4-kHz telephone channel (Ⓐ), is DSBSC modulated in the transmitter at an initial carrier frequency of 100 kHz (Ⓑ). An upper sideband filter passes the sideband between 100 to 104 kHz, but rejects the lower sideband between 96 to 100 kHz. An attenuated sample of the carrier signal is added to the upper sideband (*carrier reinsertion*) to produce the signal at (Ⓒ). This signal is up-converted in a second balanced modulator with a carrier signal at 2.9 MHz. The result is an upper sideband at 3.0 MHz and a lower sideband at 2.8 MHz (Ⓓ). The lower sideband is removed by the band-pass filter to leave only the upper sideband at 3.000 to 3.004 MHz (Ⓔ), which is then amplified and transmitted.

At the receiver, the 3-MHz SSB signal is down-converted to the 100-kHz IF using double conversion so that the sideband remains an upper sideband (Ⓕ). An upper sideband filter passes the 100- to 104-kHz USB to the demodulator. The local oscillator for the demodulator is tuned to 100 kHz, but is part of a phase locked loop system. A sample of the IF is band-pass-filtered to extract the 100-kHz (approximately) pilot carrier signal, which is then used as the reference to lock the local oscillator PLL. The result is a local oscillator signal that may not be exactly 100 kHz, but will be locked to the received signal and will produce proper demodulation of the signal (Ⓖ). The bias signal from the PLL can also be used to provide tuning correction (AFC) on the receiver tuning to center the signal on the IF.

Independent Sideband

For single sideband transmission, the carrier and one sideband are removed from the modulated signal. It is possible to replace the removed sideband with another sideband of information created by modulating a different input signal on the same carrier, giving what is known as *independent sideband* or ISB transmission. Both of the input signals have frequencies in the same audio spectrum range, but in the transmitted signal each signal occupies a different group of frequencies. The process of distributing signals in the frequency spectrum so that they do not overlap is called *frequency division multiplexing*, or FDM.

Figure 9.6.2 shows the block diagram for an ISB transmitter that provides for four multiplexed radio telephone channels modulated on the same carrier. Each telephone channel is limited to frequencies in the 250- to 3000-Hz band as shown in (Ⓐ), somewhat less than is normal for telephone circuits. Four identical band-pass filters on the signal lines provide this limiting. Channels 2 and 3 are applied to the inputs of two balanced modulators that share a common 6.25-kHz subcarrier oscillator to create two lower sideband signals in the 3.25- to 6-kHz band and upper sidebands in the 6.5- to 9.25-kHz band. Lower sideband pass filters remove the upper sidebands. The lower sideband from channel 2 is added to the baseband signal from channel 1, and the lower sideband from channel 3 is added to the baseband signal from channel 4, forming two separate channels, CH 1, 2 and CH 3, 4. The FDM spectra of these two signals are shown in (Ⓑ).

The two signals (1, 2 and 3, 4) are applied to the two inputs of an ISB modulator made of two balanced modulators with a common 100-kHz carrier oscillator. The first modulator produces upper and lower sidebands from channel 1, 2 on 100 kHz, and a USB pass filter strips the lower sideband to leave the upper sideband in the range from 100 to 106 kHz. The second modulator, with an LSB pass filter, produces the lower sideband from CH 3, 4 between 94 and 100 kHz. The two outputs and a low-level sample of the 100-kHz carrier for a pilot are added to produce the complete ISB signal at the 100-kHz carrier frequency. Its spectrum

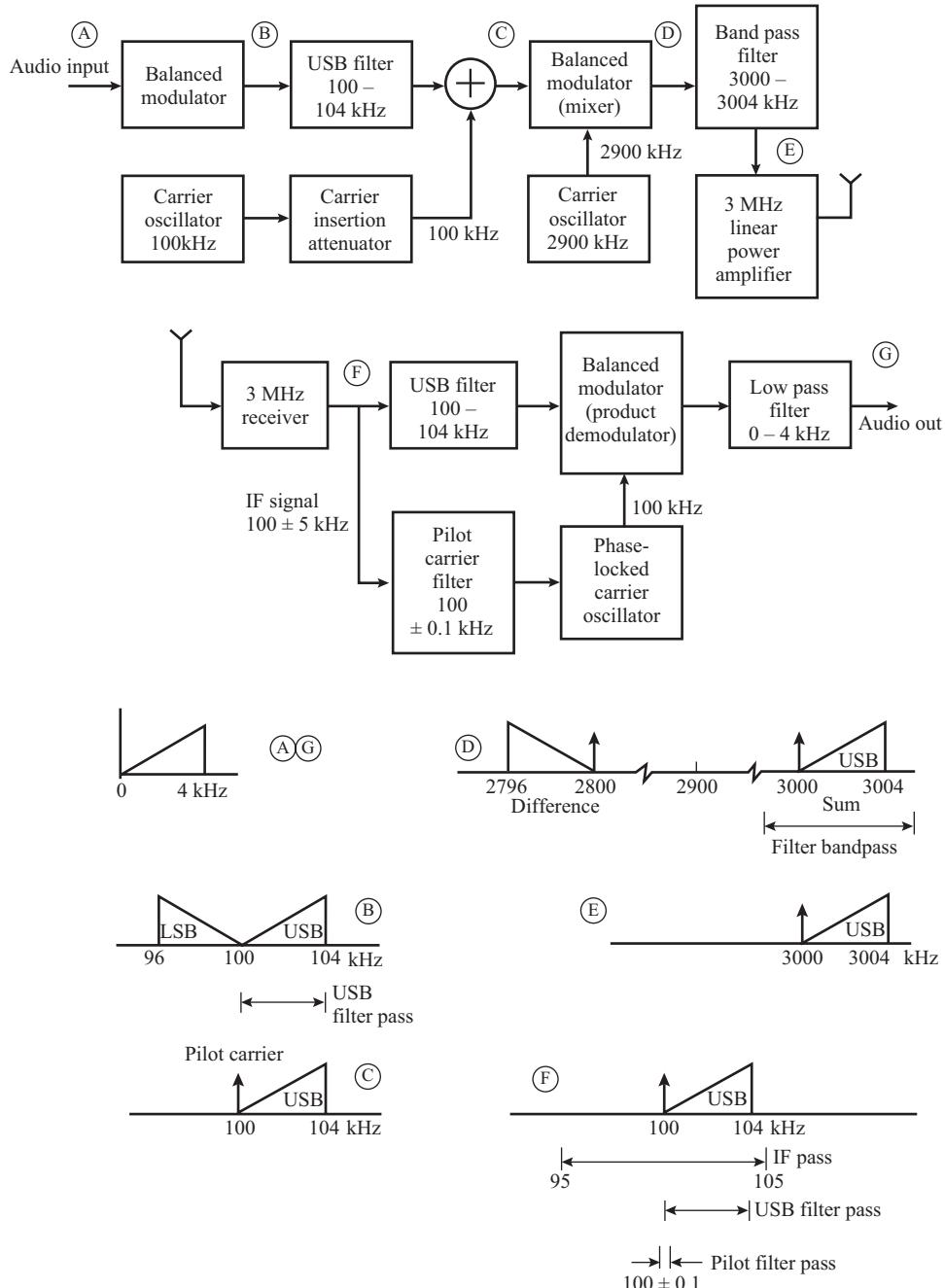


Figure 9.6.1 SSB pilot carrier radio system showing the transmitter, receiver, and signal spectra at various points in the system.

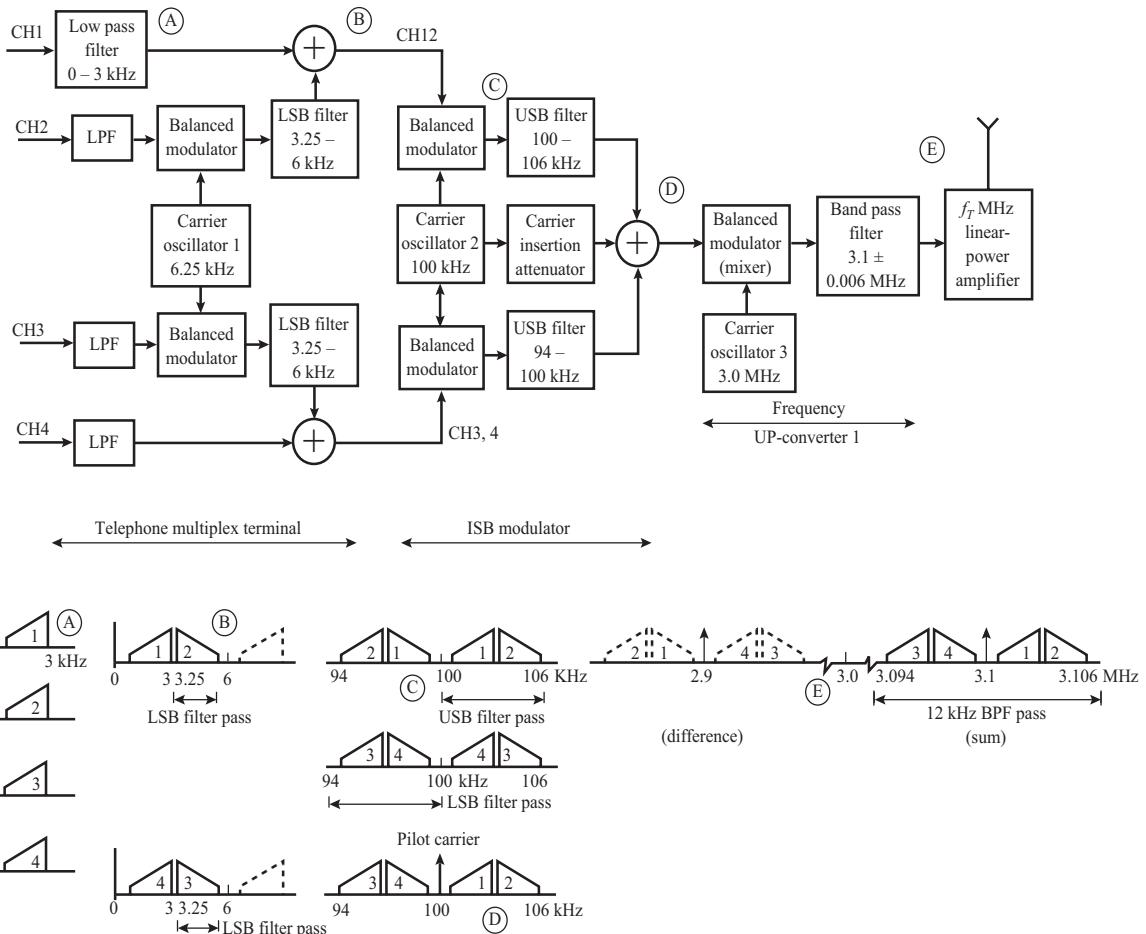


Figure 9.6.2 Four-channel ISB radio telephone transmitter. (a) Transmitter block diagram. (b) Signal spectra at various points.

is shown in (D). This signal has the four input channel signals spread out in frequency, two on either side of the 100-kHz pilot carrier, with none overlapping the others.

This 100-kHz modulated signal could be transmitted directly over a high-frequency cable circuit or, as in this case, it can be raised to a radio frequency for radio transmission. The rest of the transmitter provides an up-conversion to the final transmitting carrier frequency at 3.1 MHz and power amplification to the antenna. The up-converter is a balanced modulator with a 3-MHz carrier oscillator and a band-pass filter centered at 3.1 MHz with a band-pass of 12 kHz to remove the lower sideband (difference frequency component) from the output. A linear power amplifier tuned to the 3.1-MHz transmitting frequency brings the signal up to the antenna power level.

Frequency Division Multiplexing

Frequency division multiplexing is the process of combining several information channels by shifting their signals to different frequency groups within the frequency spectrum so that they can all be transmitted

simultaneously on a common transmission facility. The combination of SSBSC modulation and frequency up-conversion or down-conversion makes this possible.

Radio transmission was the earliest application of FDM, where it became necessary to crowd many signals into the same spectrum of frequencies. The same techniques are used for FDM whether the signal is transmitted by radio or cable. SSB is most often used because of its conservation of the available spectrum.

The International Telegraph and Telephone Consultative Committee or CCITT, based in Geneva, Switzerland, has by international agreement made specific frequency assignments for use in cable and radio telephone systems that use FDM. This agreement assures that systems in various parts of the world use compatible channel frequencies and can communicate with each other. Figure 9.6.3 shows how these channels are grouped for a typical cable system.

Modulation is done in several cascaded stages. First, pregroups of three 4-kHz-wide telephone channels are modulated, each pregroup using the same frequency assignments. The three-channel pregroups are combined four at a time to form groups, each containing 12 channels. Groups are combined five at a time to form 60-channel supergroups, and finally the supergroups are combined 16 at a time to form 960-channel master groups. A total bandwidth of approximately 4.5 MHz is required for each master group. This is more than the indicated 3.84 MHz since spaces are left between supergroups to facilitate separation by filtering.

Radio microwave transmission systems also provide for two or more 6-MHz television channels. A modified master group containing only 12 supergroups (or 720 channels) and requiring only 3 MHz of bandwidth is often used. Two such modified master groups can be stacked in each television channel allocation. A microwave system with a 12-MHz modulating bandwidth capability then could carry either two television channels or 2880 telephone channels.

The pregroup modulator provides three carrier oscillators at 12, 16, and 20 kHz. USB modulation moves the first channel from the 0- to 4-kHz baseband to the first channel slot, 12 to 16 kHz. The second channel goes in the 16- to 20-kHz slot and the third goes in the 20- to 24-kHz slot. All three carriers are suppressed. The three signals are added and passed on to one of the group modulator inputs. A separate pregroup modulator is required for every three channels, so a 960-channel system would require 320 of them.

Each group modulator provides four carriers at 84, 96, 108, and 120 kHz. Four pregroup signals are moved by LSB modulation into four consecutive frequency slots at 60 to 72, 72 to 84, 84 to 96, and 96 to 108 kHz. Because of the LSB modulation, the order of frequencies within each channel slot is reversed from what it was in the baseband signals; that is, a 100-Hz tone would appear at the top edge of the channel and a 3-kHz tone would appear at the bottom. Each carrier frequency is located 12 kHz above the upper edge of the pregroup slot, one pregroup slot away. Again, all four carriers are suppressed. The outputs from the four internal modulators are added and become one input to a supergroup modulator, with frequencies ranging from 60 to 108 kHz.

Each supergroup modulator provides carriers at 420, 468, 516, 564, and 612 kHz. LSB modulation places the five group signals in the slots 312 to 360, 360 to 408, 408 to 456, 456 to 504, and 504 to 522 kHz. Again the order of frequencies within each slot is reversed by the LSB modulation and all five carriers are suppressed. The five outputs are combined to become one input on a master group modulator, each with frequencies from 312 to 552 kHz.

Each master group modulator provides 16 carrier frequencies of 612, none, 1116, 1364, 1612, 1860, 2108, 2356, 2604, 2852, 3100, 3348, 3596, 3844, 4092, and 4340 kHz. The second supergroup is not converted, but directly fed from the group modulator output since its frequencies already fit the assigned slot. Lower sideband modulation moves each of the other 15 supergroups into 240-kHz-wide slots. The first three slots are separated by 12 kHz (three baseband channels) and the remainder by 8 kHz (two channels). All carriers are suppressed, and a single pilot carrier is transmitted at 60 kHz to provide demodulation synchronization. The result is a signal that contains 960 FDM channels each 4 kHz wide in a frequency range from 60 to 4028 kHz.

A different master group layout is often used, especially in North America. It allows the modulation of 10 supergroups in the frequency range from 564 to 3084 kHz, sometimes with the unmodulated slot at

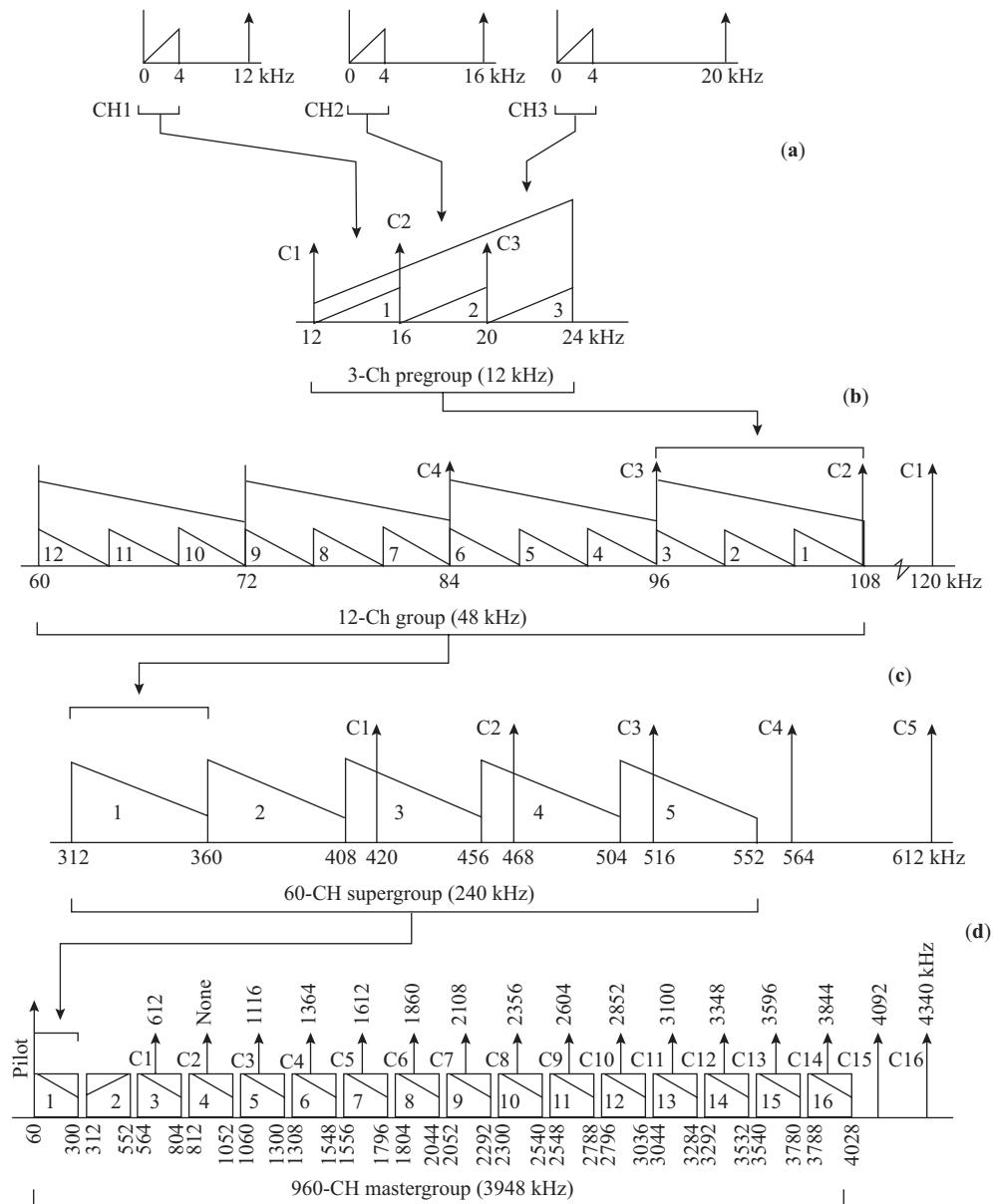


Figure 9.6.3 Cable carrier system frequency allocations. (a) The first modulation forms a three-channel pregroup. (b) Four pregroups form a group of 12 channels. (c) Five groups form a 60-channel supergroup. (d) Sixteen supergroups form a 960-channel master group. (Courtesy of Howard Sams & Co., Inc.)

312 to 552 kHz assigned as an eleventh slot. Separation of the slots is a standard 8 kHz or two channels, except for a 56-kHz break between the sixth and seventh channels. The 10 slots allow the system to carry a total of 600 channels, as opposed to 960, and two such master groups can be included in a standard 6-MHz television slot. The first six supergroup slot assignments are completely compatible with those of the CCITT system.

The spectrum is not completely filled in any FDM system. This is done to allow separation of the supergroups by filtering and to reduce intergroup interference. It also allows for provision of spare channels to simplify maintenance problems. Demodulation at the receiving end proceeds in the reverse order, first separating the master groups, then breaking each master group into its supergroups, then demodulating the groups into pregroups, and finally extracting the individual baseband channel signals.

9.7 Signal-to-Noise Ratio for SSB

Let the received signal be an upper side frequency for sinusoidal modulation, given by

$$e(t) = E_{\max} \cos(\omega_{IF} + \omega_m)t \quad (9.7.1)$$

As in Section 8.14, the noise reaching the detector is narrow-pass bandwidth-limited noise, described by Eq. (4.20.3). In this case, however, the center frequency is given by $\omega_c = 2\pi f_c = 2\pi(IF + W/2) = \omega_{IF} + \pi W$, as shown in Fig. 9.7.1. The noise input to the detector is therefore

$$n_{(t)} = n_I(t) \cos(\omega_{IF} + \pi W)t - n_Q(t) \sin(\omega_{IF} + \pi W)t \quad (9.7.2)$$

The signal plus noise input to the detector is therefore

$$e_{\text{det}}(t) = e(t) + n(t) \quad (9.7.3)$$

Demodulation (coherent detection) takes place as described in Section 9.2, giving the baseband output for the signal alone by Eq. (9.2.7), where k is the detector gain coefficient, as

$$e_{sBB}(t) = \frac{kE_{\max}}{2} \cos \omega_m t = AE_{\max} \cos \omega_m t \quad (9.7.4)$$

Coherent detection applied to the noise described by Eq. (9.7.2) yields a similar baseband output for the noise, in which πW replaces ω_m and E_{\max} is replaced by $n_I(t)$ and $n_Q(t)$ for the respective noise terms. The baseband noise output is therefore

$$e_{nBB}(t) = A [n_I(t) \cos \pi W t - n_Q(t) \sin \pi W t] \quad (9.7.5)$$

Now referring to Eq. (4.20.4), the in-phase and quadrature components in Eq. (9.7.5) can be interpreted as originating from “band-pass” noise with a center frequency $f_c = W/2$ and a band-pass extending from 0 to W , with a power spectral density of kT_s . [Note from Fig. 4.20.3 that this is equivalent to a power spectral density of $2kT_s$ distributed over a bandwidth of $W/2$ as expressed in Eq. (9.7.5).] Now where A_p is

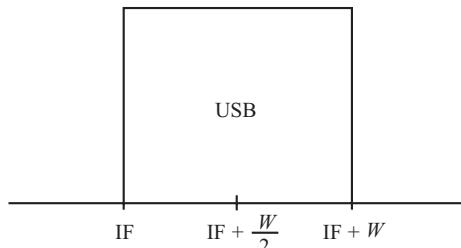


Figure 9.7.1 Band-pass noise in an SSB receiver.

the receiver gain from the antenna terminals to the detector input and A_m is the demodulator multiplier coefficient, which are common to both signal and noise, the available noise power at the detector output is

$$P_{no} = (A_p A_m) k T_s W \quad (9.7.6)$$

The available signal power at the detector output is

$$P_{so} = (A_p A_m) \frac{\left(E_{o \max} / \sqrt{2} \right)^2}{4 R_{\text{out}}} \quad (9.7.7)$$

Hence the output signal-to-noise ratio at the detector output is

$$\begin{aligned} \left(\frac{S}{N} \right)_o &= \frac{P_{so}}{P_{no}} \\ &= \frac{E_{o \max}^2}{8 R_{\text{out}} k T_s W} \end{aligned} \quad (9.7.8)$$

Since the received signal voltage at the detector input is sinusoidal of maximum value $E_{r \max}$, the available received power is

$$\begin{aligned} P_R &= (A_p) \frac{\left(E_{r \max} / \sqrt{2} \right)^2}{4 R_s} \\ &= (A_p) \frac{E_{r \max}^2}{8 R_s} \end{aligned} \quad (9.7.9)$$

The noise spectral density is $k T_s$ over an IF bandwidth $B_{IF} = W$ (which is the same as the baseband bandwidth), so the available noise power at the input to the detector is, from Eq. (8.14.11),

$$P_{n \text{REF}} = (A_p) k T_s W \quad (9.7.10)$$

Hence the reference signal-to-noise ratio (as defined in Section 8.14) is

$$\begin{aligned} \left(\frac{S}{N} \right)_{\text{REF}} &= \frac{P_R}{P_{n \text{REF}}} \\ &= \frac{E_{r \max}^2}{8 R_s k T_s W} \end{aligned} \quad (9.7.11)$$

The figure of merit as introduced in Eq. (8.14.14) is for the SSB case

$$\begin{aligned} R_{\text{SSB}} &= \frac{(S/N)_o}{(S/N)_{\text{REF}}} \\ &= \frac{R_s}{R_{\text{out}}} \end{aligned} \quad (9.7.12)$$

This is the same as that for DSBSC as given by Eq. (8.14.23). However, this is accomplished in the SSB case with a bandwidth reduction of half.

9.8 Companded Single Sideband

Companding is a technique that is used extensively in the transmission of speech signals to reduce the effects of channel noise. The speech signal is *compressed* in volume range relative to a fixed level (typically at -10 dBm). The compression ratio is typically 1 to 2 applied to the decibel difference between the reference level and the average signal level. At the output of the channel, the signal is expanded by the same ratio referred to the same fixed level. The term *compander* is used to describe the unit that does the compression and expansion. This term is coined from *compressor-expander*.

The main advantage of companding is that it reduces the “idle” noise on the channel, which allows for an increase in the total number of channels on a multiplexed carrier system. *Idle noise* is the background noise heard during the quiet intervals between speech bursts. This noise is overridden by the signal during the speech bursts.

The noise reduction is the result of the expander action at the output. To see this, let P_{ns} be the idle noise power from the source. This is increased by factor x in the compressor so that the noise power presented to the channel input is xP_{ns} . At the channel output, before expansion, the total noise power is $(xP_{ns} + P_{nch})$ where P_{nch} is the noise added by the channel. Expansion reduces the total noise by the factor x , so the expanded output noise is

$$P_{n \text{ out}} = \frac{xP_{ns} + P_{nch}}{x} = P_{ns} + \frac{P_{nch}}{x} \quad (9.8.1)$$

The channel noise P_{nch} is the combined result of channel thermal noise, intermodulation distortion, and adjacent and co-channel interference. All these components except the thermal noise increase with the number of channels in a multiplexed carrier system. Hence, for a given level of performance, the noise contribution from additional channels can be offset by the companding action. In other words, the use of companding allows the number of channels in a multiplexed system to be increased. For satellite communications, companding allows an increase in the number of multiple access channels on the system.

PROBLEMS

- 9.1.** Use trigonometric identities to verify Eq. (9.2.6).
- 9.2.** Explain in words how the information signal is recovered from an SSB carrier by a demodulator circuit.
- 9.3.** Develop Eq. (9.2.6) for demodulating a USF signal f_m on a carrier at f_c .
- 9.4.** Modulating frequencies of 0.5, 2.0 and 2.5 kHz are applied to an SSB modulator using a carrier frequency of 100 kHz. (a) Sketch the spectrum of the output signal, noting the positions of the three modulating frequencies, if the modulator is set for USB modulation. (b) Repeat (a) for LSB modulation. (c) Comment on the order of the modulating signal frequencies in the two sidebands.
- 9.5.** The USB signal in Problem 9.4(a) is demodulated by a synchronous detector using a local oscillator set to 99 kHz. (a) Find the frequencies of the three demodulated components and sketch the spectrum of the demodulator output showing their positions. (b) Comment on the quality of the output signal compared to that of the original modulating signal.

- 9.6.** The LSB signal of Problem 9.4(b) is demodulated by a detector whose local oscillator is at 97 kHz. Repeat Problem 9.5 for this case.
- 9.7.** The balanced modulator of Fig. 9.3.1 uses FETs with $IDSS = 10 \text{ mA}$ and $V_p = -12 \text{ V}$. (a) Find the two coefficients a and b for the modulator. (b) If the input signal is 10 mV (peak sine) and the oscillator signal is 1 V (peak sine), find the magnitude of the DSB signal component of the transformer primary effective current.
- 9.8.** The N -channel FETs in Fig. 9.3.1 have a characteristic given by $IDSS = 10 \text{ mA}$ and $V_p = -0.8 \text{ V}$, and the impedance presented by the primary of the output transformer is $1 \text{ k}\Omega$ at the operating frequencies. A signal of 0.01 V (rms sine) at 1 kHz is modulated on a 0.05-V (rms sine), 500-kHz carrier. Find the magnitude and frequency of each component of the net primary voltage.
- 9.9.** Describe how the doubly balanced diode ring modulator of Fig. 9.3.2 acts to produce a DSBSC signal.
- 9.10.** Use the Fourier expansion for a unity-magnitude square wave [Eq. (2.7.1)] to show that both the baseband modulating frequencies and the carrier frequency are suppressed in the output of the doubly balanced ring modulator.
- 9.11.** The modulator of Fig. 5.10.5 feeds a load resistance of $22\text{k}\Omega$ and has a gain adjusting resistor of $1.2\text{k}\Omega$. The square-wave carrier source saturates the switching transistors at 2 MHz, and a sinusoidal modulating signal of 5mV rms at 4 KHz is applied. Find the rms magnitude and frequency for each component appearing in the load voltage.
- 9.12.** A telephone channel requires a 4-kHz bandwidth. Specify the upper and lower cutoff frequencies for the sideband filter on the modulator output if it is to use the upper sideband on carrier frequencies of (a) 12kHz, (b) 16 kHz, and (c) 20 kHz.
- 9.13.** The modulator in Fig. 9.4.1 uses a sideband filter with a band pass of 100 to 104 kHz and modulates two signals at $f_1 = 1 \text{ kHz}$ and $f_2 = 3 \text{ kHz}$.
- For a carrier at 100 kHz, sketch the spectrum of the signal presented to the mixer input. Note the positions of the two modulating signals and the suppressed carrier. Is it a USB or an LSB signal?
 - Using the same filter, how could the modulator be changed to transmit the “other” sideband?
 - Sketch the spectrum of the mixer input signal for part (b), noting the positions of the modulating signals and the suppressed carrier. Is it a USB or an LSB signal?
- 9.14.** For the phase shift SSB modulator of Fig. 9.4.2, assuming single frequency modulation at f_m and a carrier at f_c , (a) develop expressions using Eq. (9.4.2) for the output voltages of the two modulators e_{o1} and e_{o2} . (b) Show that the output summation produces only the lower side frequency signal.
- 9.15.** If the two outputs in Problem 9.14(a) are subtracted instead of added, show that the difference output produces only the upper side frequency signal.
- 9.16.** Redraw Fig. 9.4.2 for the phase shift modulator up to the output of the summer, but rearrange it to produce a USB signal.
- 9.17.** If the carrier signal inputs to $BM3$ and $BM4$ are interchanged in Fig. 9.4.3, show that a USB signal is produced instead of the LSB signal.
- 9.18.** The third-method system of Fig. 9.4.3 uses an audio carrier frequency of 2.5 kHz and a radio carrier frequency of 250 kHz. Audio signals of 500 and 4000 Hz are modulated simultaneously.
 - Find the frequencies all components present in the output signal. Also find the carrier position frequency and the two band edge frequencies.
 - Sketch the spectrum and show all the frequencies in part (a).
- 9.19.** The receiver of Fig. 9.5.1 receives a USB signal containing the original modulating frequencies of 500 and 2000 Hz. It is mistuned so that the signal is 1 kHz lower than it should be. The USB filter is switched in. What audio frequencies are present in the demodulated output?

- 9.20.** The receiver of Fig. 9.5.1 receives a USB signal containing the original modulating frequencies of 500 and 2000 Hz, but this time it is mistuned 3 kHz low and also the LSB filter is switched in.
(a) What audio frequencies are present in the demodulated output?
(b) Compare the output spectrum to the original modulating spectrum and comment on the resulting signal quality.
- 9.21.** An SSB receiver has an equivalent noise temperature of 500Ω and receives a $7\text{-}\mu\text{V}$ rms sinusoidal signal in a 50Ω input resistance. It has a gain to the detector input of 106 dB and a detector gain of 2, with an output resistance of 100Ω . The baseband band-pass is 10 kHz. (a) Find the output noise power, signal power, and signal-to-noise ratio (dB). (b) Find the reference noise power, signal power, and S/N (dB). (c) Find the detector figure of merit.
- 9.22.** Explain the difference between the output S/N ratio and the reference S/N ratio of an SSB receiver.
- 9.23.** Generate SSB using the MATLAB *modulate(.)* command
- 9.24.** The telephone quality speech signals have a bandwidth of 3kHz. Obtain the saving in bandwidth when SSB is used along with FDMA when 100 speech signals are transmitted over a medium, in comparison to AM-DSB.
- 9.25.** Using the *fft(.)* function in MATLAB, show that the spectrum of SSB modulated signal contains only one side-band.
- 9.26.** Discuss the analysis of SSB signals using *Hilbert transforms*.
- 9.27.** Explain why an SSB receiver should have a nearly rectangular band pass filter(BPF) with bandwidth $B_T = W$.
- 9.28.** Bandpass Gaussian noise with variance σ_N^2 is applied to an ideal square law device, producing $y(t) = A_n^2(t)$. Find \bar{y} , \bar{y}^2 , and the PDF of $y(t)$.



Angle Modulation

10.1 INTRODUCTION

In angle modulation, the information signal may be used to vary the carrier frequency, giving rise to *frequency modulation*, or it may be used to vary the angle of phase lead or lag, giving rise to *phase modulation*. Since both frequency and phase are parameters of the carrier angle, which is a function of time, the general term *angle modulation* covers both. Frequency and phase modulation have some very similar properties, but also some marked differences. The relationship between the two will be described in detail in this chapter.

Compared to amplitude modulation, frequency modulation has certain advantages. Mainly, the signal-to-noise ratio can be increased without increasing transmitted power (but at the expense of an increase in frequency bandwidth required); certain forms of interference at the receiver are more easily suppressed; and the modulation process can take place at a low-level power stage in the transmitter, thus avoiding the need for large amounts of modulating power.

10.2 Frequency Modulation

The modulating signal $e_m(t)$ is used to vary the carrier frequency. For example, $e_m(t)$ may be applied as a voltage to a voltage-dependent capacitor, which in turn controls the frequency of an oscillator. (Some modulating circuits are described in Section 10.12). In a well-designed modulator the *change* in carrier frequency will be proportional to the modulating voltage and thus can be represented as $ke_m(t)$, where k is a constant known as the *frequency deviation constant*. The units for k are clearly *hertz/volt* or Hz/V. The instantaneous carrier frequency is therefore equal to

$$f_i(t) = f_c + ke_m(t) \quad (10.2.1)$$

where f_c is the unmodulated carrier frequency.

EXAMPLE 10.2.1

Sketch the instantaneous frequency-time curve for a 90-MHz carrier wave frequency modulated by a 1-kHz square wave that has zero dc component and peak-to-peak voltage of 20 V. The frequency deviation constant is 9 kHz/V.

SOLUTION The peak-to-peak frequency deviation is $20 \times 9 = 180$ kHz, and this is spaced symmetrically about the unmodulated carrier of 100 MHz. The resulting frequency-time curve is shown in Fig. 10.2.1.

As noted previously, the instantaneous frequency may be expressed as $f_i(t) = f_c + k e_m(t)$, and the corresponding instantaneous angular velocity is $\omega_i(t) = 2\pi f_i(t)$. The generation of the modulated carrier can be represented graphically by means of a rotating phasor as shown in Fig. 10.2.2 (a).

The phasor, of constant length $E_{c \text{ max}}$, rotates in a counterclockwise direction at an angular velocity $\omega_i(t) = 2\pi f_i(t)$. The angle turned through in time t is shown as $\theta_i(t)$, where for convenience the positive x -axis is used as the reference axis. The angle $\theta_i(t)$ is found by noting that the angular velocity is the time rate of change of angle, or

$$\frac{d\theta_i(t)}{dt} = \omega_i(t) \quad (10.2.2)$$

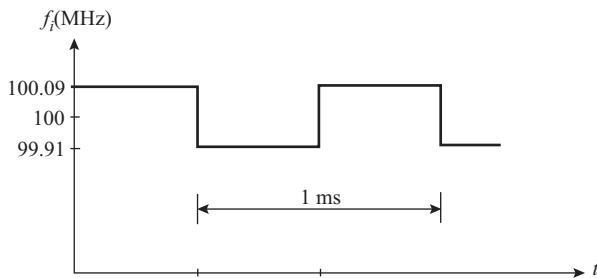


Figure 10.2.1 Instantaneous frequency-time curve for Example 10.2.1.

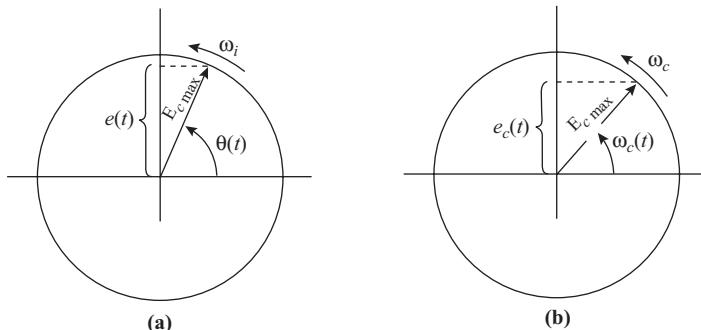


Figure 10.2.2 Rotating phasor representation of a carrier of amplitude $E_{c \text{ max}}$ rotating (a) at instantaneous angular velocity $\omega_i(t)$ and (b) at constant angular velocity ω_c .

and hence

$$\begin{aligned}
 \theta(t) &= \int_0^t \omega_i(t) dt \\
 &= \int_0^t 2\pi(f_c + ke_m(t)) dt \\
 &= 2\pi f_c t + 2\pi k \int_0^t e_m(t) dt
 \end{aligned} \tag{10.2.3}$$

Thus the modulating signal is contained in the angle, in this rather indirect way. Note that the expression for the modulated angle could not have been obtained by simply substituting f_i for f_c in the sine-wave function $E_{c \max} \sin(2\pi f_c t)$, the reason being that this is derived on the basis of a constant frequency, which is not valid for frequency modulation. By setting $e_m(t) = 0$, the unmodulated angle is seen in Fig 10.2.2 (b) to be simply $\theta(t) = 2\pi f_c t$.

EXAMPLE 10.2.2

Sketch $\theta(t)$ as a function of time for a 100-MHz carrier wave frequency modulated by a 1-kHz square wave that has zero dc component and peak-to-peak voltage of 20 V. The frequency deviation constant is 9 kHz/V.

SOLUTION The peak-to-peak frequency deviation is $20 \times 9 = 180$ kHz, and this is symmetrical about the unmodulated carrier of 100 MHz. Thus $\Delta f = \pm 90$ kHz about the carrier, where the plus sign is used for the positive half-cycles and the negative sign for the negative half-cycles of the modulating waveform. Over the positive half-cycles the integral term gives $+\Delta f \cdot t$, and over the negative half-cycles $-\Delta f \cdot t$. Thus the angle is given by $\theta(t) = 2\pi(f_c \pm \Delta f)t$, where the plus sign applies to positive half-cycles and the negative sign to negative half-cycles. The waveforms are sketched in Fig. 10.2.3.

The cosine function representing the carrier wave is given by the projection of the phasor on the x -axis and is seen to be

$$e_c = E_{c \max} \cos \theta(t) \tag{10.2.4}$$

Thus, in the unmodulated case, this reduces to the sinewave $E_{c \max} \cos 2\pi f_c t$, while for the modulated case, the full expression for $\theta(t)$, including the integral term, must be used. For the square-wave modulation in the previous example, the modulated carrier would appear as sketched in Fig. 10.2.4.

10.3 Sinusoidal FM

Many important characteristics of FM can be found from an analysis of sinusoidal modulation. For sinusoidal modulation, $e_m(t) = E_{m \ max} \cos 2\pi f_m t$ and hence

$$\begin{aligned}
 f_i(t) &= f_c + ke_m(t) \\
 &= f_c + kE_{m \ max} \cos 2\pi f_m t \\
 &= f_c + \Delta f \cos 2\pi f_m t
 \end{aligned} \tag{10.3.1}$$

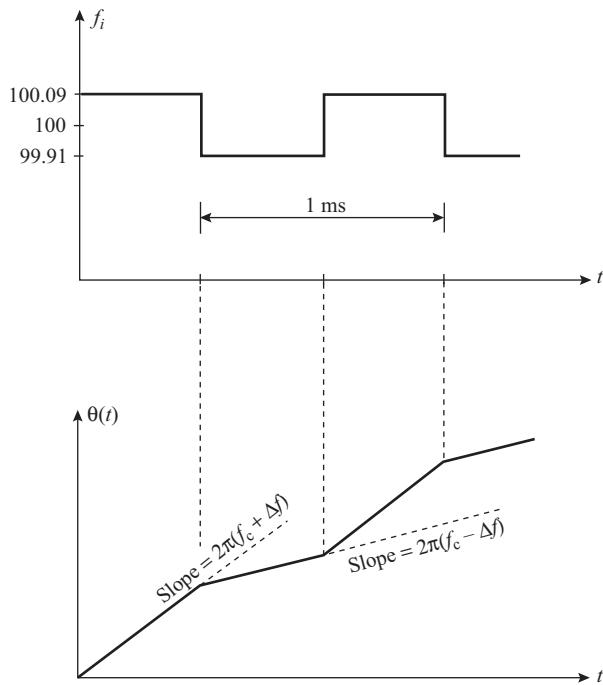


Figure 10.2.3 Solution to Example 10.2.2

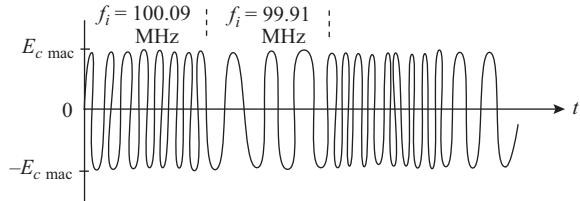


Figure 10.2.4 Square-wave FM.

where the peak *frequency deviation* Δf is proportional to the peak modulating signal and is

$$\Delta f = kE_{m \text{ max}} \quad (10.3.2)$$

The instantaneous frequency as a function of time is sketched in Fig. 10.3.1.

The expression for the sinusoidally modulated carrier therefore become

$$\begin{aligned} e(t) &= E_{c \text{ max}} \cos \theta(t) \\ &= E_{c \text{ max}} \cos \left(2\pi f_{ct} + 2\pi \Delta f \int_0^t \cos 2\pi f_m t dt \right) \\ &= E_{c \text{ max}} \cos \left(2\pi f_{ct} + \frac{\Delta f}{f_m} \sin 2\pi f_m t \right) \end{aligned} \quad (10.3.3)$$

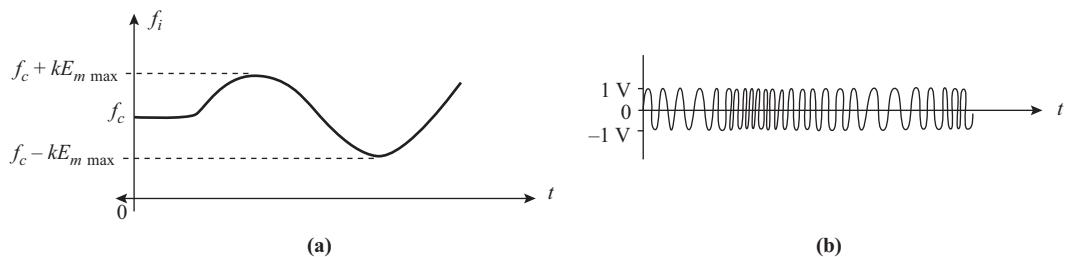


Figure 10.3.1 Instantaneous frequency–time curve for a sinusoidally frequency modulated wave.

The modulation index for FM, usually denoted by β , is defined as

$$\beta = \frac{\Delta f}{f_m} \quad (10.3.4)$$

and hence the equation for the sinusoidally modulated wave becomes

$$e(t) = E_{c \max} \cos(2\pi f_c t + \beta \sin 2\pi f_m t) \quad (10.3.5)$$

EXAMPLE 10.3.1

Determine the modulation index, and plot the sinusoidal FM wave for which $E_{c \max} = 10 \text{ V}$, $E_{m \max} = 3 \text{ V}$, $k = 2000 \text{ Hz/V}$, $f_m = 1 \text{ kHz}$, and $f_c = 20 \text{ kHz}$. On the same set of axes, plot the modulating function. The plot should extend over two cycles of the modulating function.

SOLUTION The peak deviation is $\Delta f = 2000 \times 3 = 6000 \text{ Hz}$. The modulation index is $\beta = 6 \text{ kHz/kHz} = 6$. The functions to be plotted are $e_m(t) = 3 \cos 2\pi 10^3 t$ and $e(t) = 10 \cos(4\pi 10^4 t + 6 \sin 2\pi 10^3 t)$ over a range $0 \leq t \leq 2 \text{ ms}$. The graphs, obtained using Mathcad, are shown in Fig. 10.3.2.

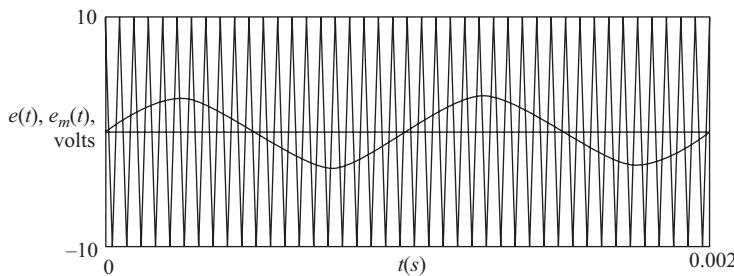


Figure 10.3.2 Solution to Example 10.3.1.

10.4 Frequency Spectrum for Sinusoidal FM

Equation (10.3.5) may be analyzed by Fourier methods in order to obtain the spectrum. The actual analysis is quite involved and only the results will be presented here. The trigonometric series contains a carrier term $J_0(\beta)E_{c \max} \cos \omega_c t$, a first pair of side frequencies $J_1(\beta)E_{c \max} \cos (\omega_c \pm \omega_m)t$, a second pair of side frequencies $J_2(\beta)E_{c \max} \cos (\omega_c \pm 2\omega_m)t$, a third pair of side frequencies $J_3(\beta)E_{c \max} \cos (\omega_c \pm 3\omega_m)t$, and so

on. The amplitude coefficients $J_n(\beta)$ are known as *Bessel functions of the first kind of order n*. Values for these functions are available in both tabular and graphical form and are also available as built-in functions in programs for calculators and computers (such as Mathcad). From the point of view of applications here, the Bessel function gives the amplitude of the carrier ($n = 0$) and side frequencies ($n = 1, 2, 3, \dots$). Some values are shown in Table 10.4.1, where for convenience $E_c \max$ is set equal to unity. The graphs of the carrier and the first three side frequencies are shown in Fig. 10.4.1 for values of β up to 10.

Table 10.4.1 Bessel Functions for a Sinusoidally Frequency-modulated Carrier of Unmodulated Amplitude, 1.0 V (Amplitude Moduli Less than $|0.01|$ not shown.)

Modulation Index β	Carrier J_0	Side Frequencies											
		1st J_1	2nd J_2	3rd J_3	4th J_4	5th J_5	6th J_6	7th J_7	8th J_8	9th J_9	10th J_{10}	11th J_{11}	12th J_{12}
0.25	0.98	0.12	0.01										
0.5	0.94	0.24	0.03										
1.0	0.77	0.44	0.11	0.02									
1.5	0.51	0.56	0.23	0.06	0.01								
2.0	0.22	0.58	0.35	0.13	0.03	0.01							
2.4	0	0.52	0.43	0.20	0.06								
3.0	-0.26	0.34	0.49	0.31	0.13	0.04	0.01						
4.0	-0.40	-0.07	0.36	0.43	0.28	0.13	0.05	0.02					
5.0	-0.18	-0.33	0.05	0.36	0.39	0.26	0.13	0.05	0.02	0.01			
5.5	0	-0.34	-0.12	0.26	0.40	0.32	0.19	0.09	0.03	0.01			
6.0	0.15	-0.28	-0.24	0.11	0.36	0.36	0.25	0.13	0.06	0.02	0.01		
7.0	0.30	0	-0.30	-0.17	0.16	0.35	0.34	0.23	0.13	0.06	0.02	0.01	
8.0	0.17	0.23	-0.11	-0.29	-0.10	0.19	0.34	0.32	0.22	0.13	0.06	0.03	0.01
8.65	0	0.27	0.06	-0.24	-0.23	0.03	0.26	0.34	0.28	0.18	0.10	0.05	0.02

As an example of the use of Table 10.4.1, it can be seen that, for $\beta = 0.05$, the spectral components are

$$\text{Carrier } (f_c)$$

$$J_0(0.5) = 0.94$$

$$\text{First-order side frequencies } (f_c \pm f_m)$$

$$J_1(0.5) = 0.24$$

$$\text{Second-order side frequencies } (f_c \pm 2f_m)$$

$$J_2(0.5) = 0.03$$

The fact that the spectrum component at the carrier frequency decreases in amplitude does *not* mean that the carrier wave is amplitude modulated. The carrier wave is the sum of all the components in the spectrum, and these add up to give a constant amplitude carrier as shown in Fig. 10.3.2. The distinction is that the modulated carrier is not a sine wave, whereas the spectrum component at carrier frequency is. (All spectrum components are either sine or cosine waves.) It will be noted from Table 10.4.1 that amplitudes can be negative in some instances. It will also be seen that for certain values of β (2.4, 5.5, 8.65, and higher values not shown), the carrier amplitude goes to zero. This serves to emphasize the point that it is the sinusoidal component of the spectrum at carrier frequency, *not* the modulated carrier, that goes to zero and that varies from positive to negative peak (1 V in this case) as the frequency varies.

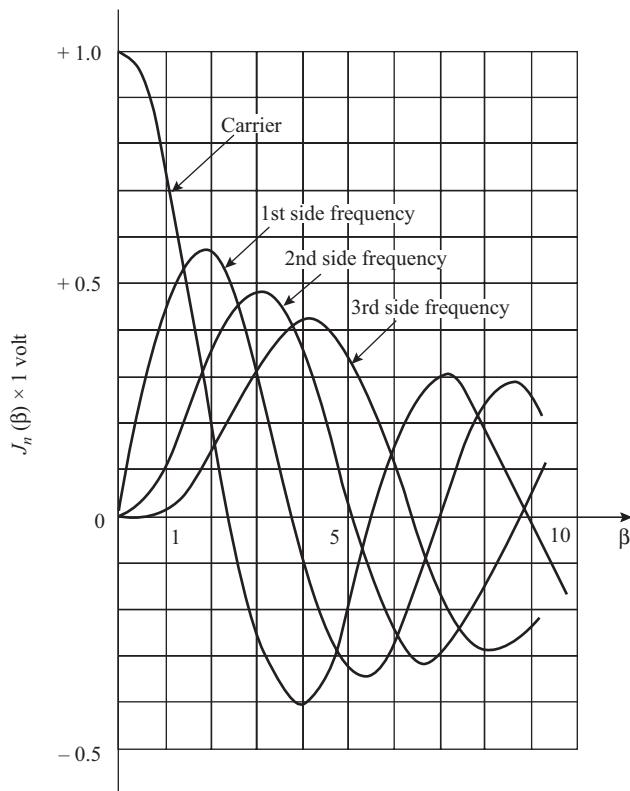


Figure 10.4.1 Graphs of the carrier amplitude and the first three side frequencies for a sinusoidally frequency modulated carrier ($E_{c \max} = 1\text{V}$).

The spectra for various values of β are shown in Fig. 10.4.2(a), (b), and (c). In each case the spectral lines are spaced by f_m , and the bandwidth occupied by the spectrum is seen to be

$$B_{\text{FM}} = 2nf_m \quad (10.4.1)$$

where n is the highest order of side frequency for which the amplitude is significant. From Table 10.4.1 it can be seen that, where the order of side frequency is greater than $(\beta + 1)$, the amplitude is 5% or less of unmodulated carrier amplitude. Using this as a guide for bandwidth requirements, Eq. (10.4.1) can be written as

$$B_{\text{FM}} = 2(\beta + 1)f_m \quad (10.4.2)$$

or, substituting for β from Eq. (10.3.4)

$$B_{\text{FM}} = 2(\Delta f + f_m) \quad (10.4.3)$$

To illustrate the significance of this, three examples will be considered:

1. $\Delta f = 75 \text{ kHz}, f_m = 0.1 \text{ kHz}$

$$\begin{aligned} B_{\text{FM}} &= 2(75 + 0.1) \\ &= 150 \text{ kHz} \end{aligned}$$

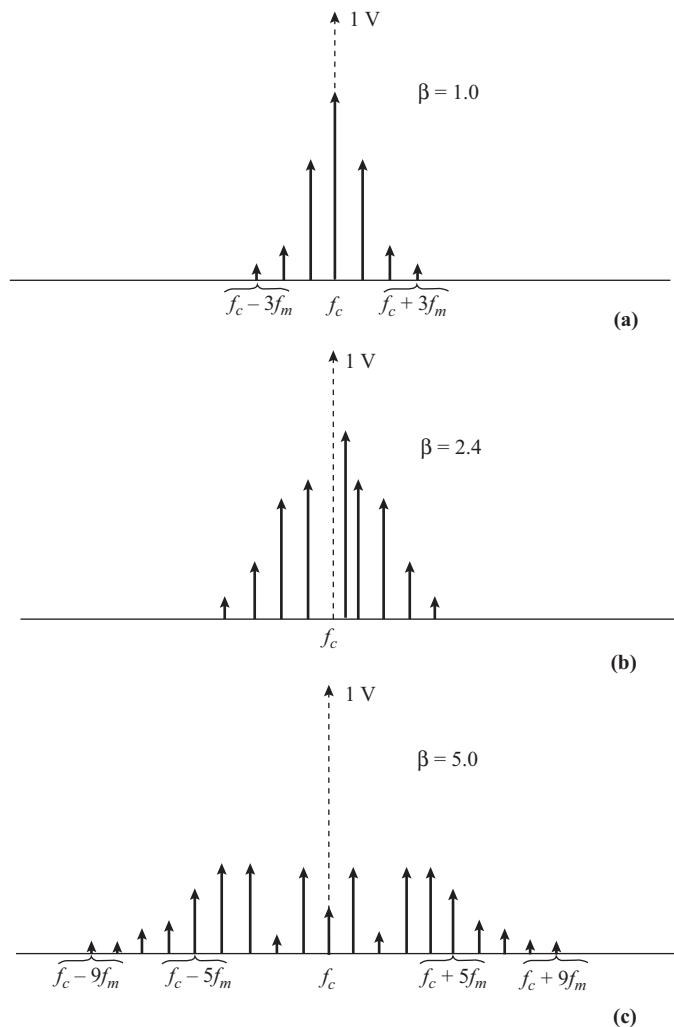


Figure 10.4.2 Spectra for sinusoidal FM with (a) $\beta = 1.0$, (b) $\beta = 2.4$ (note missing carrier), and (c) $\beta = 5.0$.

2. $\Delta f = 75 \text{ kHz}$, $f_m = 1.0 \text{ kHz}$

$$\begin{aligned} B_{\text{FM}} &= 2(75 + 1) \\ &= 152 \text{ kHz} \end{aligned}$$

3. $\Delta f = 75 \text{ kHz}$, $f_m = 10 \text{ kHz}$

$$\begin{aligned} B_{\text{FM}} &= 2(75 + 10) \\ &= 170 \text{ kHz} \end{aligned}$$

Thus, although the modulating frequency changes from 0.1 to 9 kHz, or by a factor of 100 : 1, the bandwidth occupied by the spectrum alters very little, from 150 to 170 kHz. These examples illustrate why frequency modulation is sometimes referred to as a constant-bandwidth system.

10.5 Average Power in Sinusoidal FM

The peak voltages of the spectrum components are given by $E_{n \text{ max}} = J_n(\beta)E_c \text{ max}$. Since the rms values denoted by E_n and E_c are proportional to the peak values, these are also related as

$$E_n = J_n(\beta)E_c \quad (10.5.1)$$

For a fixed load resistance R the average power of any one spectral component is $P_n = E_n^2/R$. The total average power is the sum of all such components. Noting that there is only one carrier component and a pair of components for each side frequency, the total average power is

$$P_T = P_o + 2(P_1 + P_2 + \dots) \quad (10.5.2)$$

In terms of the rms voltages this becomes

$$P_T = \frac{E_o^2}{R} + \frac{2}{R}(E_1^2 + E_2^2 + \dots) \quad (10.5.3)$$

In terms of the unmodulated carrier and the Bessel function coefficients, this is

$$\begin{aligned} P_T &= \frac{E_c^2 J_o^2(\beta)}{R} + \frac{2E_c^2}{R}(J_1^2(\beta) + J_2^2(\beta) + \dots) \\ &= \frac{E_c^2}{R}[J_o^2(\beta) + 2(J_1^2(\beta) + J_2^2(\beta) + \dots)] \\ &= P_c[J_o^2(\beta) + 2(J_1^2(\beta) + J_2^2(\beta) + \dots)] \end{aligned} \quad (10.5.4)$$

Here, the unmodulated power is $P_c = E_c^2/R$. A property of the Bessel functions is that the sum $[J_o^2(\beta) + 2(J_1^2(\beta) + J_2^2(\beta) + \dots)] = 1$, so the total average power is equal to the unmodulated carrier power. This result might have been expected because the amplitude of the wave remains constant whether or not it is modulated. In effect, when modulation is applied, the total power that was originally in the carrier is redistributed between all the components of the spectrum. As previously pointed out, at certain values of β the carrier component goes to zero, which means that in these instances the power is carried by the side frequencies only.

EXAMPLE 10.5.1

A 15-W unmodulated carrier is frequency modulated with a sinusoidal signal such that the peak frequency deviation is 6 kHz. The frequency of the modulating signal is 1 kHz. Calculate the average power output by summing the powers for all the side-frequency components.

SOLUTION The total average power output P is 15 W modulated. To check that this is also the value obtained from the sum of the squares of the Bessel functions, from Eq. (10.3.4) we have

$$\beta = \frac{\Delta f}{f_m} = \frac{6}{1} = 6$$

The Bessel function values for $\beta = 6$ are read from Table 10.4.1 and substituted in Eq. (10.5.4) to give

$$\begin{aligned} P_T &= 15[0.15^2 + 2(0.28^2 + 0.24^2 + 0.11^2 + 0.36^2 + 0.36^2 + 0.25^2 + 0.13^2 + 0.06^2 \\ &\quad + 0.02^2 + 0.01^2)] \\ &= 15(1.00) \\ &= 15 \text{ W} \end{aligned}$$

It follows that, since the average power does not change with frequency modulation, the rms voltage and current will also remain constant, at their respective unmodulated values.

10.6 Non-sinusoidal Modulation: Deviation Ratio

In the frequency-modulation process, intermodulation products are formed; that is, beat frequencies occur between the various side frequencies when the modulation signal is other than sinusoidal or cosinusoidal. It is a matter of experience, however, that the bandwidth requirements are determined by the maximum frequency deviation and maximum modulation frequency present in the modulating wave. The ratio of maximum deviation to maximum frequency component is termed the *deviation ratio*, which is defined as

$$D = \frac{\Delta F}{F_m} \quad (10.6.1)$$

where ΔF is the maximum frequency deviation and F_m is the highest frequency component in the modulating signal. The bandwidth is then given by Eq. (10.4.2) on substituting D for β , with the same limitations on accuracy, as

$$\begin{aligned} B_{\max} &= 2(D + 1)F_m \\ &= 2(\Delta F + F_m) \end{aligned} \quad (10.6.2)$$

This is known as *Carson's rule*.

EXAMPLE 10.6.1

Canadian regulations state that for FM broadcast the maximum deviation allowed is 75 kHz and the maximum modulation frequency allowed is 15 kHz. Calculate the maximum bandwidth requirements.

SOLUTION Using Eq. (10.6.2),

$$\begin{aligned} B_{\max} &= 2(\Delta F + F_m) \\ &= 2(75 + 15) \\ &= 180 \text{ kHz} \end{aligned}$$

Examination of Table 10.4.1 shows that side frequencies of 1% amplitude extend up to the ninth side-frequency pair, so Carson's rule underestimates the bandwidth required. For D equal to 5 or greater, a better estimate is given by $B_{\max} = 2(D + 2)F_m$. In this example, this would result in a maximum bandwidth requirement of 210 kHz. The economic constraints on commercial equipment limit the bandwidth capabilities of receivers to about 200 kHz.

10.7 Measurement of Modulation Index for Sinusoidal FM

Commercially available frequency deviation meters are available that enable Δf to be measured for a known value of f_m . The frequency-modulation index is then simply the ratio of these two quantities as given by Eq. (10.3.4). As a further check, the spectrum can be examined with the aid of a spectrum analyzer and the conditions at which the carrier component of the spectrum goes to zero noted. These should occur at modulation indexes of 2.4, 5.5, and so on, as shown in Table 10.4.1.

10.8 Phase Modulation

Referring once again to the expression for an unmodulated carrier, this is

$$e_c(t) = E_{c \max} \cos(\omega_c t + \phi_c) \quad (10.8.1)$$

The phase angle ϕ_c is arbitrary and is included in the general case to show that the reference line for the rotating phasor of Fig. 10.2.2 is arbitrary. Figure 10.8.1(a) shows the situation for $\phi_c = 25^\circ$.

When phase modulation is applied, it has the effect of moving the reference line (circuits for phase modulation are described in Section 10.12), as shown in Fig. 10.8.1(b). Mathematically, the phase modulation may be written as

$$\phi(t) = \phi_c + K e_m(t) \quad (10.8.2)$$

where K is the *phase deviation constant*, analogous to the frequency deviation constant k introduced for frequency modulation. It will be seen that K must have units of radians per volt when ϕ_c is measured in radians. The constant phase angle ϕ_c has no effect on the modulation process, and this term can be dropped without loss of generality. Thus the equation for the phase modulated wave becomes

$$e(t) = E_{c \max} \cos(\omega_c t + K e_m(t)) \quad (10.8.3)$$

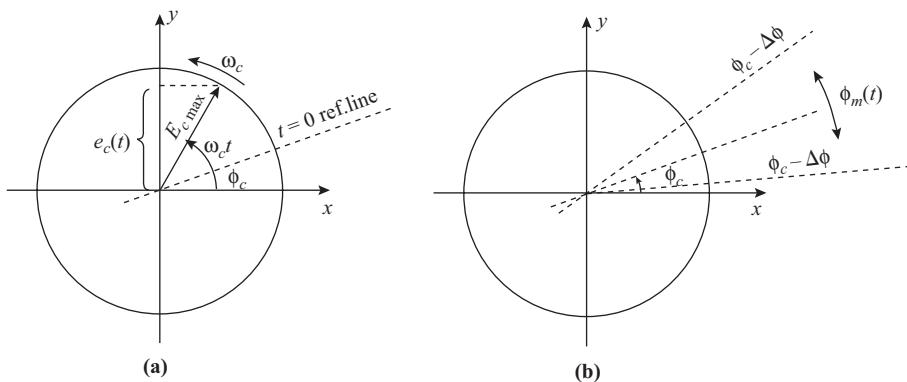


Figure 10.8.1 (a) Rotating phasor representation of a carrier of amplitude $E_{c \max}$ and phase lead $\phi_c = 25^\circ$. (b) Effect of applying phase modulation.

EXAMPLE 10.8.1

A modulating signal given by $e_m(t) = 3 \cos(2\pi 10^3 t - 90^\circ)$ volts is used to phase modulate a carrier for which $E_{c \max} = 10$ V and $f_c = 20$ kHz. The phase deviation constant is $K = 2$ rad/V. Plot the modulated waveform over two cycles of the modulating function.

SOLUTION The phase modulation function is

$$\begin{aligned}\phi_m(t) &= Ke_m(t) \\ &= 2 \times 3 \cos(2\pi 10^3 t - 90^\circ) \\ &= 60 \sin 2\pi 10^3 t\end{aligned}$$

Hence the modulated wave function is

$$e(t) = 10 \cos(4\pi 10^4 t + 6 \sin 2\pi 10^3 t)$$

This is identical to the modulated wave in Example 10.3.1 and hence the graph of Fig. 10.3.2 applies.

10.9 Equivalence between PM and FM

It is seen that for phase modulation the angular term is given by

$$\theta(t) = \omega_c t + Ke_m(t) \quad (10.9.1)$$

Now, the corresponding instantaneous frequency in general is obtained from

$$f_i(t) = \frac{1}{2\pi} \frac{d\theta(t)}{dt} \quad (10.9.2)$$

Hence the equivalent instantaneous frequency for phase modulation is, on differentiating Eq. (10.9.1),

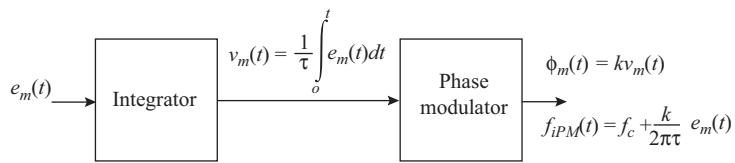
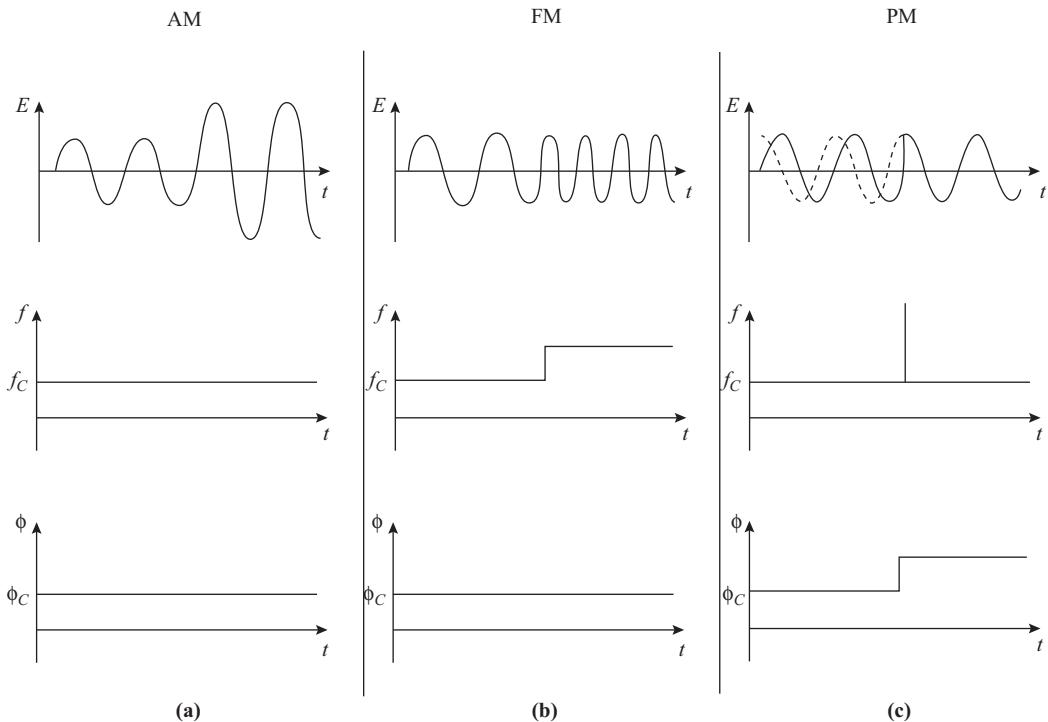
$$f_{iPM}(t) = f_c + \frac{K}{2\pi} \frac{de_m(t)}{dt} \quad (10.9.3)$$

The importance of this equation is that it shows how phase modulation may be used to produce frequency modulation. The differentiation is nullified by passing the modulating signal through an integrator before it is applied to the phase modulator, as shown in Fig. 10.9.1. The time constant of the modulator is shown as τ , so the actual voltage applied to the phase modulator is

$$v_m(t) = \frac{1}{\tau} \int_0^t e_m(t) dt \quad (10.9.4)$$

The equivalent frequency modulation is then

$$\begin{aligned}f_{iPM}(t) &= f_c + \frac{K}{2\pi} \frac{dv_m(t)}{dt} \\ &= f_c + \frac{K}{2\pi\tau} e_m(t)\end{aligned} \quad (10.9.5)$$

**Figure 10.9.1** How FM may be obtained from PM.**Figure 10.9.2** Modulating with a step waveform: (a) AM, (b) FM, (c) PM.

This is identical to the instantaneous frequency expression for frequency modulation, Eq. (10.2.1), with a frequency deviation constant given by $k = K/(2\pi\tau)$. This equivalence between FM and PM has already been illustrated in Example 10.8.1 for sinusoidal modulation. Frequency modulators that utilize the equivalence between FM and PM are described in Section 10.12.

As with frequency modulation, many important characteristics can be found from an analysis of a sinusoidally phase-modulated carrier. However, at this point it is instructive to compare the three methods of modulation, amplitude, frequency, and phase, for a modulating signal that is a step function.

In the case of amplitude modulation (Fig. 10.9.2[a]), the amplitude follows the step change, while the frequency and phase remain constant with time. The amplitude change could be observed, for example on an oscilloscope. With frequency modulation, shown in Fig. 10.9.2(b), the amplitude and phase remain constant while the frequency follows the step change. Again, this change could be observed, for example on a frequency counter.

With phase modulation, the amplitude remains constant while the phase angle follows the step change with time, as shown in Fig. 10.9.2(c). The phase change is measured with reference to what the phase angle would have been with no modulation applied. After the step change in phase, the sinusoidal carrier appears as though it is a continuation of the dashed curve shown on the amplitude-time graph of Fig. 10.9.2(c). Also, from the amplitude-time graph it is seen that the frequency of the wave before the step change is the same as after the step change. However, at the step change in phase, the abrupt displacement of the waveform on the time axis makes it appear as though the frequency undergoes an abrupt change. This is shown by the spike in the frequency-time graph in Fig. 10.9.2(c). A phase meter could be used to measure the change in phase, but this requires the reference waveform and is not as direct as the measurement of amplitude or frequency. The spike change in frequency could be measured directly on a frequency counter. In principle, the apparent change of frequency with phase modulation will occur even where the source frequency of the carrier is held constant, for example by using a crystal oscillator. In practice, it proves to be more difficult to achieve large frequency swings using phase modulation.

10.10 Sinusoidal Phase Modulation

For sinusoidal modulation, $e_m(t) = E_{m \max} \sin 2\pi f_m t$, and hence

$$\begin{aligned} Ke_m(t) &= KE_{m \max} \sin 2\pi f_m t \\ &= \Delta\phi \sin 2\pi f_m t \end{aligned} \quad (10.10.1)$$

where the *peak phase deviation* $\Delta\phi$ is proportional to the peak modulating signal and is

$$\Delta\phi = KE_{m \max} \quad (10.10.2)$$

The sine expression is used for the modulation signal rather than the cosine expression, because this brings out more clearly the equivalence in the spectra for FM and PM, as will be shortly shown. The equation for sinusoidal PM is therefore

$$e(t) = E_{c \max} \cos(\omega_c t + \Delta\phi \sin \omega_m t) \quad (10.10.3)$$

This is identical to Eq. (10.3.5) with $\Delta\phi = \beta$, and therefore the trigonometric expansion will be similar to that for sinusoidal FM, containing a carrier term, and side frequencies at $f_c \pm nf_m$. The amplitudes are also given in terms of Bessel functions of the first kind, $J_n(\Delta\phi)$. In this case the argument is the peak phase deviation $\Delta\phi$, rather than the frequency modulation index β . It follows therefore that the magnitude and extent of the spectrum components for the PM wave will be the same as for the FM wave for which $\Delta\phi$ is numerically equal to β . It also follows that the power relationships developed in Section 10.5 for sinusoidal FM apply to the equivalent sinusoidal PM case.

For analog modulating signals, phase modulation is used chiefly as a stage in the generation of frequency modulation; as previously described. It should be noted that the demodulators in analog FM receivers (even the phase discriminators described in Section 10.14) interpret the received signal as frequency modulation, real or equivalent. The effect this has on the reception of a phase modulated carrier is illustrated in Problem 10.17.

With sinusoidal phase modulation, application of Eq. (10.9.3) gives the equivalent frequency modulation as

$$\begin{aligned} f_{i eq}(t) &= f_c + \Delta\phi f_m \cos \omega_m t \\ &= f_c + \Delta f_{eq} \cos \omega_m t \end{aligned} \quad (10.10.4)$$

The equivalent peak deviation is seen to be

$$\Delta f_{eq} = \Delta\phi \cdot f_m \quad (10.10.5)$$

The other major area of application for phase modulation lies in the digital modulation of carriers.

10.11 Digital Phase Modulation

Phase modulation is very widely used in digital systems. In the simplest system, the voltage levels $\pm V$ representing the binary digits 1 and 0 may be used to multiply the carrier. If the digital signal is represented by $p(t)$ and a balanced modulator is used (see Section 8.9), the modulated signal is essentially a DSBSC wave given by

$$e(t) = Ap(t)E_{c \text{ max}} \cos \omega_c t \quad (10.11.1)$$

where A is a constant of the multiplier. When $p(t) = +V$, then $e(t) = AV E_{c \text{ max}} \cos \omega_c t$. When $p(t) = -V$, the minus sign can be interpreted as a 180° phase shift, and the modulated wave is $e(t) = -VE_{c \text{ max}} \cos (\omega_c t + 180^\circ)$. Thus the phase of the modulated signal shifts between zero and 180° in accordance with the digital modulating signal. This type of modulation is referred to as *binary phase shift keying* (BPSK), the word "keying" being a relic from the time when a Morse key was the most common method of generating a digitally modulated telegraph signal. Digital modulation methods are discussed further in Section 12.9.

10.12 Angle Modulator Circuits

In the study of angle modulator circuits, a difficulty arises in that the carrier frequency is not constant, so steady-state sinusoidal analysis of networks forming part of the circuit is not strictly valid. Another difficulty is that the angle modulation is achieved by varying inductance or capacitance as a function of the modulating signal, so these become time-varying components rather than being constants of the circuit, as assumed for steady-state analysis. These difficulties notwithstanding, a reasonably accurate picture of the principles of operation of angle modulator circuits can be obtained from what is termed *quasi-steady-state analysis*, in which it is assumed that the instantaneous values of inductance, capacitance, and carrier frequency are changing sufficiently slowly for these to be treated as constants. Steady-state analysis is then carried out for the particular instantaneous values at any given instant.

In the analyses a number of nonlinear functions arise, of the form $f(\delta) = (1 + \delta)^n$, where δ represents the fractional change in the variable and $n < 0$. For example, for the varactor diode described in the following section, δ represents the fractional change in bias voltage, and $n = -\alpha$. For the frequency equation, of the form $f = 1/(2\pi \sqrt{LC})$, δ represents the fractional change in inductance or capacitance and $n = -0.5$. For $\delta \ll 1$, the binomial expansion gives $f(\delta) \approx 1 + n\delta$, which shows that the functional variation is approximately linear in relation to the fractional change. Usually, what is required is the *transfer characteristic* showing the variation of frequency (or phase) as a function of modulating voltage. A measure of the linearity of the transfer characteristic is given by the *linear correlation coefficient*. This has a value between zero and unity, the ideal value being unity. The slope of the transfer characteristic gives the frequency (or phase) deviation constant. Both the slope and the linear correlation coefficient can be found using a computational package such as Mathcad, as will be illustrated later.

Varactor Diode Modulators

The varactor diode has already been met with in the study of voltage-controlled oscillators in Section 6.6. In a sense, voltage-controlled oscillators are frequency modulation circuits in which discontinuous (such as step changes) in frequency are usually encountered. In the present section, the use of the varactor diode in generating continuously variable changes in frequency and phase will be examined. The circuit for a Clapp FM oscillator is shown in Fig. 10.12.1.

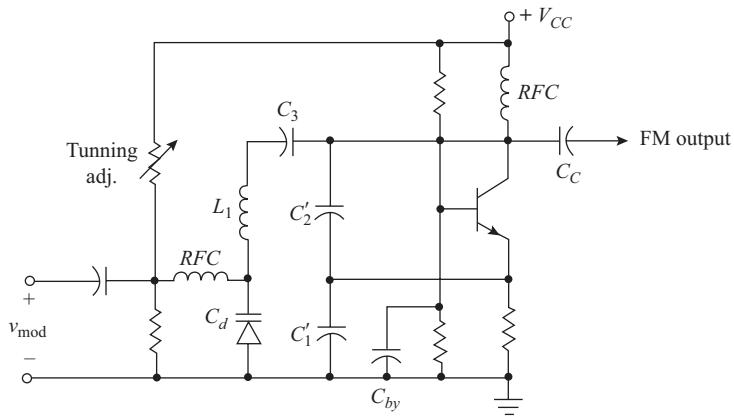


Figure 10.12.1 Clapp FM oscillator utilizing a varactor diode.

In this circuit, which is similar to that of Fig. 6.6.1(a), the fixed bias for the varactor diode is adjusted by means of the variable resistor in the bias line. The radio-frequency choke (RFC) prevents the radio-frequency signal from entering the modulating circuit. Capacitor C_3 alters the effective inductance as explained in Section 6.4 and also acts as a dc blocking capacitor. The varactor diode capacitance acts in series with C_3 .

Figure 10.12.2 provides an illustration of the modulation process. The diode is biased into its reverse-bias region at some fixed bias point $-V_R$. The modulating voltage is superimposed on this, which results in a variation of capacitance about the fixed value C_{d0} . The diode capacitance forms part of the total tuning capacitance C_T of the frequency determining circuit, and so the variation in C_d can be projected onto the C_T/C_d graph as shown. This in turn can be projected onto the C_T/f graph for which the frequency is inversely proportional to the square root of the capacitance. The resulting frequency modulation is then as shown.

An estimate of the frequency deviation constant can be obtained for small-signal conditions using quasi-steady-state analysis. In the present application, the diode voltage consists of a reverse fixed bias $-V_R$ and a time-varying component $v_m(t)$ which is the modulating voltage. Thus the applied diode voltage is a function of the modulating voltage, which will be denoted simply as v_m :

$$V_d(v_m) = -(V_R + v_m) \quad (10.12.1)$$

When this voltage is substituted in Eq. (6.6.1), the diode capacitance as a function of modulating voltage becomes

$$C_d(v_m) = \frac{C_o}{(1 - V_d(v_m)/\psi)^\alpha} \quad (10.12.2)$$

Since the diode capacitance forms part of the total tuning capacitance, this is now also a function of time. For example, if the diode is in series with some fixed capacitance C_{ser} , the total tuning capacitance is

$$C_T(v_m) = \frac{C_{ser}C_d(v_m)}{C_{ser} + C_d(v_m)} \quad (10.12.3)$$

With no modulation applied, the total tuning capacitance will have some value C_{TO} corresponding to the unmodulated frequency f_o . With modulation applied, the frequency is therefore

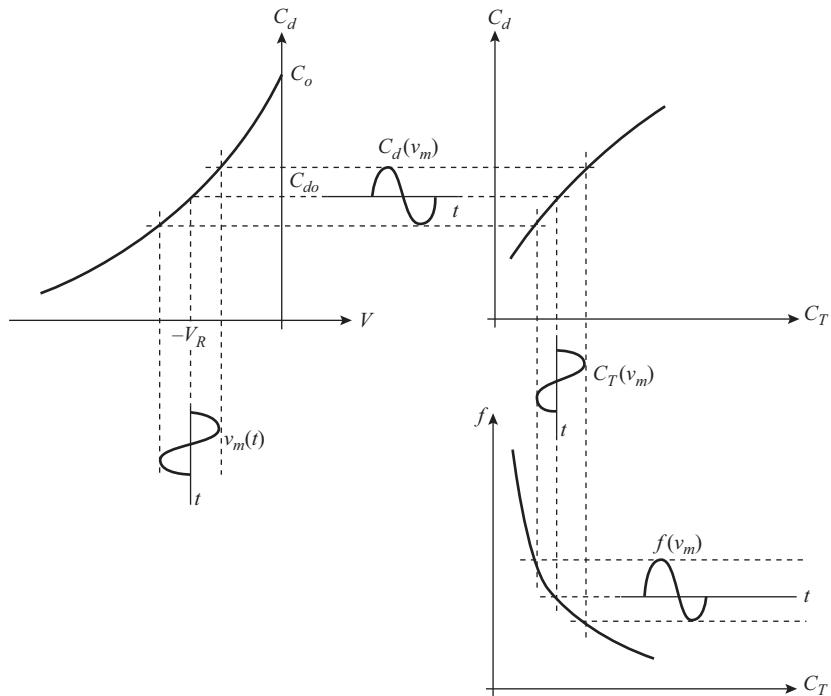


Figure 10.12.2 Graphical representation of the FM process.

$$F(v_m) = f_0 \sqrt{\frac{C_{TO}}{C_T(v_m)}} \quad (10.12.4)$$

These equations enable the transfer characteristic to be obtained. As mentioned in the introduction, the linear correlation coefficient gives a measure of the linearity of the transfer function, and the slope gives the frequency deviation constant. The computed results for the FM oscillator of Fig. 10.12.1 are shown in Fig. 10.12.3 (see Problem 10.22).

The varactor diode may also be used as a phase modulator, one possible circuit being shown in Fig. 10.12.4. Here a crystal-controlled oscillator circuit supplies a constant current signal to a tuned circuit. As before, the diode capacitance is in series with a fixed capacitance. What has changed compared to the FM circuit is that the frequency remains constant, but the phase angle of the tuned circuit is a function of the total tuning capacitance. A parallel tuned circuit with reasonable Q -factor ($Q \geq 10$) can be modeled by the dynamic resistance (see Section 1.4) in parallel with a lossless inductor L , in parallel with the total tuning capacitance C_T . Denoting the fixed frequency by ω_o , the admittance of the tuned circuit (with L and C not necessarily resonant) is

$$Y = \frac{1}{R_D} + j \left(\omega_o C_T - \frac{1}{\omega_o L} \right) \quad (10.12.5)$$

The phase angle of Y is

$$\gamma = \tan^{-1} \left(R_D \left(\omega_o C_T - \frac{1}{\omega_o L} \right) \right) \quad (10.12.6)$$

Now, since $Q = R_D/\omega_o L$, Eq. (10.12.6) can be rearranged as

$$\gamma = \tan^{-1} Q(\omega_o^2 LC_T - 1) \quad (10.12.7)$$

With no modulation applied, the tuned circuit resonates to the oscillator frequency ω_o when the tuning capacitance has some specific value $C_T = C_{TO}$. Hence the equation for phase angle becomes

$$\gamma = \tan^{-1} Q \left(\frac{C_T}{C_{TO}} - 1 \right) \quad (10.12.8)$$

With modulation, the capacitance changes, and therefore so does the phase angle. Now, since the current is constant, the voltage across the tuned circuit is

$$\begin{aligned} V_o &= \frac{I_o}{Y} \\ &= \frac{I_o}{|Y|} \angle -\gamma \end{aligned} \quad (10.12.9)$$

Hence the phase of the output voltage as a function of capacitance is

$$\begin{aligned} \phi &= -\gamma \\ &= \tan^{-1} Q \left(1 - \frac{C_T}{C_{TO}} \right) \end{aligned} \quad (10.12.10)$$

It will be noted that $\phi = 0$ when $C_T = C_{TO}$. As with the FM circuit, a plot of the transfer characteristic (in this instance ϕ/v_m) is obtained by evaluating in sequence Eqs. (10.12.1) through (10.12.3) and

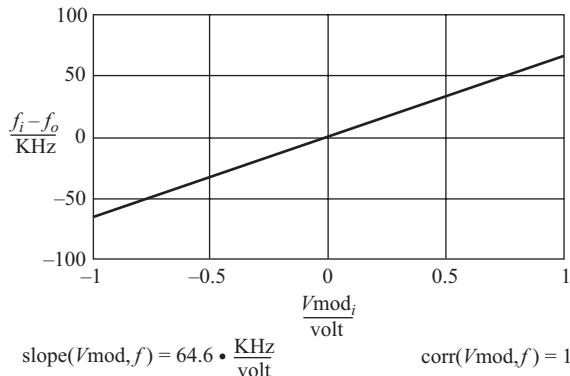


Figure 10.12.3 Plot of the transfer characteristic for the Clapp FM oscillator shown in Fig. 10.12.1. (Circuit details are given in Problem 10.22.) The linear correlation coefficient is unity, and the slope (rounded off) is $k = 65$ kHz/V.

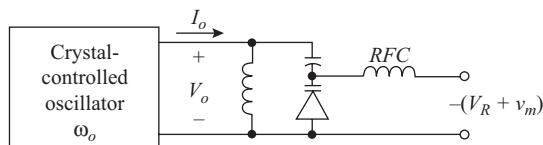


Figure 10.12.4 Varactor diode phase modulator.

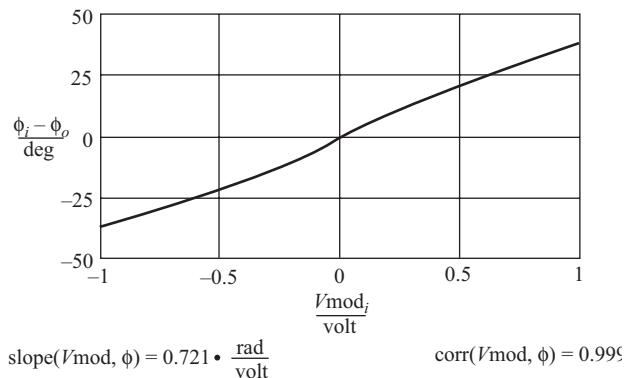


Figure 10.12.5 Transfer characteristic for the phase modulator of Fig. 10.12.4. (See Problem 10.24 for circuit details.) The linear correlation coefficient is 0.999, and the phase shift constant is 0.721 rad/V.

(10.12.10) for a range of values of modulating voltage. Figure 10.12.5 shows one such characteristic (see Problem 10.24).

Transistor Modulators

Transistor circuits can also be designed that behave as variable reactances (varactors) with the reactance controlled by the modulating signal. One such circuit utilizing a JFET is shown in Fig. 10.12.6.

The frequency determining circuit of the oscillator is connected to an external circuit consisting of a JFET, radio frequency chokes (RFCs), and impedances Z_1 and Z_2 . The small-signal oscillator voltage across the external circuit is shown as v_o , and the corresponding current entering this circuit is i_o . Hence the admittance presented by the external circuit to the oscillator is

$$\begin{aligned} Y &= \frac{i_o}{v_o} \\ &= \frac{i_d}{v_o} + \frac{1}{Z_1 + Z_2} \end{aligned} \quad (10.12.11)$$

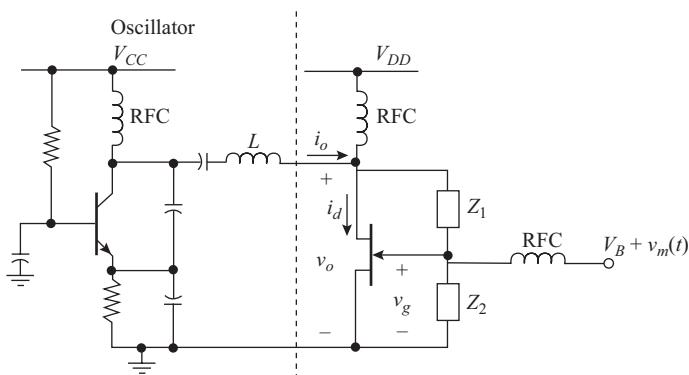


Figure 10.12.6 JFET reactance modulator.

This assumes that the admittance of the RFC in the drain circuit is small enough to be ignored. For the JFET, $i_d = g_m v_g$, and from the circuit, $v_g = v_o Z_2 / (Z_1 + Z_2)$. Substituting these relationships in the admittance equation gives

$$\begin{aligned} Y &= \frac{g_m Z_2}{Z_1 + Z_2} + \frac{1}{Z_1 + Z_2} \\ &= \frac{g_m Z_2 + 1}{Z_1 + Z_2} \end{aligned} \quad (10.12.12)$$

Thus the admittance is a function of the transconductance g_m , which in turn can be made to depend on the modulating voltage. In particular, the imaginary part of Y , which is the susceptance, can be made to alter the tuning and hence the frequency of the oscillator.

In practice, it will always be the case that $|Z_1| \gg |Z_2|$ and $|g_m Z_2| \gg 1$. The expression for admittance then reduces to

$$Y \approx g_m \frac{Z_2}{Z_1} \quad (10.12.13)$$

Four combinations of Z_1 and Z_2 may be used as follows: C_1, R_2 ; R_1, C_2 ; R_1, L_2 ; and L_1, R_2 . Each gives a different susceptance as a function of g_m . Consider the first arrangement for which Z_1 is a capacitive reactance $1/j\omega_o C_1$ and Z_2 a resistance R_2 . This results in an admittance

$$\begin{aligned} Y &= j\omega_o g_m C_1 R_2 \\ &= j\omega_o g_m \tau \end{aligned} \quad (10.12.14)$$

Here, $\tau = C_1 R_2$ is the time constant of the Z_1, Z_2 branch. The condition $|Z_1| \gg |Z_2|$ requires that $\omega_o \tau \ll 1$, and therefore the circuit is useful at relatively low frequencies. Also, the admittance is seen to be equal to a capacitive susceptance for which the equivalent capacitance is

$$C_{eq} = g_m \tau \quad (10.12.15)$$

The C_{eq} or L_{eq} component for the other combinations can be derived in a similar fashion, and these are listed in Table 10.12.1

TABLE 10.12.1

Time Constant, τ	Condition	L_{eq} or C_{eq}
$C_1 R_2$	$\omega \tau \ll 1$	$C_{eq} = g_m \tau$
$R_1 C_2$	$\omega \tau \gg 1$	$L_{eq} = \tau / g_m$
$R_1^{-1} L_2$	$\omega \tau \ll 1$	$C_{eq} = g_m \tau$
$L_1 R_2^{-1}$	$\omega \tau \gg 1$	$L_{eq} = \tau / g_m$

EXAMPLE 10.12.1

Assuming that “very much greater than” means a 10 : 1 ratio at least, determine the highest frequency for the reactance modulator for which the C_1, R_2 time constant is 0.3 μ s. Given that the lowest value encountered for the transconductance is 2 mS, determine the actual values of C_1 and R_2 . What is the value of the equivalent tuning capacitance for these values?

SOLUTION Denote the “much greater than” condition by $p = 10$, and the “much less than” condition by $q = 1/p$; then

$$f_{\max} = \frac{q}{2\pi\tau} \approx 53 \text{ kHz}$$

$$|Z_2| = \frac{P}{g_m(\min)} = 5 \text{ k}\Omega$$

Since Z_2 is a resistance, $R_2 = |Z_2| = 5 \text{ k}\Omega$. Check: $g_m |Z_2| = 9$ which meets the “very much greater than” criterion. It therefore follows that

$$C_1 = \frac{\tau}{R_2} = 60 \text{ pF}$$

$$C_{eq} = g_m \tau = 600 \text{ pF}$$

For the JFET, the transconductance is a linear function of the gate–source bias voltage, so by making the modulating voltage part of the bias, modulation of C_{eq} or L_{eq} is achieved. Denoting the modulating signal voltage by $v_m(t)$ and the fixed bias as V_B , then the gate–source bias voltage is given by

$$V_{GS} = V_B + v_m(t) \quad (10.12.16)$$

The transconductance for a JFET is linearly dependent on V_{GS} , the relationship being given by

$$g_m = g_{mo} \left(1 - \frac{V_{GS}}{V_P} \right) \quad (10.12.17)$$

Here, V_p is the pinch-off voltage and g_{mo} is the transconductance at zero bias, both of these being constants for a given JFET. It should be noted that for a JFET the parameters normally specified are V_p and the current at zero bias I_{DSS} , and $g_{mo} = -2I_{DSS}/V_p$.

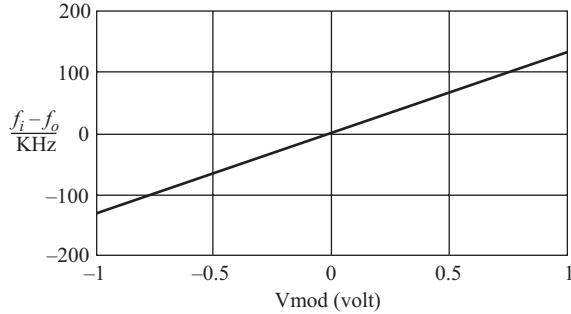
Since L or C may be varied, then denoting the LC combination with no modulation applied as $(LC)_o$ and the corresponding oscillator frequency as f_o , and $(LC)_m$ as the modulated value, the modulated frequency is given by

$$f(v_m) = f_o \sqrt{\frac{(LC)_o}{(LC)_m}} \quad (10.12.18)$$

Again it will be seen that a number of nonlinear relationships are involved, but as remarked in the introduction, an approximately linear relationship can be achieved for the transfer characteristic by keeping the fractional variation in the variables small compared to unity. The linear correlation coefficient and the frequency deviation coefficient can be found from the transfer characteristic as described for the varactor diode modulator. Figure 10.12.7 shows the result for the modulator specified in Problem 10.27.

The transistor may also be used as a phase modulator; one circuit utilizing a JFET is shown in Fig. 10.12.8. In this circuit, a crystal-controlled oscillator is shown to emphasize the fact that the external circuit does not affect the frequency. What has to be shown is that the phase of the output voltage v_d is a function of the JFET g_m , which in turn is a function of the modulating voltage. Assuming the gate current is negligible, then

$$i_o = (v_o - v_d) j \frac{\omega_o C}{2} \quad (10.12.19)$$



$$\text{slope}(V_{\text{mod}}, f) = 112.7 \frac{\text{kHz}}{\text{volt}} \quad \text{corr}(V_{\text{mod}}, f) = 1$$

Figure 10.12.7 Transfer curve for the JFET FM modulator of Problem 10.27. The linear correlation coefficient is unity, and the slope (rounded off) is 113 kHz/V.

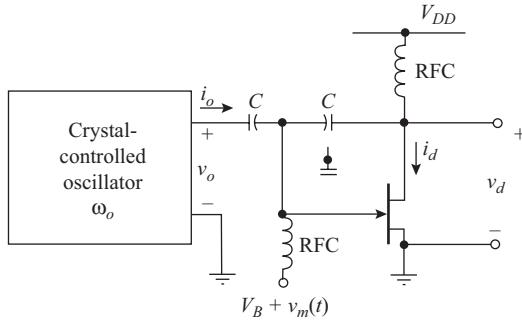


Figure 10.12.8 JFET phase modulator.

As before, the drain current is given by $i_d = g_m v_g$, and the RF gate voltage in this case is

$$v_g = \frac{v_o + v_d}{2} \quad (10.12.20)$$

Hence, on the assumption that $i_o \cong i_d$,

$$g_m \frac{v_o + v_d}{2} = (v_o - v_d) j \frac{\omega_o C}{2} \quad (10.12.21)$$

From this, the output voltage is obtained as

$$v_d = -v_o \frac{g_m - j \omega_o C}{g_m + j \omega_o C} \quad (10.12.22)$$

The modulus of the complex term in the numerator is equal to that in the denominator, so variations in g_m do not affect the amplitude, which is an advantage. It also follows that $|v_d| = |v_o|$. In polar form, Eq. (10.12.19) becomes

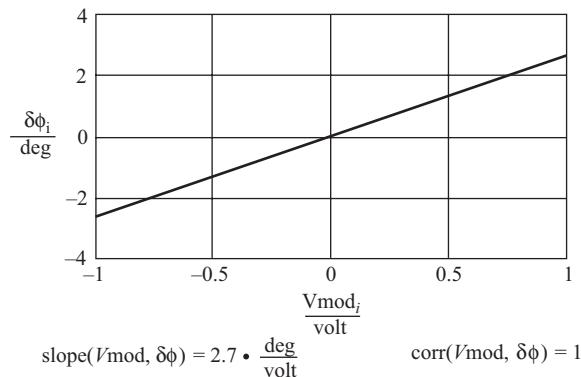


Figure 10.12.9 Transfer curve for the phase modulator of Problem 10.35. The slope of the curve is $2.67^\circ/\text{V}$, and the linear correlation coefficient is unity.

$$\begin{aligned} |v_d| \angle \phi &= |v_o| \frac{(\angle 180^\circ - \gamma)}{\angle \gamma} \\ \therefore \phi &= 180^\circ - 2\gamma \end{aligned} \quad (10.12.23)$$

where

$$\gamma = \tan^{-1} \frac{\omega_o C}{g_m} \quad (10.12.24)$$

Hence the output voltage phase angle ϕ in radians, as a function g_m , is

$$\phi = \pi - 2 \tan^{-1} \frac{\omega_o C}{g_m} \quad (10.12.25)$$

Denoting the unmodulated phase angle as ϕ_o the change in phase angle is given by

$$\delta\phi = \phi - \phi_o \quad (10.12.26)$$

A plot of the transfer characteristic (in this instance phase angle versus modulating voltage) for the modulator detailed in Problem 10.35 is shown in Fig. 10.12.9.

10.13 FM Transmitters

Direct frequency modulation can be employed using any of the FM circuits described in the previous section. Usually, however, direct FM at the final carrier frequency is not feasible because of the problem of maintaining high-frequency stability of the carrier while at the same time obtaining adequate frequency deviation. A commonly used technique is to frequency modulate a subcarrier oscillator and then use a combination of frequency multiplication and mixing to achieve the desired final carrier frequency with specified maximum deviation.

The distinction between frequency multiplication and mixing is important in FM systems. With frequency multiplication, the instantaneous frequency is multiplied. For example if the instantaneous frequency of an FM oscillator is $f_i = f_o + \Delta f$, when passed through a frequency multiplier this becomes $n f_i = n f_o + n \Delta f$, where n is the multiplying factor. Frequency multiplication can be achieved by passing the signal through a

class C amplifier and tuning the output to the desired harmonic. (Recall that with a class C amplifier the output current is in the form of pulses at the input frequency, and the output circuit can be tuned to a harmonic of this.)

With frequency mixing, the deviation is not altered. For example, if a signal with instantaneous frequency $f_i = f_o + \Delta f$ is passed through a mixer, which is also fed by a local oscillator f_x , the output circuit can be tuned to the difference frequency $f_o + \Delta f - f_x = f_{IF} + \Delta f$. Figure 10.13.1 shows how a combination of frequency multiplication and mixing may be used to increase deviation without increasing the nominal FM oscillator frequency.

An *LC* oscillator may be directly frequency modulated to produce relatively large deviation, but then it may be difficult to maintain stability of the center (unmodulated) frequency. Stability can be improved by using automatic frequency control (AFC). Figure 10.13.2 shows one possible arrangement. A sample of the output signal, taken from the final driver stage, is mixed with a signal from a stable crystal oscillator. The difference frequency output from the mixer is fed into the discriminator circuit, where it is demodulated (discriminator circuits are described in Section 10.14). The demodulated output consists of the original modulating signal plus any variations caused by drifting of the nominal frequency of the *LC* oscillator. The low-pass filter removes the signal component while leaving the drift component, which is fed back as a control bias to the oscillator. This is connected in such a way as to shift the nominal frequency in the desired direction to reduce the drift.

Class C power amplifiers can be used to amplify FM signals since the amplitude distortion introduced by the class-C operation has no effect on the modulation (unlike AM, where class C cannot be used once the carrier is modulated). This makes for more efficient FM transmitters compared to AM.

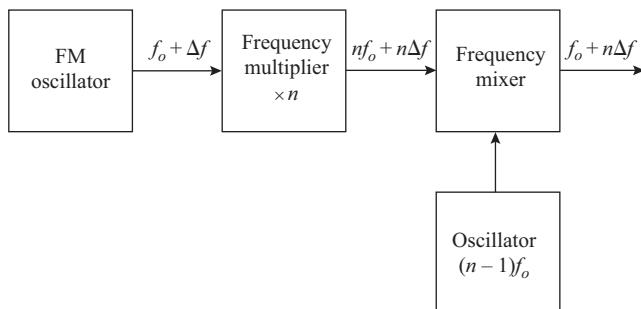


Figure 10.13.1 Use of frequency multiplication and mixing to increase the frequency deviation.

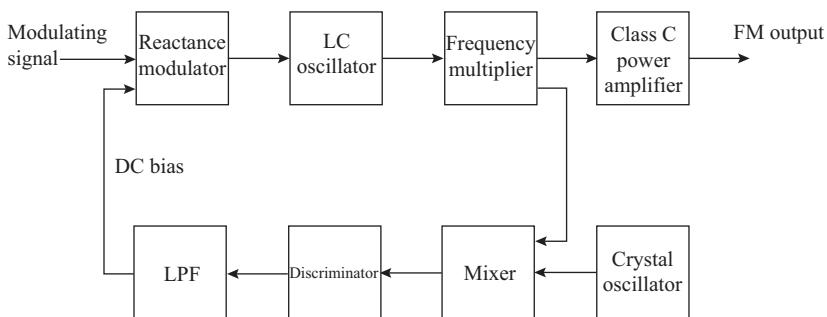


Figure 10.13.2 Frequency-stabilized FM oscillator.

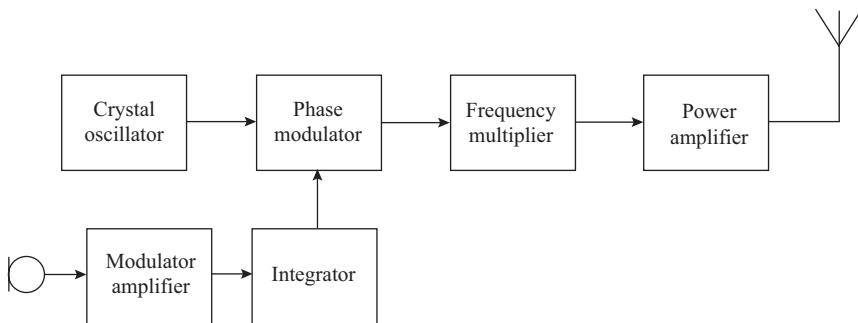


Figure 10.13.3 FM achieved through phase modulation.

Phase modulation may be used to indirectly frequency modulate an oscillator as described in Section 10.9. This allows a crystal-controlled oscillator to be used, as shown in the block schematic of Fig. 10.13.3. This method is widely used in VHF and UHF radio telephone equipment.

A very popular indirect method of achieving FM is known as the *Armstrong method* after its inventor. In this method, the initial modulation takes place as an amplitude-modulated DSBSC signal (Section 8.9) so that a crystal-controlled oscillator can be used if desired. To see the connection between DSBSC AM and FM, it is first necessary to analyze a low-level phase modulated signal, which provides the key to the connection. Equation (10.8.3) for phase modulation is repeated here:

$$e(t) = E_c \max \cos (\omega_c t + Kv_m(t)) \quad (10.13.1)$$

The trigonometric expansion for this is

$$e(t) = E_c \max [\cos \omega_c t \cdot \cos (Kv_m(t)) - \sin \omega_c t \cdot \sin (Kv_m(t))] \quad (10.13.2)$$

Now the peak phase excursion in the angle $Kv_m(t)$ can be kept small such that $\cos (Kv_m(t)) \approx 1$ and $\sin (Kv_m(t)) \approx Kv_m(t)$. The expansion for the phase modulated wave therefore becomes

$$e(t) \approx E_c \max [\cos \omega_c t - (Kv_m(t)) \sin \omega_c t] \quad (10.13.3)$$

The first term of the expansion is seen to be the unmodulated carrier, and the second term is a DSBSC term similar to that of Eq. (8.9.1), but with the carrier shifted by 90° . Thus, by arranging for the summation of such a DSBSC signal with a carrier term, phase modulation can be achieved through AM.

Figure 10.13.4 shows how this might be done. The crystal oscillator generates the subcarrier, which can be low, say on the order of 100 kHz. One output from the oscillator is phase shifted by 90° to produce the sine term, which is then DSBSC modulated in the balanced modulator by $v_m(t)$. This is combined with the direct output from the oscillator in the summing amplifier, the result then being the phase-modulated signal given by Eq. (10.13.3). The modulating signal is passed through an integrator, which as shown in Section 10.9 results in an equivalent frequency modulation. At this stage, the equivalent frequency deviation will be low, so the arrangement of Fig. 10.13.1 is used to increase the peak deviation.

EXAMPLE 10.13.1

An Armstrong transmitter is to be used for transmission at 152 MHz in the VHF band, with a maximum deviation of 15 kHz at a minimum audio frequency of 100 Hz. The primary oscillator is to be a 100-kHz

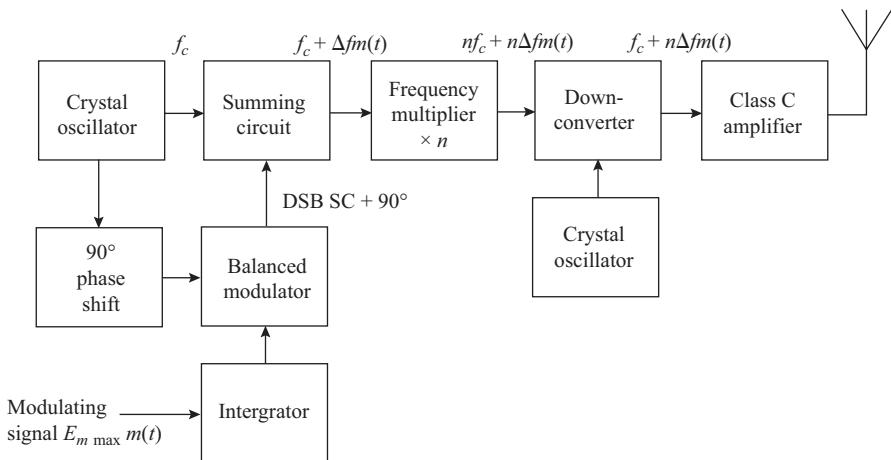


Figure 10.13.4 Armstrong method for FM.

crystal oscillator, and the initial phase-modulation deviation is to be kept to less than 12° , to avoid audio distortion. Find the amount by which the frequency must be multiplied to give proper deviation, and specify a combination of doublers and triplers that will give this. Also, specify the mixer crystal and any multiplier stages needed.

SOLUTION The maximum phase deviation of the modulator is

$$\Delta\phi_{\max} = 12^\circ \cong 0.21 \text{ rad}$$

From Eq. (10.10.5),

$$\Delta f_{\max} = \Delta\phi_{\max} f_{\min} = 0.21 \times 100 = 21 \text{ Hz}$$

The frequency deviation increase required is

$$N = \frac{\Delta f_{\max \text{ allow}}}{\Delta f_{\max}} = \frac{15,000}{21} \cong 714 \quad (\text{use } 729 = 3^6)$$

The modulated waveform will be passed through a chain of six tripler stages, giving a final deviation of

$$21 \times 729 = 15.31 \text{ kHz}$$

at a frequency of $0.1 \times 729 = 72.9$ MHz. The deviation is slightly high, but a slight attenuation of the audio signal will compensate for this. The mixer oscillator signal is

$$f_0 = 152 - 72.9 = 79.1 \text{ MHz}$$

f_0 is best obtained by using two tripler stages from an 8.7889-MHz crystal oscillator.

FM Broadcast

The prime requisite of FM broadcasting is excellent fidelity, since music provides the chief program material. Frequency modulation acts in several ways to improve this fidelity. First, since FM broadcasting takes place in the VHF band from 88 to 108 MHz, a much wider baseband can be used. The baseband width presently in use

is 50 Hz to 15 kHz, with a maximum allowable deviation of ± 75 kHz. Channel spacing is 200 kHz, and power outputs of as much as 100 kW are used.

Until recently single-channel monophonic FM broadcasts were common, but nearly all FM stations are now transmitting two-channel stereo programs. In the near future this may be further changed to provide four-channel stereo. Also, some FM stations are frequency division multiplexing an additional channel on their carrier for the purpose of providing background music for public buildings, a system licensed as a *subsidiary communications authorization*, or SCA.

The transmitters used are typically of the Armstrong type discussed previously. Initial deviations are kept small to limit modulation distortion, and many stages of frequency multiplication are used to bring the deviation up to that required at the output. Again, to provide the necessary power output, the output stage is a push-pull parallel class C amplifier. Water-cooled vacuum tubes are used in this stage.

The main differences between the FM systems already discussed and the broadcast system lies in the composition of the audio signal presented to the modulator. It is a composite signal carrying several signals, and these will be discussed.

First, if monophonic FM transmission is to be used, only one channel is needed, and the single audio channel is applied directly to the modulator input. It must be fully compensated to provide good fidelity in the band from 50 to 15,000 Hz.

Two-channel stereo is accomplished as follows. The two audio channels are not simply frequency division multiplexed before modulation. They are first mixed to provide two new signals, one of which is a *balanced* monophonic signal. The first is the sum of the two input channels, and the second is the difference of the two. The sum channel is modulated directly in the baseband assignment between 50 and 15,000 Hz. The difference signal is DSBSC modulated in the 23- to 53-kHz slot about a carrier at 38 kHz. A pilot carrier at 19 kHz is also transmitted. The sum signal in the audio portion of the band can be demodulated by a monophonic FM receiver to provide normal monophonic reception. A receiver with a stereo demodulator can also retrieve the difference signal and combine the two to produce the original L and R channel signals. Figure 10.13.5(a) shows the block diagram of the premodulation mixing circuits.

The L and R channels are summed and passed through a low-pass 15-kHz filter to form the monophonic portion of the baseband signal. The R channel is inverted and then added to the L signal to yield the difference signal $L - R$. This signal is DSBSC-modulated on the 38-kHz carrier by a balanced modulator and passed through a 23- to 53-kHz band-pass filter to remove any unwanted signal components. The two channel bands and the 19-kHz pilot carrier are added together to produce the final baseband signal (Fig. 10.13.5[b]). This final composite signal is presented to the modulator input of the FM transmitter.

An additional channel is often modulated on the same carrier for service to commercial operations, such as music to stores and public buildings. This channel is limited to a signal bandwidth of 7.5 kHz and is multiplexed to lie in the range from 53 to 75 kHz, with a pilot carrier at 67 kHz. The sideband frequencies are shown in Fig. 10.13.5(b). This auxiliary channel does not interfere in any way with the ordinary broadcast.

10.14 Angle Modulation Detectors

Basic Detection of FM Signals

To detect an FM signal, it is necessary to have a circuit whose output voltage varies *linearly* with the frequency of the input signal. The slope detector is a very basic form of such a circuit, although its linearity of response is not good. Figure 10.14.1(a) shows the basic arrangement. By tuning the circuit to receive the signal on the slope of the response curve Fig. 10.14.1(b), the carrier amplitude V is caused to vary with frequency. In this case the circuit is tuned so that its resonant frequency f_0 is lower than the carrier frequency f_c . When the signal

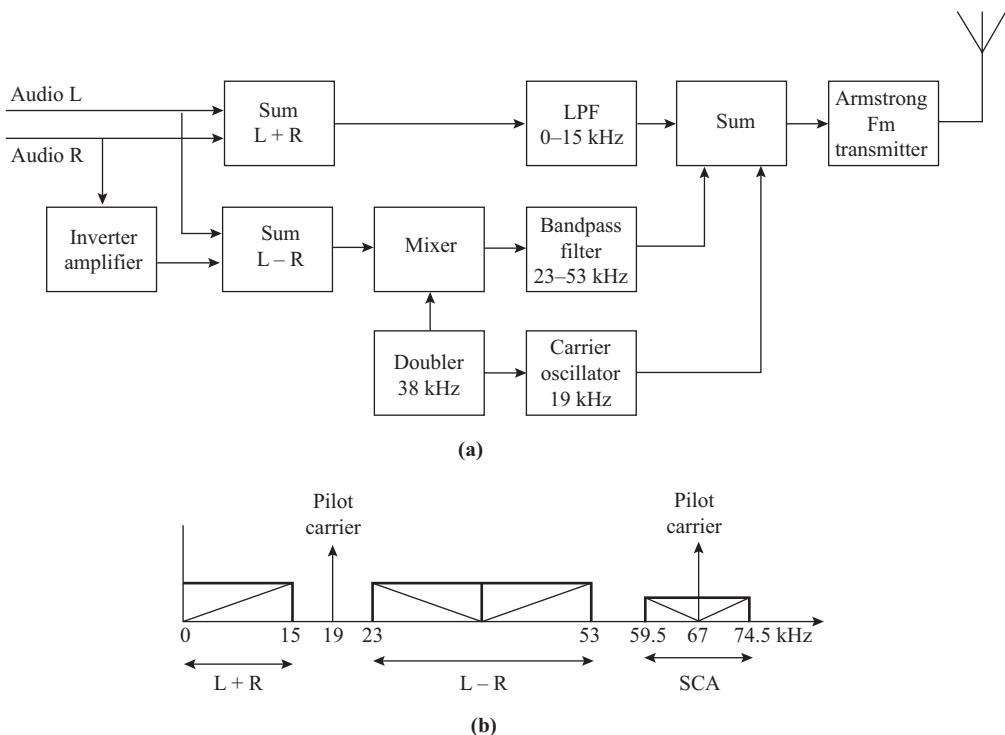


Figure 10.13.5 FM stereo broadcast transmitter: (a) block schematic; (b) baseband spectrum showing position of the auxiliary 7.5-kHz channel.

frequency increases above f_c , with modulation the amplitude of the carrier voltage drops. When the signal frequency decreases below f_c , the carrier voltage rises. The change of voltage results because of the change in the magnitude of the impedance in the tuned circuit as a function of frequency and results in an effective conversion of frequency modulation into amplitude modulation. The modulation is recovered from the amplitude modulation by means of a normal envelope detector. However, the linear range on the voltage/frequency-transfer characteristic is limited.

Linearity can be improved by utilizing the arrangement of Fig. 10.14.1(c), a circuit known as the *Round-Travis detector* or *balanced slope detector*. This circuit combines two circuits of the type shown in Fig. 10.14.1(a) in a balanced configuration. One slope detector is tuned to f_{01} above the incoming carrier frequency and the other to f_{02} below the carrier frequency, and the envelope detectors combine to give a differential output. This means that by showing the V_1 response as positive that V_2 response can be shown as negative on the same axes, and the output $V_0 = |V_1| + |V_2|$ will have an S shape when plotted against frequency, shown in Fig. 10.14.1(d). This S curve is characteristic of FM detectors. When the incoming signal is unmodulated, the output is balanced to zero; when the carrier deviates towards f_{01} , $|V_1|$ increases while $|V_2|$ decreases, and the output goes positive; when it deviates toward f_{02} , $|V_1|$ decreases while $|V_2|$ increases, and the output goes negative.

Foster–Seeley Discriminator

The phase shift between primary and secondary voltages of a tuned transformer is a function of frequency, and the Foster–Seeley discriminator utilizes this frequency–phase dependence for the recovery of the modulating

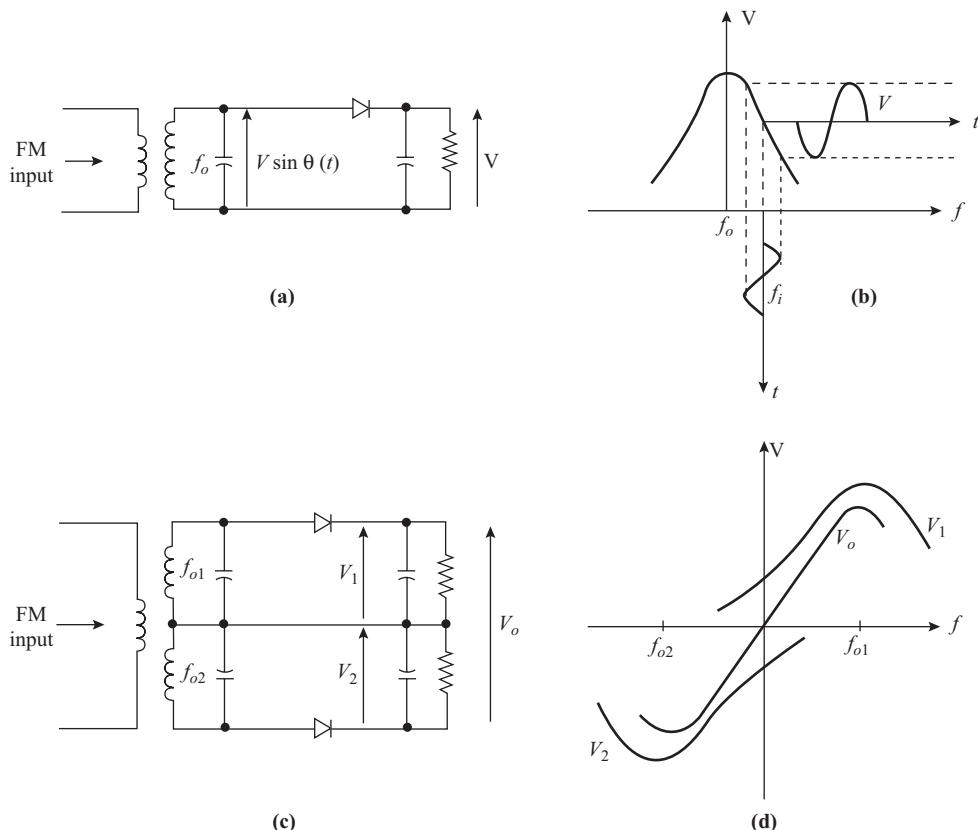


Figure 10.14.1 (a) FM slope detector; (b) magnitude of output voltage plotted against frequency; (c) balanced double-tuned slope (or Round Travis) detector; (d) output voltage plotted against frequency, showing the S-shaped transfer function curve.

signal. Figure 10.14.2(a) shows a tuned transformer (see Section 1.8), the voltage transfer function (VTF) being given by Eq. (1.8.10). This is plotted in Fig. 10.14.2(b) and (c) for a typical transformer as a function of the fractional frequency change from resonance. Recalling that the VTF is the ratio of secondary voltage to primary voltage, it will be seen that the secondary voltage lags the primary voltage by nearly 90° at resonance and that phase shift increases as frequency decreases, and decreases as frequency increases.

Figure 10.14.3(a) shows the basic arrangement for the Foster–Seeley discriminator. The primary voltage is tightly coupled through capacitor C_3 and the RFC to the center tap on the secondary (in some circuits a tertiary winding is used for this coupling). The coupling is tight enough that practically all the primary voltage appears between the center tap and ground. Figure 10.14.3(b) shows the generator equivalent circuit, where it will be seen that the voltage across diode D_1 is $V_{D1} = V_1 + 0.5 V_2$ and that across diode D_2 is $V_{D2} = V_1 - 0.5 V_2$.

In terms of the voltage transfer function, $V_{D1} = V_1(1 + VTF/2)$ and $V_{D2} = V_1(1 - VTF/2)$. The magnitude of these voltages are plotted in Fig. 10.14.4 (normalized to $V_1 = 1$ V), where it will be seen that, as a result of the phase relationship shown in Fig. 10.14.2(c), $|V_{D1}|$ varies in the opposite sense to that for $|V_{D2}|$. The RF voltages are rectified by the diode circuits, and the outputs developed across each diode load will be almost equal to the peak RF voltage at the respective inputs. Thus $V_{o1} \cong |V_{D1}|$ and $V_{o2} \cong |V_{D2}|$. The total output voltage is the *difference* between these, or $V_o \cong |V_{D1}| - |V_{D2}|$. This is plotted in Fig. 10.14.4(b).

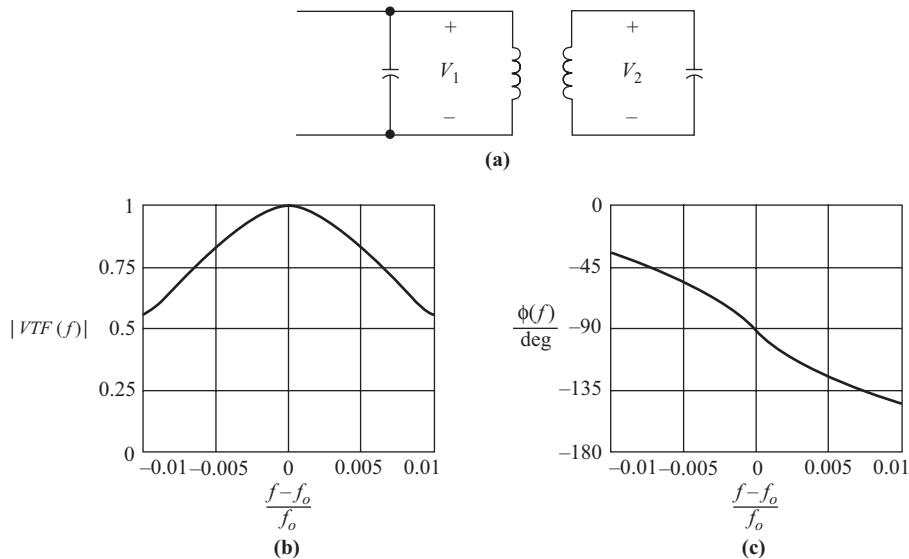


Figure 10.14.2 (a) Tuned transformer. (b) Magnitude of the VTF and (c) phase of the VTF

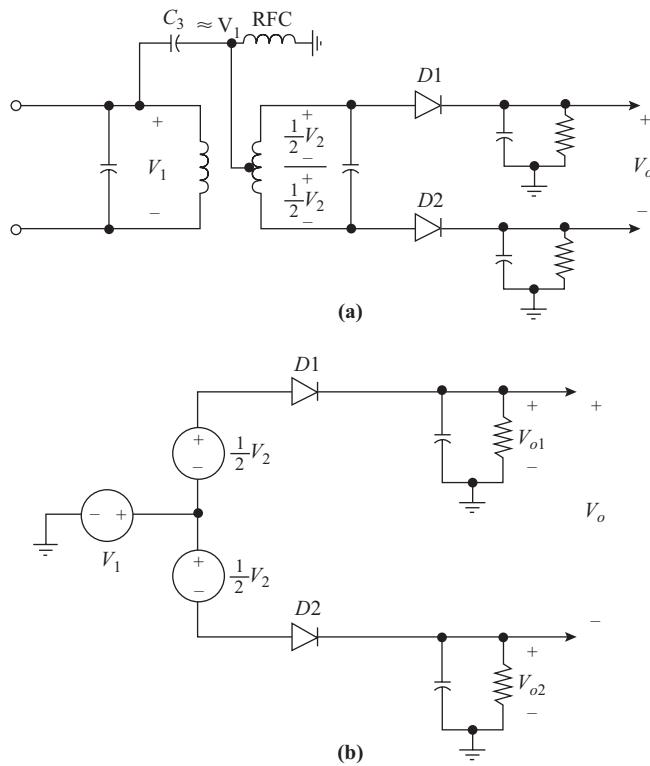


Figure 10.14.3 (a) Basic Foster-Seeley discriminator. (b) Voltage generator equivalent circuit.

The difference-voltage/frequency curve has the characteristic S shape, and over the linear portion of the curve, the output voltage is proportional to the frequency deviation.

Amplitude variations are reduced to negligible proportions by amplitude limiting the signal before applying it to the discriminator. There are many practical variations of the circuit, and Fig. 10.14.5 shows it being used with integrated circuits. In this application, amplifier block MFC6010 provides gain and amplifier block MC1355 provides gain limiting. Note that the primary voltage is coupled into the secondary by means of a tertiary winding that is very closely coupled to the primary. Also, the ground point is shifted so that the output may be taken with respect to ground, a much more practical arrangement.

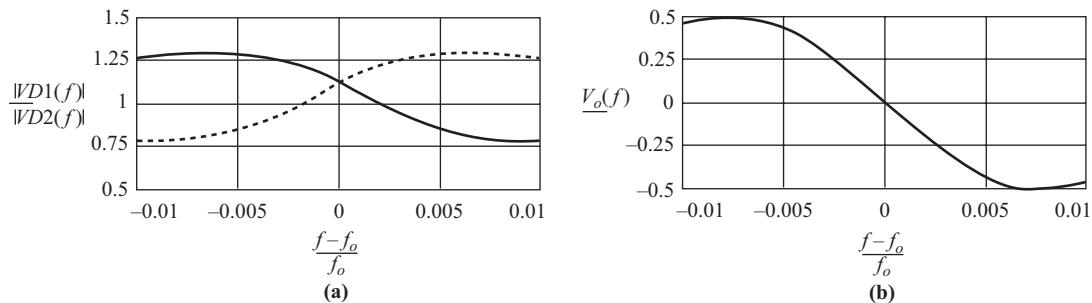


Figure 10.14.4 Magnitude of the voltage applied to (a) diode D_1 and diode D_2 . (b) Output voltage, which is equal to the difference in the diode voltage magnitudes.

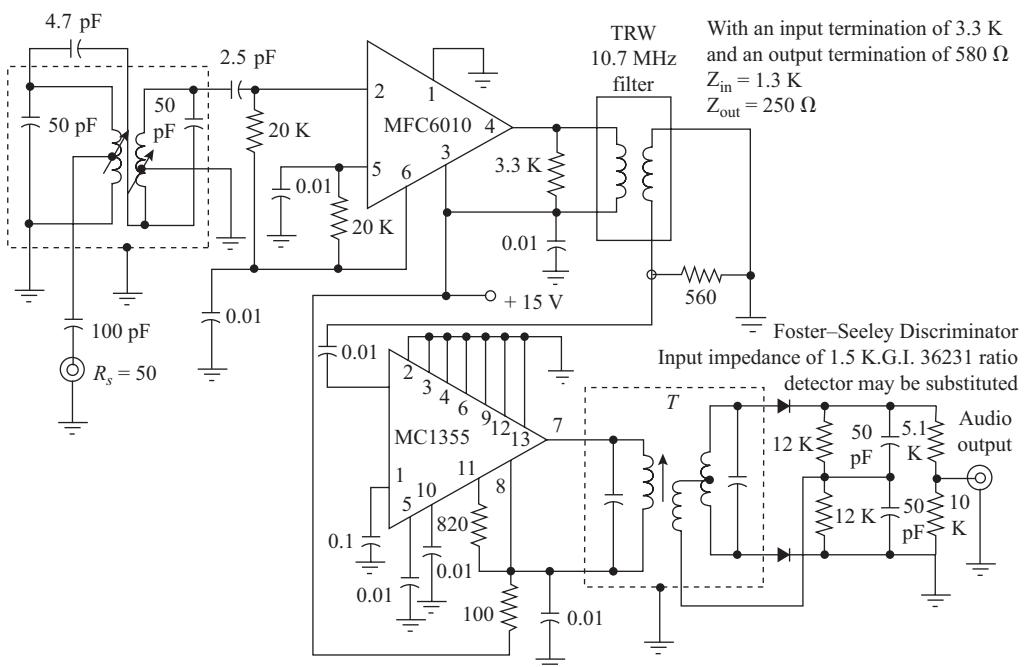


Figure 10.14.5 FM receiver unit utilizing integrated-circuit amplifiers and a Foster-Seeley discriminator. (Courtesy Motorola, Inc.)

Ratio Detector

A very simple change can be made to the Foster–Seeley discriminator that improves the limiting action, but at the expense of reducing the output. One diode is reversed so that the two diodes conduct in series (Fig. 10.14.6); the detector circuits then provide a damping action, which tends to maintain a constant secondary voltage, as will be shown.

Diodes D_1 and D_2 and associated loads RC form envelope detectors as before, and the frequency-to-phase-to-amplitude chain of events occurs as in the Foster–Seeley discriminator. Now, however, the polarity of voltage in the lower capacitor is reversed, so the sum voltage appears across the combined loads (rather than the difference voltage as in the Foster–Seeley). Hence, as V_{01} increases, V_{02} decreases and V_0' remains constant (and, of course, V_0' remains constant as V_{01} decreases and V_{02} increases). Therefore, a large capacitor (in practice usually an electrolytic type) can be connected across the V_0' points without affecting the voltage, except to improve the “constancy.”

From the circuit of Fig. 10.14.6, two equations can be written for the output voltage V_o , as follows:

$$V_0 = \frac{1}{2}V_0' - V_{02}$$

and

$$V_0 = -\frac{1}{2}V_0' + V_{01}$$

Adding,

$$2V_0 = V_{01} - V_{02}$$

Therefore,

$$\begin{aligned} V_0 &= \frac{1}{2}(V_{01} - V_{02}) \\ &= \frac{1}{2}(|V_{D1}| - |V_{D2}|) \end{aligned} \quad (10.14.1)$$

The output voltage is one-half that of the Foster–Seeley circuit.

Limiting action occurs as a result of variable damping on the secondary of the transformer. For example, if the input voltage amplitude ($V_{1 \text{ max}}$) were to suddenly increase, as would occur with a noise spike of voltage, the voltage V_0' could not follow immediately, since it is held constant by means of the large capacitor. The

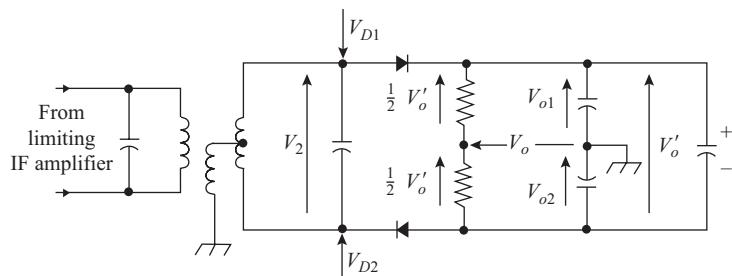


Figure 10.14.6 Ratio detector circuit.

voltage across the diodes in series is $V_2 - V'_0$ and since V_2 will increase with V_1 , the diodes conduct more heavily (that is, more current flows). This results in heavier damping of the secondary (which is also reflected into the primary), which reduces the Q factor. This in turn tends to offset the increase in V_2 by reducing the gain of the limiting amplifier feeding the circuit.

If the input voltage should decrease suddenly, diode conduction is reduced and so is the damping. The gain of the limiting amplifier therefore increases, tending to offset the original decrease in voltage. The fact, too, that the diode load, as seen by the two diodes in series, has a long time constant and therefore cannot respond to fast changes in voltage helps in the limiting process.

Of course, the limiting action will not be effective for slow changes in amplitude, as V'_0 will gradually adjust to the new level determined by input voltage amplitude.

Figure 10.14.7 shows the circuit diagram of a high-quality FM stage utilizing a ratio detector in conjunction with integrated-circuit amplifiers. Both amplifying blocks MC1355 provide limiting gain.

Quadrature Detector

This type of FM detector also depends on the frequency dependence of the phase angle of a tuned circuit, but has the advantage that only a single tuned circuit is required. As a result, it is becoming increasingly popular in integrated FM strips.

The impedance of a parallel tuned circuit can be expressed as $|Z_p(f)\angle\phi(f)$, where both the magnitude and phase are functions of frequency [see Eq. (1.4.2)]. Figure 10.14.8 shows how the magnitude and phase vary with frequency for a typical circuit used in the quadrature detector tuned to 10.7 MHz.

It will be seen that, to a close approximation, the phase angle varies as a linear function of frequency. In the quadrature detector, the voltage developed across such a tuned circuit by the FM carrier current is applied to one input of a multiplier, and the voltage across an inductance in series with the tuned circuit is applied to the other input, as shown in Fig. 10.14.9(a).

The quasi-steady-state analysis then proceeds as follows. Let I represent the constant current at angular frequency ω ; then the voltage across the inductance is $V_L = j\omega LI = \omega LI\angle 90^\circ$. Likewise, the voltage across the parallel tuned circuit is $V_p = |Z_p|I\angle\phi$. The output from the multiplier is $V_o = KV_L V_p$, where K is the multiplier constant. At the output of the multiplier the low-pass filter passes the average output voltage, while removing the RF components. Now the average value of the output is

$$V_{o \text{ avg}} = K|V_L| |V_p| \cos(90^\circ - \phi) \quad (10.14.2)$$

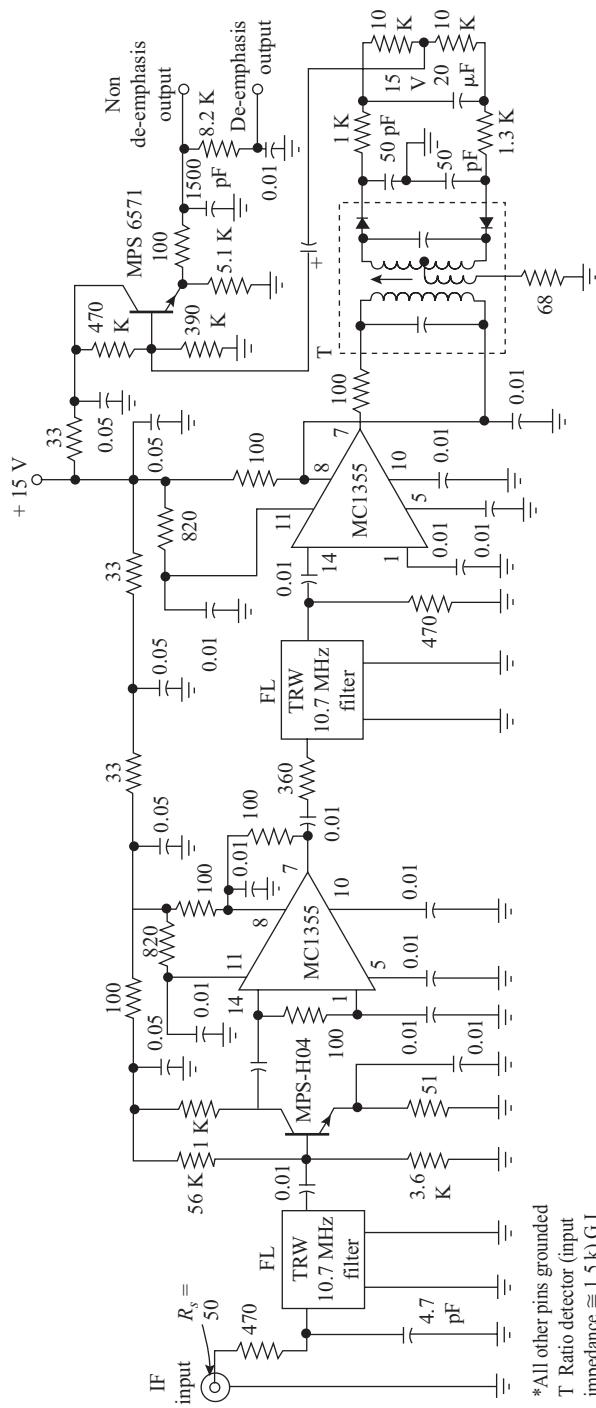
(This is similar to the average power equation $VI \cos \theta$, where θ is the phase angle between voltage and current.) Equation (10.14.2) can be rewritten as

$$V_{o \text{ avg}} = K|V_L| |V_p| \sin \phi \quad (10.14.3)$$

Thus the average output voltage is proportional to the $\sin \phi$, and for small values of phase angle, $\sin \phi \approx \phi$. As shown in Fig. 10.14.8(b), the phase angle varies almost linearly with frequency, and hence the average output voltage varies in the same manner. Figure 10.14.9(b) shows the average output voltage computed as a function of the frequency change from resonance for the quadrature detector of Fig. 10.14.10 (see Problem 10.45).

Phase Locked Loop Demodulator

The block schematic for a phase locked loop (PLL) is shown in Fig. 10.14.11. The phase detector, which is basically a balanced modulator, produces an average (or low-frequency) output voltage that is a linear function of the phase difference between the two input signals. The low frequency component is selected by the low pass filter which also removes much of the noise. The filtered signal is amplified through amplifier A



*All other pins grounded
T Ratio detector (input impedance $\approx 1.5\text{ k}\ \Omega$) G.I.
#36231 or equivalent.

Figure 10.14.7 High-quality FM receiver unit utilizing integrated-circuit amplifiers and ratio detector. (Courtesy Motorola, Inc.)

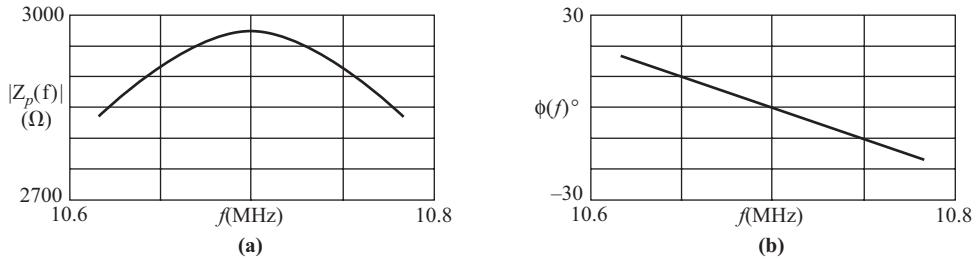


Figure 10.14.8 (a) Impedance magnitude and (b) phase angle for a parallel tuned circuit.

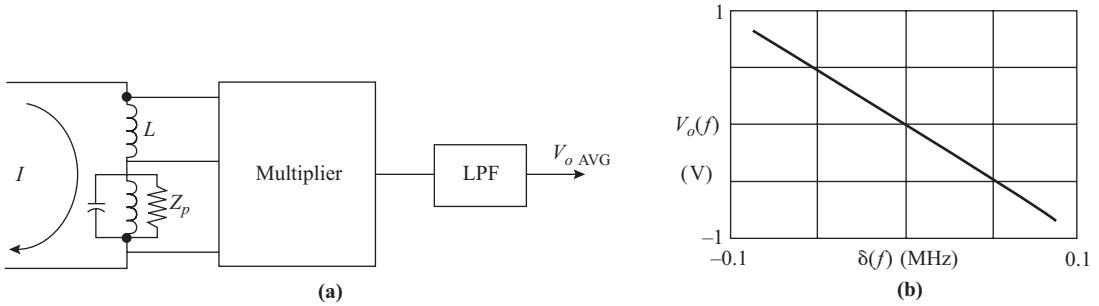


Figure 10.14.9 (a) Circuit arrangement for a quadrature detector. (b) Average output voltage.

and passed as a control voltage to the VCO where it results in frequency modulation of the VCO frequency. When the loop is in lock the VCO frequency follows or tracks the incoming frequency. For example, when the instantaneous frequency increases, the control voltage will cause the VCO frequency to increase.

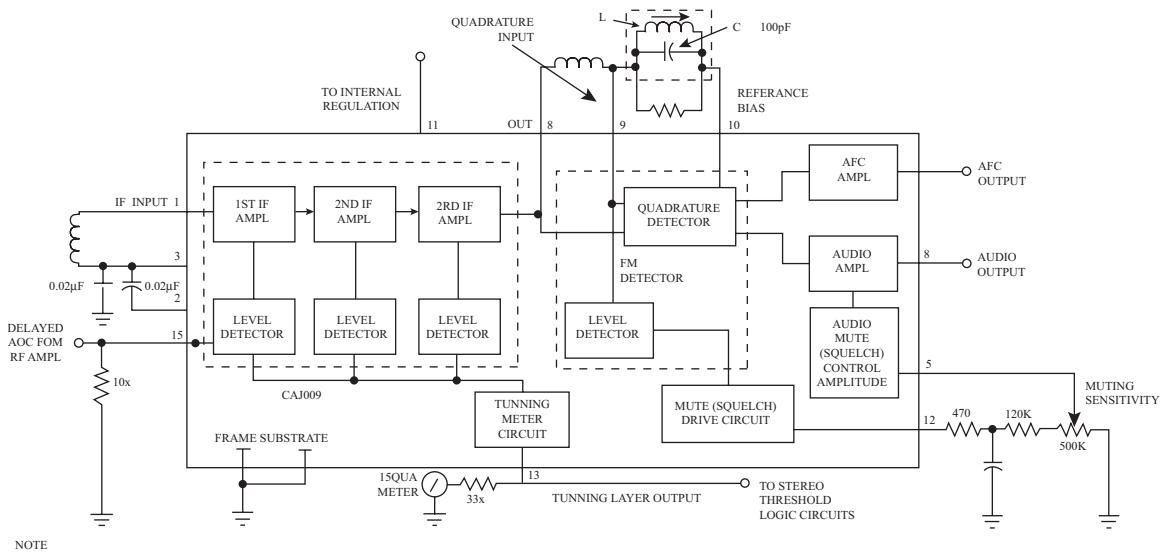
As stated in Eq. (10.2.1) the instantaneous frequency of an FM wave is $f_i(t) = f_c + ke_m(t)$. For the VCO, the instantaneous frequency can be written as $f_{vco}(t) = f_0 + k_{vco}V_c(t)$, where f_0 is the free-running frequency, and k_{vco} is the frequency deviation constant of the VCO. For the VCO frequency to track the instantaneous incoming frequency

$$f_{vco} = f_i \quad (10.14.4)$$

or $f_0 + k_{vco}V_c(t) = f_c + ke_m(t)$. From this it is seen that

$$V_c(t) = \frac{f_c - f_0 + ke_m(t)}{k_{vco}} \quad (10.14.5)$$

The VCO can be tuned so that $f_0 = f_c$ (see Section 6.6), and hence the control voltage is equal to a constant times the modulating signal voltage. From Fig. 10.14.11, it is seen that $V_c(t)$ is also taken as the output voltage, which therefore is the demodulated output.



NOTE

All resistor values are typical and in ohms

"L tunes with 100pF (C) at 10.7MHz

Quie 75.1G1 Ex22741 or equivalents

Figure 10.14.10 Block diagram of Signetics CA3089 FM IF system incorporating the quadrature detector. The System is available in a single 16-lead dual-in-line package. Externally required components are shown outside the main block. (Permission to reprint granted by Signetics Corporation, a subsidiary of U.S. Philips Corp., 811 E. Arques Avenue, Sunnyvale, CA 94086.)

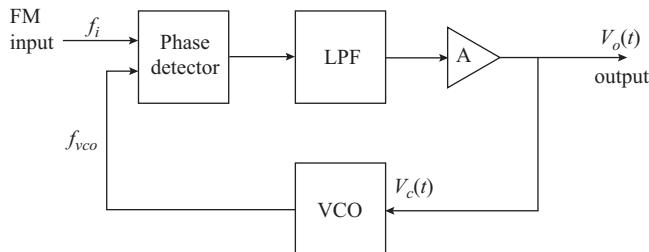


Figure 10.14.11 Phase-locked loop FM demodulator.

10.15 Automatic Frequency Control

Automatic frequency control (AFC) is highly desirable in tunable FM receivers to prevent distortion arising from oscillator frequency drift. The characteristic S curve for FM detectors (see Section 10.14) requires that the quiescent operating point be at the zero origin, which is at the center of the S curve [see, for example, Fig. 10.14.4(c)]. Any oscillator frequency drift in one direction or the other shifts the quiescent point away from the center. The peak frequency deviation on the side to which the point shifts may therefore enter the curved region at the extreme of the S-curve with resulting distortion in the output voltage.

The steady or average voltage output from a discriminator is zero if the IF is centered on the zero origin of the frequency input, even when frequency modulation is present. An offset in the IF, such as caused by a drift in oscillator frequency, results in a dc voltage offset. This can be used to bias a varactor diode forming part of the oscillator tuning in such a way as to reduce the frequency shift. The automatic frequency control obtained in this way is a further illustration of the VCO principles discussed in Section 6.6.

10.16 Amplitude Limiters

Amplitude limiters are amplifier circuits that are used to eliminate amplitude modulation and amplitude-modulated noise from received FM signals before detection. This step is necessary because most of the discriminator circuits respond in some degree to amplitude variations in the FM signal, introducing an unwanted source of noise.

The limiter works in such a manner that limiting begins when the input signal becomes larger than that required to drive the amplifier from cutoff to saturation (that is, across its active range). Figure 10.16.1(a) shows such an amplifier. It uses a bipolar transistor and double-tuned IF transformers, with the output transformer supplying the detector circuit. Resistors R_1 , R_2 , and R_E provide dc bias under zero signal conditions to maintain the transistor in the active region. R_{dc} acts to reduce the effective supply voltage and limit the saturation value of the collector current to a low value so that saturation will occur at a low-input-signal level.

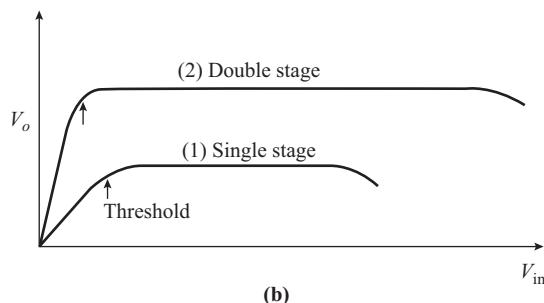
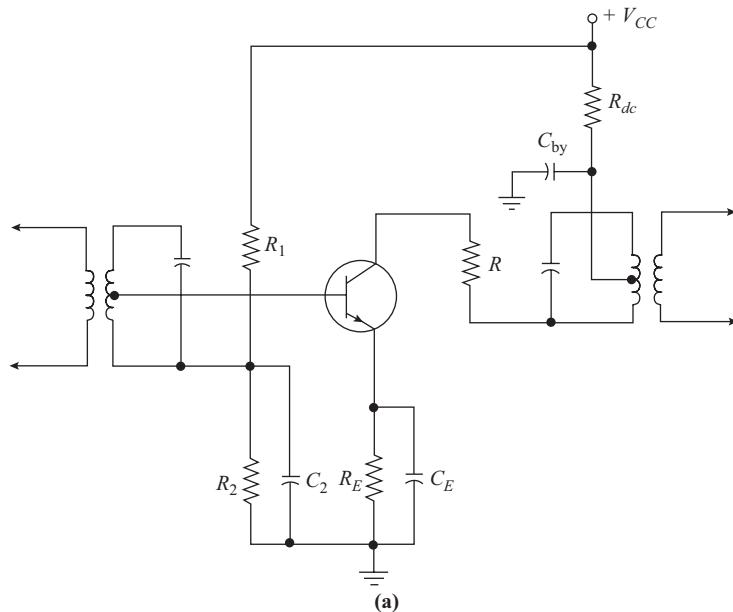


Figure 10.16.1 Amplitude limited circuit: (a) schematic; (b) limited response showing the expanded range of a double-stage limiter.

For very low signal levels, the circuit acts as a normal class A amplifier. For larger signals, which drive the transistor into cutoff, the base-leak circuit formed by C_2 and R_2 charges up, driving the operating point toward cutoff. At the same time, the positive peaks drive the base current beyond the saturation point, and the collector-current waveform has both the positive and negative peaks clipped off to produce an approximately rectangular waveform. The fundamental of this rectangular current waveform drives the tuned circuit of the output transformer in the “flywheel” manner typical of the class C amplifier. The point at which the signal-input range exceeds the active range of the transistor forms the threshold of the limiter, beyond which limiting action takes place. Further increase of input signal does not significantly increase the collector-current magnitude, and its fundamental component remains almost constant. If a very large increase in input occurs, such as by a large noise pulse, the bias circuit will drive even further into cutoff, and the conduction angle of the amplifier will decrease. The limiter is then said to have been “captured” by the noise and will not be released until the bias capacitor discharges. This effect puts an upper limit on the limiting range, and it means that if large noise pulses are received, they could cause the limiter to respond and reduce the desired input signal. The shape of the limiter response is indicated in Fig. 10.16.1(b).

If the receiver is to respond to a very large range of input signals and noise signals, it may be necessary to increase the range of the limiter. This may be accomplished in two ways. First, for large-signal conditions, AGC may be provided so that the limiter is not driven beyond its limiting range. Second, a second stage of limiting may be provided. Two-stage limiting is much more effective and provides very good protection against large-amplitude noise pulses. Since there is more gain in the limiter circuits, the threshold at which limiting occurs is lower, and less preamplifier gain is required.

It was stated that most types of FM detectors require amplitude limiting in order to function properly. The ratio detector (see Section 10.14) is an exception to this, because it has a degree of inherent limiting built into it. In critical applications it is still necessary to provide additional limiting, but the ratio detector performs well enough otherwise.

10.17 Noise in FM Systems

Noise in an FM receiver can be referred to the input as shown in Fig. 10.17.1(a). With FM receivers the limiter stages described in Section 10.16 help to reduce impulse-type noise, so FM has this advantage over AM.

Another advantage arises from the nature of the noise modulation process. The noise at the receiver input cannot directly frequency modulate the incoming carrier since its frequency is fixed at a distant transmitter, which may in fact be crystal controlled. The noise phase modulates the carrier at the receiver and, as will be shown, this leads to a reduction in output compared to the AM situation.

An important advantage with FM reception is that an improvement in signal-to-noise ratio can be achieved by increasing the frequency deviation. This requires an increased bandwidth, but at least the option is there for an exchange of bandwidth for signal-to-noise ratio. This aspect of FM reception will be explained in detail in this section.

Figure 10.17.1(b) shows the signal and noise voltages at the FM detector. The noise, having passed through a band-pass filter, can be represented by narrow-band noise as described in Section 4.20. This will be analyzed shortly. At this point it need only be noted that the power spectral density as given by Eq. (4.20.1) is kT_s , and hence the available noise power at the detector input for a bandwidth W is $P_{n\text{REF}} = kT_s W$. The rms signal voltage is $E_c = E_{c\text{ max}}/\sqrt{2}$, and therefore the signal power at the detector input is $P_R = E_{c\text{ max}}^2/(8R_s)$. The reference signal-to-noise ratio, (introduced in Eq. [8.14.13]), is for the FM case

$$\left(\frac{S}{N}\right)_{\text{REF}} = \frac{P_R}{P_{n\text{REF}}}$$

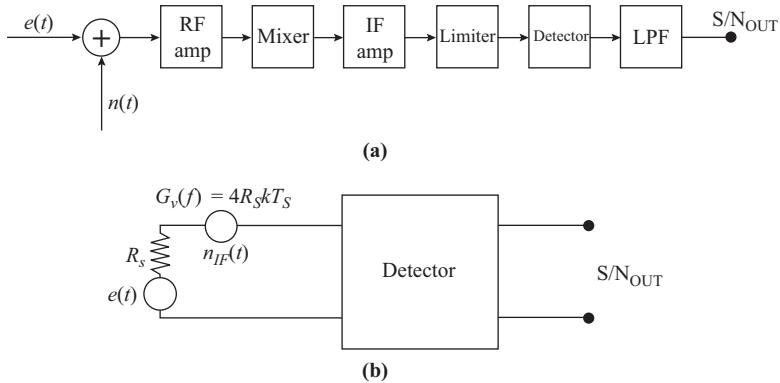


Figure 10.17.1 (a) Block schematic for an FM receiver, (b) Signal and noise at the detector.

$$= \frac{E_c^2 \max}{8R_S k T_s W} \quad (10.17.1)$$

Here again it is emphasized that, although the reference signal-to-noise ratio is determined at the input side of the detector, the bandwidth W is that determined by the LPF at the output side.

Consider now the signal output from the band-pass filter in the absence of noise, when a sinusoidally modulated carrier is received. The instantaneous frequency is

$$f_i = f_{IF} + \Delta f \cos \omega_m t \quad (10.17.2)$$

The FM detector converts this to a signal output given by

$$v_m(t) = C\Delta \cos \omega_m t \quad (10.17.3)$$

where C , a constant, is the frequency-to-voltage coefficient of the detector. The rms voltage output is $E_m = C\Delta f / \sqrt{2}$, and hence the available power output is

$$P_{so} = \frac{(C\Delta f)^2}{8R_{out}} \quad (10.17.4)$$

The next part of the analysis requires finding the noise at the output of the detector. The noise voltage output from a band-pass system is given by Eq. (4.20.3), which is rewritten here as

$$n_{IF}(t) = n_1(t) \cos \omega_{IF} t - n_Q(t) \sin \omega_{IF} t \quad (10.17.5)$$

The incoming carrier in this case is frequency modulated and can be written as

$$e(t) = E_c \max \cos(\omega_{IF} t + \phi_m(t)) \quad (10.17.6)$$

where $\phi_m(t)$ is the equivalent phase modulation produced by the signal (see Section 10.9). An analysis of the detector output when noise and modulating signal are present *simultaneously* is very complicated, but fortunately the additional noise terms in the output, resulting from the interaction of $\phi_m(t)$ and the noise modulation, lie outside the baseband. As a result, the noise analysis can be made assuming that an unmodulated carrier is received. In this case, the input to the detector is

$$e_{det}(t) = e_c(t) + n_{IF}(t)$$

$$= (E_{c \max} + n_1(t)) \cos \omega_{IFt} - n_Q(t) \sin \omega_{IFt} \quad (10.17.7)$$

This should be compared with Eq. (8.14.3) for the AM case. As with the AM case, the waveform can be expressed in an equivalent form: $e_{\text{det}}(t) = R(t) \cos (\omega_{IFt} + \psi(t))$; but in this situation it is the phase angle $\psi(t)$, rather than the envelope $R(t)$, that is of interest. A trigonometric analysis of the equation gives

$$\psi(t) = \tan^{-1} \frac{n_Q(t)}{E_{c \max} + n_I(t)} \quad (10.17.8)$$

Also, as in the AM case, the analysis will be limited to the situation where the carrier amplitude is much greater than the noise voltage for most of the time. Under these circumstances the noise phase angle becomes

$$\begin{aligned} \psi(t) &\cong \tan^{-1} \frac{n_Q(t)}{E_{c \max}} \\ &\cong \frac{n_Q(t)}{E_{c \max}} \end{aligned} \quad (10.17.9)$$

The total carrier angle is $\theta(t) = \omega_{IFt} + \psi(t)$, and from Eq. (10.9.2) the equivalent frequency modulation is

$$\begin{aligned} f_{\text{eq}}(t) &= f_{IF} + \frac{1}{2\pi} \frac{d\psi(t)}{dt} \\ &= f_{IF} + \frac{\dot{n}_Q(t)}{2\pi E_{c \max}} \end{aligned} \quad (10.17.10)$$

where the dot notation is used for the differential coefficient. The noise voltage output is therefore

$$v_n(t) = \frac{C \dot{n}_Q(t)}{2\pi E_{c \max}} \quad (10.17.11)$$

To find the noise power output, it is best to work in the frequency domain. A result of Fourier analysis shows that, if a voltage waveform $v(t)$ has a power spectral density $G(f)$, then the power spectral density for $\dot{v}(t)$ is $\omega^2 G(f)$. Equation (4.20.4) gives the spectral density for $\dot{n}_Q(t)$ as $2kT_s$, and hence the spectral density for $n_Q(t)$ is $\omega^2 2kT_s$. Combining this with Eq. (10.17.11) and simplifying gives, for the power spectral density of $v_n(t)$,

$$G_v(f) = \frac{C^2 f^2 2kT_s}{E_{c \max}^2} \quad (10.17.12)$$

From the definition of power spectral density (see Section 2.17), the noise power output is

$$\begin{aligned} P_{no} &= \int_0^W G_v(f) df \\ &= \frac{W^3 C^2 2kT_s}{3 E_{c \max}^2} \end{aligned} \quad (10.17.13)$$

Combining this with Eq. (10.17.4) and simplifying gives, for the output signal-to-noise ratio,

$$\left(\frac{S}{N} \right)_o = \frac{P_{so}}{P_{no}}$$

$$= \frac{3\Delta f^2 E_c^2 \max}{16R_{\text{out}} k T_s W^3} \quad (10.17.14)$$

The signal-to-noise figure of merit introduced in Section 8.14 is, for FM,

$$\begin{aligned} R_{\text{FM}} &= \frac{(\text{S/N})_o}{(\text{S/N})_{\text{REF}}} \\ &= 1.5 \cdot \left(\frac{\Delta f}{W} \right)^2 \frac{R_s}{R_{\text{out}}} \\ &= 1.5 \beta_W^2 \frac{R_s}{R_{\text{out}}} \end{aligned} \quad (10.17.15)$$

Here β_W is the modulation index calculated for the highest baseband frequency W . Note that this is not the same in general as the modulation index for the signal, which is $\beta = \Delta f/f_m$.

The information on FM signal-to-noise ratio is often presented in another way, using the carrier-to-noise ratio as the input parameter at the detector. The carrier-to-noise ratio is similar to the $(\text{S/N})_{\text{REF}}$ ratio except that the total IF bandwidth is used rather than W . Denoting the carrier-to-noise ratio by (C/N) , Eq. (10.17.1) is modified to

$$\left(\frac{\text{C}}{\text{N}} \right) = \frac{E_c^2 \max}{8R_s k T_s B_{\text{IF}}} \quad (10.17.16)$$

Also, applying Carson's rule, $B_{\text{IF}} = 2(\beta_W + 1)W$ gives

$$\left(\frac{\text{C}}{\text{N}} \right) = \frac{E_c^2 \max}{16R_s k T_s (\beta_W + 1)W} \quad (10.17.17)$$

The ratio of output signal-to-noise ratio to carrier-to-noise ratio is known as the receiver (or detector) processing gain and is

$$\begin{aligned} R_{\text{PG}} &= \frac{(\text{S/N})_o}{(\text{C/N})} \\ &= 3\beta_W^2 (\beta_W + 1) \frac{R_s}{R_{\text{out}}} \end{aligned} \quad (10.17.18)$$

It is important to keep in mind that this result applies only when the carrier is much larger than the noise, and the derivation is for a sinusoidally modulated wave. When the carrier level is reduced below what is termed the *threshold level*, it is found that the output signal-to-noise ratio worsens very rapidly. The reason for this is that the noise phase modulation introduces random jumps of 2π radians, which result in output noise spikes. This can be illustrated by means of the phasor diagram shown in Fig. 10.17.2.

Figure 10.17.2(a) shows $E_c \max \gg A_n(t)$, where the phase modulation $\psi(t)$ of the resultant phasor does not show excessive variation even when the noise phase angle $\theta(t)$ varies through 360° . When, however, $E_c \max < A_n(t)$, the resultant phasor may rotate around the origin as shown in Fig. 10.17.2(b), introducing an additional 2π radians phase change in $\psi(t)$. Figure 10.17.2(c) shows the phase change ψ and the derivative $d\psi/d\phi$ both as functions of ϕ for $E_c \max = 10 \text{ A}$, and Figure 10.17.2(d) shows the same functions for $E_c \ max = 0.9 \text{ A}$, where the spike in the derivative is evident. In reality, A and ϕ are random functions of time, and it is only when these combine to rotate the resultant phasor around the origin that a noise spike occurs. The likelihood of this happening increases as the carrier gets smaller relative to the noise. (For the functional relationships between ψ and ϕ , see Problems 10.54 and 10.55.)

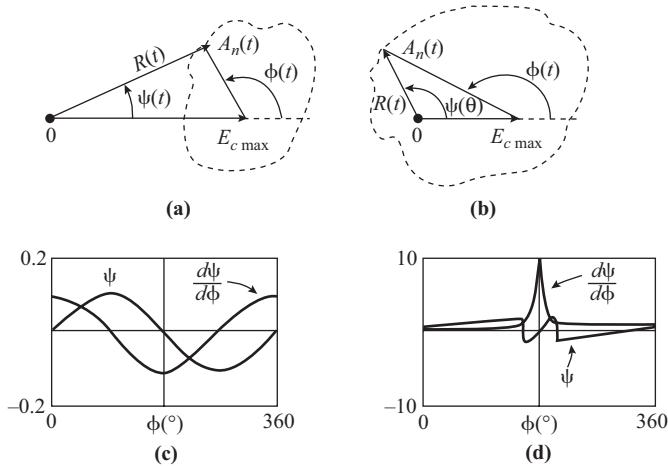


Figure 10.17.2 Phasor diagram for (a) $E_c \max \gg A_n(t)$ and (b) $E_c \max < A_n(t)$. Phase change and its derivative for (c) $E_c \max = 10 \text{ A}$ and (d) $E_c \max = 9 \text{ A}$.

The threshold level is formally defined as the C/N ratio at which the output S/N ratio is 1 dB below the value predicted by Eq. (10.17.18). It is found to be relatively insensitive to the modulation index and occurs at $C/N \approx 10 \text{ dB}$. Thus for satisfactory reception the carrier-to-noise ratio should be above this value. Equation (10.17.18) in decibels is, for $R_s = R_{\text{out}}$,

$$\left(\frac{S}{N}\right)_o \text{dB} = \left(\frac{C}{N}\right) \text{dB} + (R_{PG}) \text{dB} \quad (10.17.19)$$

Power ratios are used for the decibel calculations. Figure 10.17.3 shows a plot of output S/N as a function of C/N in decibels for $\beta_W = 5$. The *threshold margin* is the difference between the threshold level and the actual operating point, as shown in Fig. 10.17.3.

10.18 Pre-emphasis and De-emphasis

Equation (10.17.12) shows the quadratic dependence of the noise power spectral density as a function of frequency. This can be expressed as a decibel equation by taking a suitable reference, which will be chosen as the spectral density at $f = 1 \text{ Hz}$.

$$\frac{G_v(f)}{G_v(1 \text{ Hz})} = \left(\frac{f}{1 \text{ Hz}}\right)^2$$

or

$$10 \log \left(\frac{G_v(f)}{G_v(1 \text{ Hz})} \right) = 20 \log \left(\frac{f}{1 \text{ Hz}} \right)$$

or

$$[G_v(f)] \text{ dB} = 20 \log f \quad (10.18.1)$$

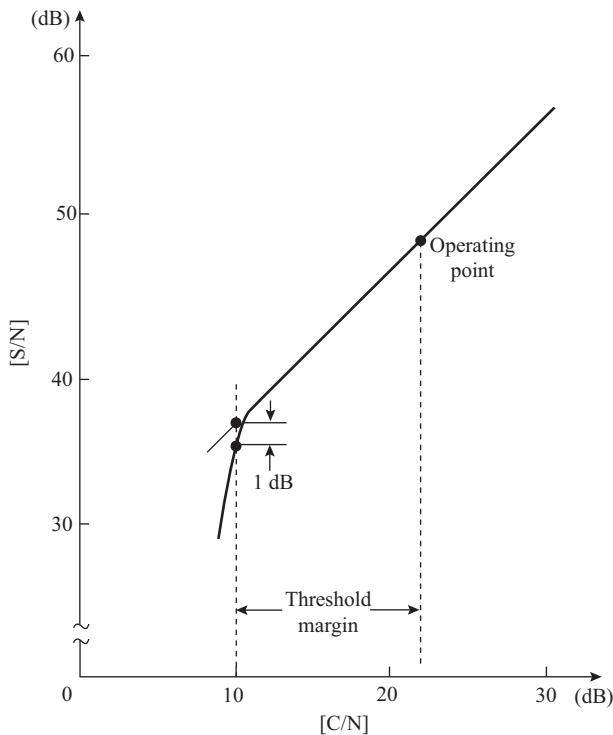


Figure 10.17.3 Output S/N versus input C/N for $\beta_W = 5$.

The noise spectral density is seen to increase at the rate of 6 dB per octave. The low noise levels at low frequencies show why the signal-to-average noise power for FM is inherently better than AM, for which the noise spectral density is constant. However, a disadvantage arises in that, for speech, the clarity of speech (or what is termed the articulation efficiency) depends on the high-frequency content of the speech waveform, and the rising noise characteristic of FM tends to mask this.

The situation can be improved by including a *de-emphasis* network following the FM detector, which attenuates the noise at the rate of 6 dB per octave. Since the network will also attenuate the modulating signal spectrum in a similar fashion, it is necessary to include a compensating network at the transmitter that pre-emphasizes the modulating spectrum by 6 dB per octave, and naturally this is known as a *pre-emphasis network*. For broadcast applications it is essential that manufacturers work to a common pre-emphasis/de-emphasis set of characteristics. Figure 10.18.1 shows typical pre-emphasis and de-emphasis networks, along with their relative responses. The values used in calculating the response curves are given in Problems 10.60 and 10.62.

10.19 FM Broadcast Receivers

FM commercial broadcasting in North America takes place in the VHF band between 88 and 108 MHz. Within this band, allotted frequencies are spaced 200 kHz apart and are allowed a maximum frequency deviation of ± 75 kHz around the carrier frequency. Propagation at VHF is restricted to line of sight, and coverage is usually only for a radius of about 50 miles around the transmitter location. The programs broadcast on these channels in the past have been mostly music, and the basic modulating frequency bandwidth is 15 kHz, as opposed to the 5 kHz used on AM stations.

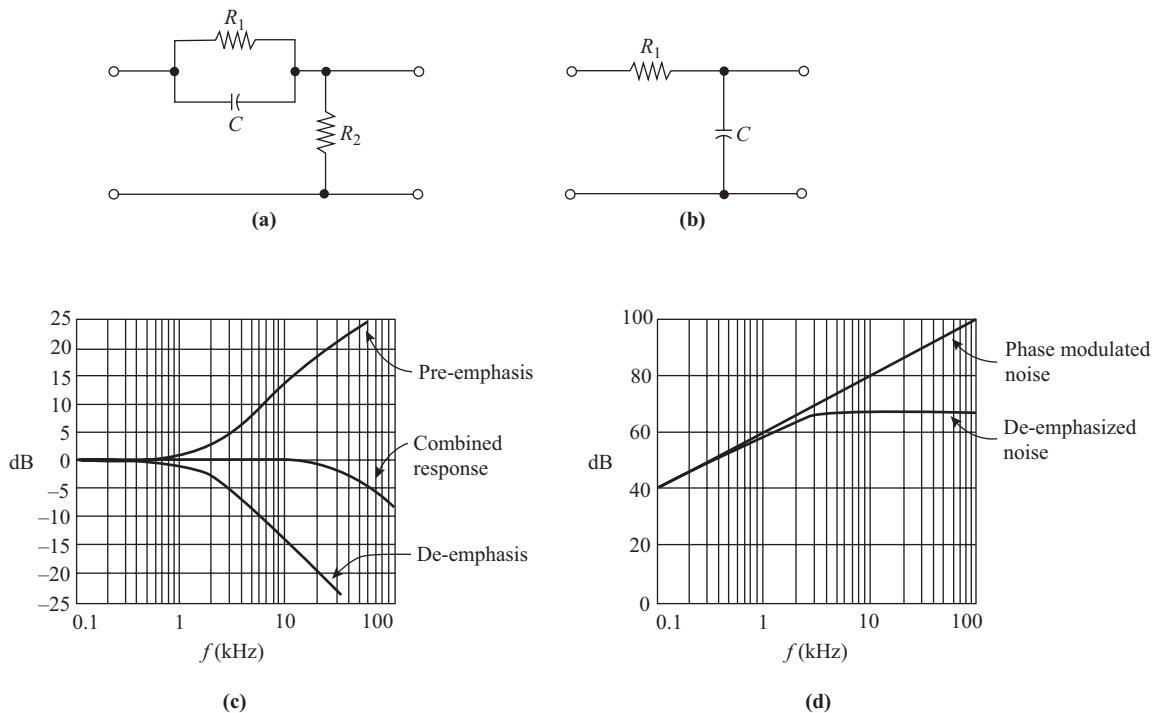


Figure 10.18.1 (a) Standard pre-emphasis network having a lower corner frequency of 2.12 kHz and (b) a typical de-emphasis network. (c) Pre-emphasis and de-emphasis curves and the combined response. (d) Noise before and after de-emphasis. All curves are referenced to their 1-Hz values.

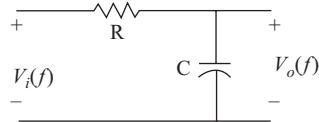
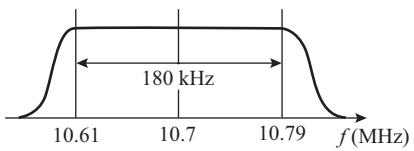
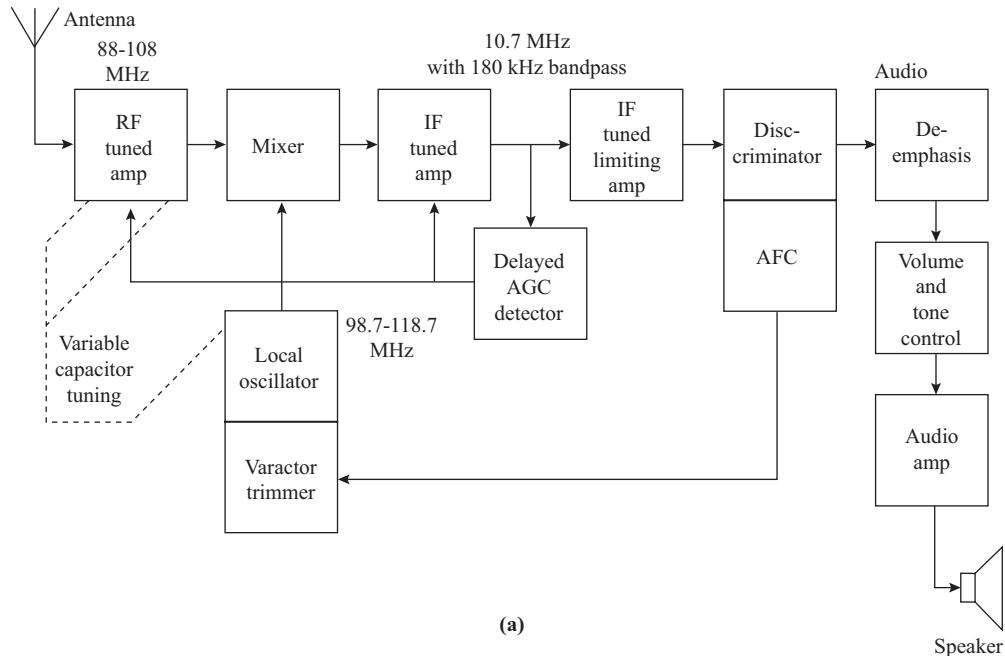
Figure 10.19.1(a) shows the block schematic of a typical FM broadcast receiver of the monaural or single-channel variety. It is a superheterodyne circuit, with a tuned RF amplifier so that maximum sensitivity is typically between $10\mu\text{V}$. The RF-stage-tuned circuits and the local oscillator are tuned by a three-ganged variable capacitor controlled from a panel knob. The oscillator frequency can be varied from 98.7 to 118.7 MHz, yielding an intermediate frequency (IF) of 10.7 MHz.

The IF amplifier section is comprised of several high-gain stages, of which one or more are amplitude limiters. The schematic shown here has one high-gain nonlimiting input stage, followed by one amplitude-limiting stage. All stages are tuned to give the desired band-pass characteristic, which is shown in Fig. 10.19.1(b). This is centered on 10.7 MHz and has a 180-kHz bandwidth to pass the desired signal. Amplitude limiting is usually arranged to have an onset threshold of about 1 mV at the limiting-stage input, corresponding to the quieting level of input signal, which may be set at $10\mu\text{V}$ or lower.

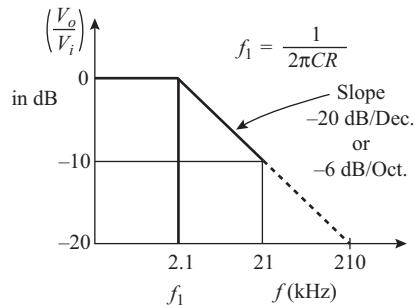
The FM detector may be any one of several types of FM detectors described in Section 10.14, perhaps incorporating automatic frequency control as described in Section 10.15.

AGC is not usually supplied in less expensive FM receivers. These receivers have sufficient amplifier gain so that the last stage operates in saturation for most signals to obtain the necessary amplitude limiting for good detection. AGC may be provided to control the RF and early IF stages so that saturation of the non-limiting IF stages does not occur on strong signals.

Figure 10.19.1 shows the block diagram of an FM receiver that uses AGC. In this case, a sample of the IF signal is extracted just before the input to the limiting IF amplifier. This sample is applied to a special



(b)



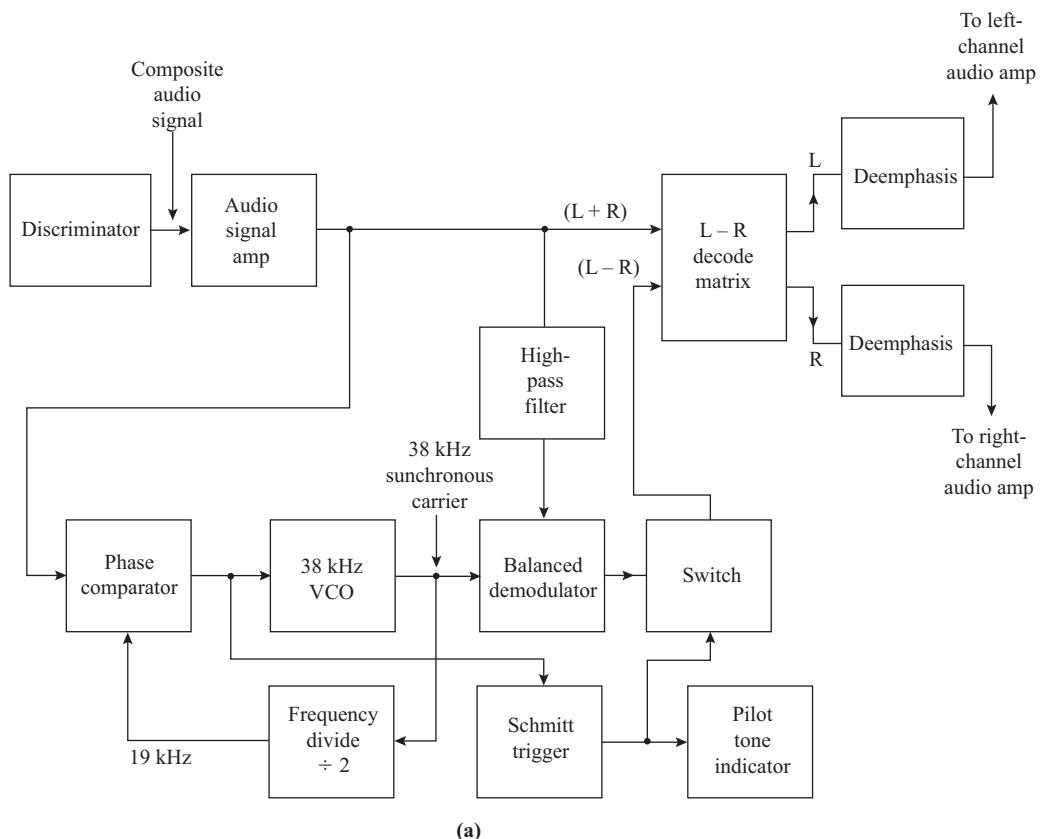
(c)

Figure 10.19.1 (a) Block diagram of a typical FM monaural broadcast receiver, (b) IF amplifier band-pass characteristics, (c) De-emphasis network and transfer function.

detector used only to obtain the AGC signal and is a peak amplitude detector similar to that in Fig. 7.9.1. The derived AGC signal is applied to control the RF preamplifier and the first IF amplifier. Its time constant is similar to that used in AM receivers.

10.20 FM Stereo Receivers

All new FM broadcast receivers are being built with provision for receiving stereo, or two-channel broadcasts. The left (L) and right (R) channel signals from the program material are combined to form two different signals, one of which is the left-plus-right signal and one of which is the left-minus-right signal. The (L - R) signal is double-sideband suppressed carrier (DSBSC) modulated about a carrier frequency of 38 kHz, with the LSB in the 23- to 38-kHz slot and the USB in the 38- to 53-kHz slot. The (L + R) signal is placed directly in the 0- to 15-kHz slot, and a pilot carrier at 19 kHz is added to synchronize the demodulator at the receiver. The composite signal spectrum is shown in Fig. 10.20.1(b).



(a)

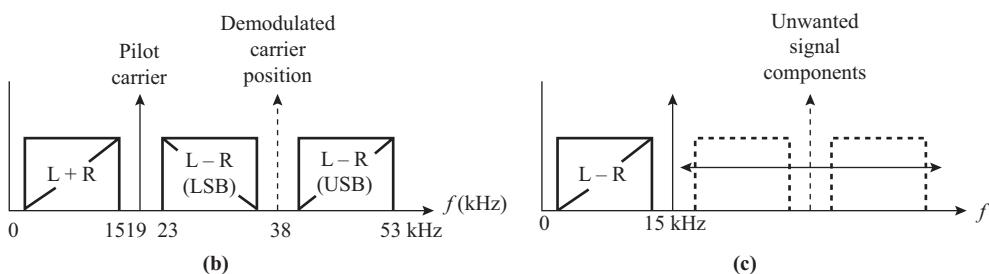


Figure 10.20.1 (a) Block diagram of an FM signal stereo channel decoder system; (b) spectrum of the composite audio signal from the FM detector; (c) spectrum of the demodulated left minus right (L - R) signal after removal of higher frequencies.

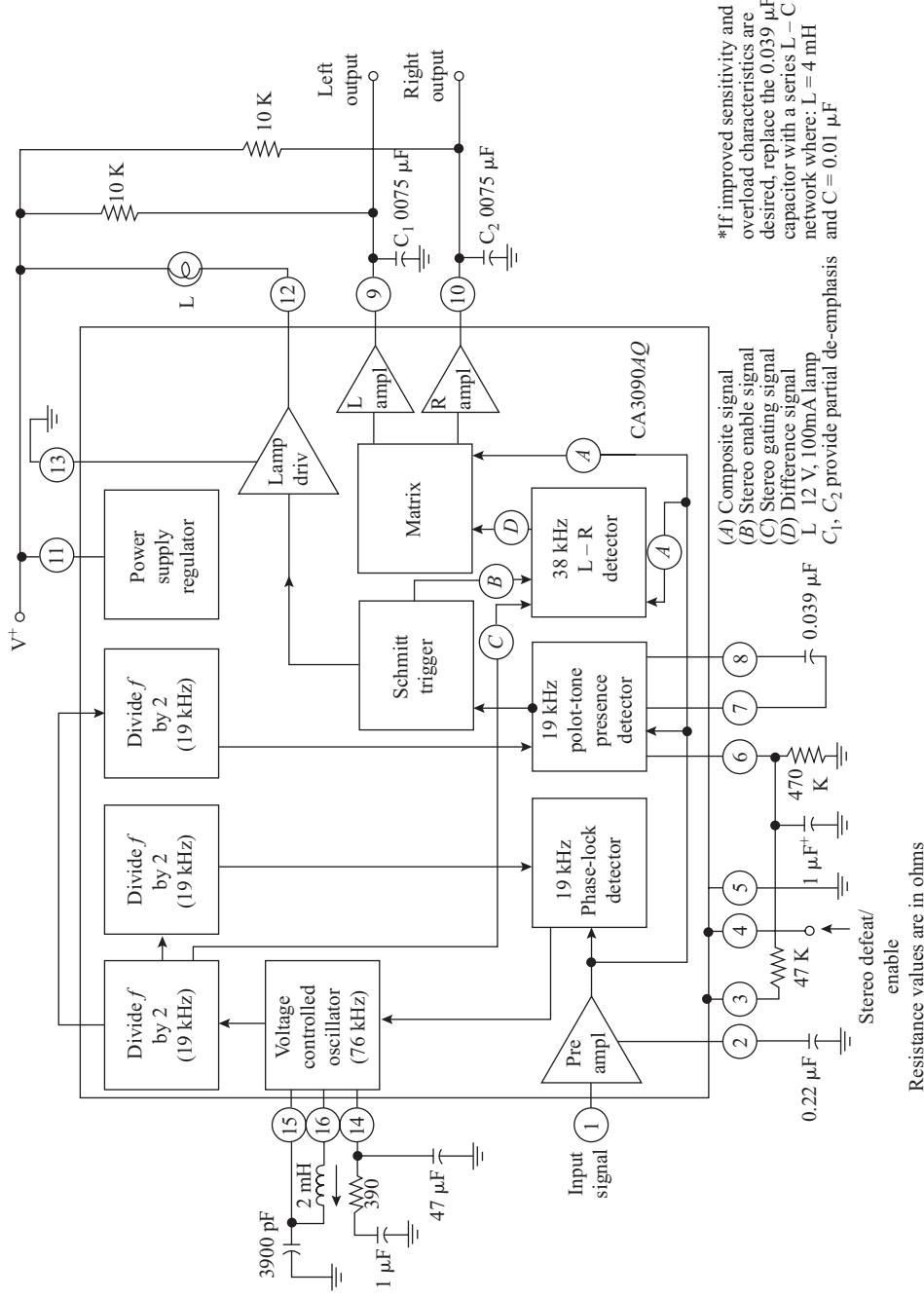


Figure 10.20.2 Functional block diagram of a typical stereo decoder IC chip, the RCA CA3090AQ. The circuit is mounted in a 14-pin dual-in-line package (DIP). (Courtesy of RCA, Data File No. 684.)

Figure 10.20.1(a) shows the block schematic of a stereo channel decoder circuit. The output from the FM detector is a composite audio signal containing the frequency-multiplexed ($L + R$) and ($L - R$) signals and the 19-kHz pilot tone. This composite signal is applied directly to the input of the decode matrix.

The composite audio signal is also applied to one input of a phase-error detector circuit, which is part of a phase locked loop 38-kHz oscillator. The output drives the 38-kHz voltage-controlled oscillator, whose output provides the synchronous carrier for the demodulator. The oscillator output is also frequency divided by 2 (in a counter circuit) and applied to the other input of the phase comparator to close the phase locked loop. The phase-error signal is also passed to a Schmitt trigger circuit, which drives an indicator lamp on the panel that lights when the error signal goes to zero, indicating the presence of a synchronizing input signal (the 19-kHz pilot tone).

The outputs from the 38-kHz oscillator and the filtered composite audio signals are applied to the balanced demodulator, whose output is the ($L - R$) channel. The ($L + R$) and ($L - R$) signals are passed through a matrix circuit that separates the L and R signals from each other. These are passed through de-emphasis networks and low-pass filters to remove unwanted high-frequency components and are then passed to the two channel audio amplifiers and speakers. On reception of a monaural signal, the pilot-tone indicator circuit goes off, indicating the absence of pilot tone, and closes the switch to disable the ($L - R$) input to the matrix. The ($L + R$) signal is passed through the matrix to both outputs. An ordinary monaural receiver tuned to a stereo signal would produce only the ($L + R$) signal, since all frequencies above 15 kHz are removed by filtering, and no demodulator circuitry is present. Thus the stereo signal is compatible with the monaural receivers.

Much of the circuitry for FM receivers is now available in integrated circuit form. One of the first pieces of circuitry to be produced in ICs was the stereo decoder. Figure 10.20.2 shows a typical IC decoder chip, the RCA CA3090AQ. This chip contains all the circuits necessary to accomplish the functions of Fig. 10.20.1(a) and comes with the chip mounted in a 14-pin DIP. Only a few external components are needed to make the circuit work. These include bias resistors and de-emphasis capacitors on the output, reactive components (L and C) for the VCO, and some bypass capacitors and bias resistors. The rest is built into the chip.

PROBLEMS

- 10.1.** A carrier wave has an unmodulated frequency of 3 MHz. Sketch the instantaneous frequency-time curve when it is frequency modulated by a 1000-Hz sawtooth waveform, which varies 10 V peak-to-peak about a zero dc value. The frequency deviation constant is 5 kHz per volt.
- 10.2.** Sketch the modulation component of the angle $\theta(t)$ as a function of time for the waveform in Problem 10.1.
- 10.3.** A carrier is frequency modulated by a ramp voltage that starts at zero and rises linearly to 7 V in 3 ms, when it abruptly returns to zero. Given that the frequency deviation constant is 5 kHz per volt and the unmodulated carrier frequency is 100 kHz, plot the modulated angle $\theta(t)$ over a time base of 5 ms, starting 1 ms before modulation is applied.
- 10.4.** A carrier is frequency modulated by a sinusoidal signal of 15 V peak and frequency of 3 kHz, the frequency deviation constant being 1 kHz/V. Calculate (a) the peak frequency deviation and (b) the modulating index.
- 10.5.** A carrier of amplitude 5 V and frequency 90 MHz is frequency modulated by a sinusoidal voltage of amplitude 5 V and frequency of 15 kHz. The frequency deviation constant is 1 kHz/V. Sketch the spectrum of the modulated wave.

- 10.6.** A sinusoidal carrier has an amplitude of 5 V and frequency of 25 kHz. It is frequency modulated by a sinusoidal voltage of amplitude 5 V and frequency of 1000 Hz. The frequency deviation constant is 2.5 kHz/V. Plot accurately to scale the modulated waveform, showing two complete cycles of the modulating wave, and determine the modulation index.
- 10.7.** Draw accurately to scale the frequency spectrum for the modulated wave of Problem 10.6, stating clearly the amplitude level beyond which side frequencies are ignored.
- 10.8.** An FM carrier is limited to a maximum deviation of 75 kHz. Compare the bandwidths when the modulating signal is sinusoidal of frequency (a) 100 Hz, (b) 1 kHz, and (c) 10 kHz, the carrier reaching maximum deviation in each case. State how you determine the bandwidth limits.
- 10.9.** Explain why it is that in some circumstances the bandwidth required for a sinusoidally frequency modulated carrier is greater than twice the modulating frequency. Determine the bandwidth occupied by a sinusoidally frequency modulated carrier for which the modulating index is 2.4 and $f_m = 3$ kHz, stating what limits are used to determine bandwidth.
- 10.10.** Using the results of Table 10.4.1 and Eq. (10.5.4), verify that the power remains constant for an FM wave for modulation indexes (a) 0.25, (b) 2.4, (c) 5.5, and (d) 7.
- 10.11.** State clearly the difference between *modulation index* and *deviation ratio*.
- 10.12.** State Carson's rule. A 1-kHz square wave is used to frequency modulate a carrier, producing a peak deviation of 75 kHz. Assuming that the square-wave harmonics up to and including the eleventh need to be taken into account, calculate the deviation ratio and, using Carson's rule, the bandwidth.
- 10.13.** Given that the peak deviation is 75 kHz, determine, using Carson's rule, the bandwidths required for sinusoidal FM indexes of (a) 0.25, (b) 2.4, (c) 5.5, and (d) 7. Compare the results with those obtained using Table 10.4.1.
- 10.14.** Explain what is meant by phase modulation. The phase deviation constant in a phase modulation system is 0.01 rad/V. Calculate the peak phase deviation when a sinusoidal modulating signal of 10 V peak is applied. Does the peak phase deviation depend on the modulating frequency?
- 10.15.** A sinusoidal carrier has an amplitude of 10 V and frequency of 20 kHz. It is phase modulated by a ramp voltage, which, starting from an arbitrary zero time reference, is zero for the first five RF cycles of carrier and then rises abruptly to 3 V. The phase deviation constant is 0.5 rad/V. On the same set of axes, plot accurately to scale the modulating and the modulated voltage waveforms, showing 10 complete cycles of the modulated carrier wave.
- 10.16.** A ramp voltage is used to phase modulate a 20-MHz carrier. The ramp rises linearly from zero to 5 V in 250 μ s, when it abruptly returns to zero. Calculate the instantaneous frequency over the modulating period. The phase deviation constant is π rad/V.
- 10.17.** A sinusoidal signal of 10 V peak is used to phase modulate a carrier, the phase deviation constant being 0.7 rad/V. Determine (a) the peak phase deviation, and (b) the equivalent peak frequency deviation for a modulating frequency of 12 kHz. Repeat the calculations for the modulating frequency increased to 15 kHz, all other factors remaining unchanged.
- 10.18.** A 3000-Hz sinusoidal signal of 5 V peak is used to phase modulate a carrier, the phase deviation constant being 1 rad/V. Determine the practical bandwidth required for the modulated signal.
- 10.19.** A binary sequence 101101 is used to phase modulate a carrier, the bit period being equal to 2.5 times the carrier period. Given that the modulated carrier amplitude is 1 V peak, draw accurately to scale the modulated waveform showing the full modulating sequence.
- 10.20.** A linear graded *pn* junction diode has a zero bias capacitance of 20 pF and a built-in potential of 0.5 V. Plot the variation of capacitance as a function of reverse-bias voltage for the range from 0 to -10 V.

- 10.21.** An abrupt *pn* junction diode has a zero bias capacitance of 20 pF and a built-in potential of 0.5 V. Plot the variation of capacitance as a function of reverse-bias voltage for the range from 0 to -10 V.
- 10.22.** For the varactor diode in the modulator circuit of Fig. 10.12.1, $C_o = 20$ pF, $\psi = 0.5$ V, and $\alpha = 0.5$. The diode is operated around a fixed bias of -7 V. The series capacitance of the oscillator tuned circuit, excluding the varactor diode, is $C_{ser} = 17.65$ pF. With the diode in circuit, but unmodulated, the oscillator frequency is 2.5 MHz. Plot the transfer characteristic and find the linear correlation coefficient for a modulation voltage which varies between ± 1 V about the fixed bias. Find also the frequency deviation constant.
- 10.23.** For the varactor diode modulator circuit of Fig. 10.12.1, $C'_1 = C'_2 = 250$ pF, $C_3 = 25$ pF, and $L = 750$ μ H. For the varactor diode, $C_o = 15$ pF, $\psi = 0.5$ V, and $\alpha = 0.5$. The diode is operated around a fixed bias of -7 V. Plot the transfer characteristic and find the linear correlation coefficient. Find also the frequency deviation constant. Assume the modulation voltage varies between ± 1 V about the fixed bias.
- 10.24.** Plot the ϕ/v_m curve for the phase modulator of Fig. 10.12.4 for which $Q = 30$ and $C_{ser} = 20$ pF. For the varactor diode, $C_o = 20$ pF at $V = -15$ V, $\psi = 0.5$ V, and $\alpha = 0.5$. Determine the linear correlation coefficient and phase shift constant.
- 10.25.** For the phase modulator circuit of Fig. 10.12.4, the fixed series capacitance is 25 pF and the inductance is 900 μ H. For the varactor diode $C_o = 20$ pF, $\alpha = 0.5$, and $\psi = 0.5$ V. The diode is operated around a fixed bias of -15 V. The circuit *Q*-factor is 30. Plot the change in phase angle as a function of modulating voltage over the range ± 1 V about the fixed bias.
- 10.26.** Using Mathcad or otherwise, determine the phase shift coefficient and the linear correlation coefficient for the modulator of Problem 10.25.
- 10.27.** Values for a JFET varactor modulator are $R_1 = 100$ k Ω and $C_2 = 5$ pF. For the JFET, $V_p = -15$ V, $I_{DSS} = 15$ mA, and the fixed bias is $V_B = -1$ V. The oscillator inductance is $L = 1$ mH, and with the modulator connected but no modulation applied, the frequency is 4 MHz. The modulating voltage varies continuously between ± 1 V. Assuming that $\omega\tau = 10$ meets the requirement that $\omega\tau \gg 1$, determine the minimum frequency at which the circuit should be used. Verify that the condition $g_m |Z_2| \gg 1$ is met. Plot the transfer function and determine the linear correlation coefficient and the frequency deviation constant.
- 10.28.** On a common set of axes, plot the imaginary part $\text{Im}(Y)$ of Eq. (10.12.12) over the frequency range from 3 to 5 MHz in steps of 0.1 MHz. Plot separate curves for g_m values 1, 2, 3, 4, and 5 mS, with $R_1 = 1.5$ k Ω and $C_1 = 7.5$ pF in each case. Do these curves confirm that the admittance presents a capacitive susceptance? On a second common set of axes, plot $\text{Im}(Y)/\omega$ over the same frequency range and determine the slope of each curve. What conclusions can be drawn from these curves?
- 10.29.** From the data given in Problem 10.28, and using a frequency of 4 MHz and the g_m values given, calculate the equivalent capacitance (a) as determined from the susceptance part of Eq. (10.12.12) and (b) from the approximation, Eq. (10.12.15). Compare the two sets of values and comment.
- 10.30.** Given that $R_2 = 5$ k Ω and $C_1 = 25$ pF, and assuming that $\omega\tau$ has to be no greater than 0.1, determine a suitable maximum frequency for which the circuit of Fig. 10.12.6 acts as a capacitance. Plot the equivalent capacitance as a function for g_m over the range from 2.5 to 5 mS, using steps of 0.25 mS. Show that the approximation for C_{EQ} given in Table 10.12.1 is adequate for practical purposes.
- 10.31.** Given that $R_1 = 50$ k Ω and $L_2 = 300$ μ H, and assuming that $\omega\tau$ has to be no greater than 0.1, determine a suitable maximum frequency for which the circuit of Fig. 10.12.6 acts as a capacitance. Using the results of Table 10.12.1 and eq. (10.12.12), plot and compare the equivalent capacitance as a function for g_m over the range from 2.5 to 5 mS in steps of 0.25 mS. Verify that $g_m |Z_2| \gg 1$ for all values of g_m at the maximum usable frequency.

- 10.32.** Given that $L_1 = 2 \text{ mH}$ and $R_2 = 5 \text{ k}\Omega$, determine a suitable minimum frequency for which the circuit of Fig. 10.12.6 acts as an inductance that is a function of g_m . Plot the equivalent inductance as a function for g_m over the range from 2.5 to 5 mS in steps of 0.25 mS. Verify that $g_m|Z_2| \gg 1$.
- 10.33.** Given that $R_1 = 25 \text{ k}\Omega$ and $C_2 = 20 \text{ pF}$, determine a suitable minimum frequency for which the circuit of Fig. 10.12.6 acts as an inductance that is a function of g_m . Plot the equivalent inductance as a function for g_m over the range from 2.5 to 5 mS using steps of 0.25 mS. Verify that $g_m|Z_2| \gg 1$.
- 10.34.** Using Eq. (10.12.10), plot the phase angle of a parallel tuned circuit as a function of tuning capacitance for a range of $\pm 10\%$ about the value required for resonance. The circuit is resonant at a capacitance value of 45 pF. The circuit Q -factor is 30.
- 10.35.** The oscillator of Fig. 10.12.8 generates a frequency of 4 MHz, and for the phase modulator, $C = 200 \text{ pF}$. The JFET parameters are $V_p = -15 \text{ V}$, and $I_{DSS} = 15 \text{ mA}$. The fixed bias is $V_B = -1 \text{ V}$, and the modulating voltage varies by up to $\pm 1 \text{ V}$ about this. Plot the change in phase angle as a function of modulating voltage, and determine the phase deviation constant and the linear correlation coefficient.
- 10.36.** A 100-kHz carrier is frequency modulated to produce a peak deviation of 800 Hz. This FM signal is passed through a 3 by 3 by 4 frequency multiplier chain, the output of which is mixed with an oscillator signal and the difference frequency taken as the new output. Determine the frequency of the oscillator required to produce a 100-kHz FM output and the peak deviation of the output.
- 10.37.** A 3-V, 1-MHz carrier is phase modulated, and over one cycle of the carrier, the modulating voltage may be assumed constant at 1 V. On a common set of axes, plot and compare one cycle of the modulated waveform derived using Eqs. (10.13.1) and 10.13.3) for a phase modulation constant $K = 0.5 \text{ rad/V}$. Repeat for $K = 0.5 \text{ rad/V}$.
- 10.38.** An Armstrong FM modulator uses a primary oscillator of 1 MHz, and the maximum phase deviation for good linearity is limited to 10° with sinusoidal modulation. What is the corresponding peak frequency deviation at a modulating frequency of 100 Hz? If the transmitted signal is to have a maximum peak frequency deviation of 30 kHz at a carrier frequency of 120 MHz, specify the frequency multiplication factor needed, and suggest a chain of doublers and triplers to accomplish this. For the chain you suggest, specify the local oscillator frequency needed for the final frequency conversion. The minimum modulating frequency is 100 Hz.
- 10.39.** Explain the method by which two-channel stereo signals are transmitted by an FM broadcast transmitter. Suggest a way in which four-channel stereo might be broadcast.
- 10.40.** A synchronously tuned transformer has identical primary and secondary circuits for which $Q = 70$ and $C = 200 \text{ pF}$. The circuits are separately resonant at 1 MHz and are critical coupled. Plot the amplitude and phase of the voltage transfer function (VTF) as functions of the percentage change in frequency for a range of $\pm 1\%$ about the resonant frequency.
- 10.41.** The circuit of Problem 10.40 is used in a Foster-Seeley discriminator. Plot the diode voltages and the output voltage as functions of the percentage change in frequency for a range of $\pm 1\%$ about the resonant frequency.
- 10.42.** A synchronously tuned transformer has identical primary and secondary circuits for which $Q = 68$ and $C = 200 \text{ pF}$. The circuits are separately resonant at 1 MHz, and the coupling factor is $kQ = 0.9$. The transformer provides the input to a ratio detector. Plot the diode voltages and the output voltage as functions of the percentage change in frequency for a range of $\pm 1\%$ about the resonant frequency.
- 10.43.** The tuned circuit in a quadrature detector has a capacitance of 150 pF and a Q -factor of 25. Plot the magnitude and phase angle of the impedance as a function of percentage change in frequency for a $\pm 1\%$ frequency range about the IF, which is 10.7 MHz.
- 10.44.** What are the dimensions for the constant K in Eq. (10.14.2)? Show that the output voltage is given by $V_o = K I^2 |Z_p| |Z_L| \sin \phi$, where I is the current shown in Fig. 10.14.9(a).

- 10.45.** The quadrature detector of Fig. 10.14.10 uses an inductance of $L_2 = 27 \mu\text{H}$. For the tuned circuit, $C = 100 \text{ pF}$ and $Q = 20$. The IF is 10.7 MHz. For calculation purposes, assume the multiplier constant $K = 1 \text{ V}^{-1}$ and $I = 1 \text{ mA}$. Plot the variation in output voltage as a function of percentage change in frequency for a $\pm 1\%$ frequency range about the IF.
- 10.46.** A phase-locked-loop (PLL) FM detector is tuned to demodulate an FM carrier at 500 kHz. The carrier is modulated by a 3-V, 800-Hz sine wave. The frequency deviation constant of the VCO is 150 Hz/V, and for the incoming signal the frequency deviation constant is 120 Hz/V. Plot the output voltage over one cycle of the modulating waveform.
- 10.47.** For the phase-locked-loop FM detector of Problem 10.46, the free-running frequency is offset by 100 Hz from the unmodulated incoming carrier frequency. Plot the output voltage over one cycle of the modulating waveform in this case.
- 10.48.** A synchronously tuned transformer has identical primary and secondary circuits for which $Q = 65$ and $C = 200 \text{ pF}$. The circuits are separately resonant at 1 MHz and the coupling is $kQ = 0.7$. Assuming $V_1 = 1 \text{ volt}$ determine the detected output voltage when an unmodulated carrier at 1MHz is present. What should the ouput ideally be? The transformer is used in a Foster–Seeley discriminator. Plot the diode voltages and the output voltage as functions of the percentage change in frequency for a range of $\pm 1\%$ about the resonant frequency.
- 10.49.** Explain why automatic frequency control (AFC) is desirable in FM receivers. The mixer oscillator in an FM receiver drifts from its correct value such that it produces a -13-kHz error in the IF, which is at 10.7 MHz. A synchronously tuned transformer forms the input to a Foster–Seeley detector in the receiver. The transformer utilizes identical primary and secondary circuits, each having a Q-factor of 70 and tuning capacitance of 100 pF. Each circuit is resonant separately at 10.7 MHz, and the coupling is $kQ = 0.65$. Assuming the primary voltage $V_1 = 1 \text{ V}$ for computations, determine the dc offset relative to the detector output at 10.7 MHz.
- 10.50.** Explain the function of amplitude limiters in FM receivers and the advantage to be obtained in using these. Would such limiters be effective in reducing impulse noise introduced as phase modulation in the receiver?
- 10.51.** A received carrier has a peak voltage of $1 \mu\text{V}$ and the antenna (source) resistance is 50Ω . Calculate the available received power. The equivalent noise temperature of the receiver is 290 K. Calculate the available noise power spectral density. Calculate the available signal-to-noise power ratio in decibels for a noise bandwidth of 5 kHz and compare with the reference signal-to-noise ratio given by Eq. (10.17.1).
- 10.52.** The equivalent noise temperature of an FM receiving system is 450 K. The antenna resistance is 50Ω . The received signal has an unmodulated peak voltage of $5 \mu\text{V}$, and when modulated by a sine wave, the peak deviation is 150 kHz. The baseband bandwidth is 15 kHz. Calculate (a) the output signal-to-noise ratio of the system in decibels, and (b) the receiver figure of merit, also in decibels. Assume $R_{\text{OUT}} = 50 \Omega$.
- 10.53.** The equivalent noise temperature of an FM receiving system is 350K and the IF bandwidth is 200 kHz. The antenna resistance is 50Ω . Calculate the received carrier-to-noise ratio in decibels when the received signal has an unmodulated peak voltage of $5 \mu\text{V}$. When modulated by a sine wave, the peak deviation of the received signal is 150 kHz. The baseband bandwidth is 15 kHz. Calculate the ratio of SNR/CNR in decibels, where SNR is the output signal-to-noise ratio and CNR is the carrier-to-noise ratio. Assume $R_{\text{OUT}} = 50 \Omega$.
- 10.54.** Referring to Fig. 10.17.1(a), show that the noise phase modulation ψ as a function of the noise angle ϕ is given by

$$\psi(\phi) = \tan^{-1} \frac{\sin \phi}{x + \cos \phi}$$

where $x = E_c \max / A_n$.

- 10.55.** Using the result of Problem 10.54, on a common set of axes plot $\psi(\phi)$ and $d\psi(\phi)/d\phi$ as functions of ϕ for $x = 10$ for the range $0 \leq \phi \geq 360^\circ$. Repeat for $x = 0.9$. Comment on the results. (Mathcad or a similar package should be used to obtain the plots.)
- 10.56.** Explain what is meant by the FM threshold. An FM receiver has an audio bandwidth of 20 kHz. It receives an FM signal that is modulated by a 10-kHz sine wave, the peak deviation being 200 kHz. The carrier-to-noise ratio is 23 dB. Does the output signal-to-noise ratio depend on the modulating frequency? Calculate the output signal-to-noise ratio in decibels.
- 10.57.** Given that the threshold carrier-to-noise ratio for an FM system is 10 dB, calculate the threshold margin required to ensure an output signal-to-noise ratio of 56 dB when receiving a sinusoidally modulated signal for which the peak deviation is 120 kHz. The baseband bandwidth of the receiver is 15 kHz.
- 10.58.** Explain what is meant by the processing gain of an FM detector. The postdetection bandwidth of an FM receiver is 4 kHz (which may also be assumed to equal the highest modulating frequency), and the pre-detector (IF) bandwidth is 30 kHz. Determine (a) the detector processing gain in dB. (b) The receiver works with a 7-dB margin above the threshold of 10 dB. Determine the output signal-to-noise ratio in decibels.
- 10.59.** Explain why de-emphasis is used in FM reception. What is the function of pre-emphasis in an FM transmitter?
- 10.60.** For the pre-emphasis and de-emphasis networks shown in Fig. 10.18.1, the R_I , C_I time constant is 75 μ s, and the resistor values are $R_I = 10 \text{ k}\Omega$, $R_2 = 5 \text{ k}\Omega$. Determine the lower corner frequency. On a common set of axes, plot the pre- and de-emphasis characteristics (in decibels) for the frequency range from 0.1 to 100 kHz (on a logarithmic scale). On the same set of axes, plot the combined response curve.
- 10.61.** Show that the relative response for the de-emphasis network of Fig. 10.18.1(b), referred to the response at 1 Hz, is given by $(1 + j2\pi\tau)/(1 + j2\pi f\tau)$, where τ is the time constant of the network. Plot the relative response in decibels for the frequency range from 0.1 to 10 kHz (on a logarithmic scale) for $\tau = 75 \mu\text{s}$.
- 10.62.** An FM receiver uses the de-emphasis network of Fig. 10.18.1(b) with $\tau = 75 \mu\text{s}$. Plot the relative noise spectral density in decibels at the output, without and with the de-emphasis network in place, for the frequency range from 0.1 to 10 kHz (on a logarithmic scale).
- 10.63.** Explore the MATLAB functions *besselj* and *bessely*.
- 10.64.** Generate FM and PM waveforms using simple MATLAB programs.
- 10.65.** Generate FM and PM using the MATLAB *modulate(.)* function.
- 10.66.** A carrier wave has an unmodulated frequency of 6MHz. Sketch the instantaneous frequency time curve when it is frequency modulated by a 4000Hz sawtooth waveform. which is varied 5V peak-to-peak about zero dc value. The frequency deviation constant is 8kHz per volt. Use MATLAB for coding.
- 10.67.** A carrier is frequency modulated by a ramp voltage that starts at zero and rises to 10V in 2ms, when it abruptly drops to zero. Given that the frequency deviation constant 5kHz/V and the unmodulated sinusoidal carrier frequency is 400kHz, plot the modulated angle over time base of 10ms. Use MATLAB for coding.
- 10.68.** An FM transmitter has a carrier oscillator with a rest frequency of 4.5MHz. The oscillator shifts the frequency by $\pm 1.6 \text{ kHz}$ when a 1.6V p-p message signal is applied. The frequency multiplier section following the oscillator has three frequency triplers. Find the transmitted carrier rest frequency, the deviation, and the percentage modulation at the antenna.



Pulse Modulation

11.1 Introduction

Just as the amplitude, frequency, or phase of a sinusoidal carrier can be modulated with an information signal, so the amplitude, frequency, or phase (or position) of pulses in a pulse train can also be modulated. In the field of telecommunications the most widely used form is *pulse code modulation*, or PCM. This is a variant of pulse amplitude modulation in which the pulse amplitudes are transmitted in binary code (and hence the method is sometimes referred to as *coded pulse modulation*). Because of its great importance in telecommunications systems, PCM will be described in considerable detail in a later section. The basics of some of the other forms of pulse modulation are also covered.

11.2 Pulse Amplitude Modulation (PAM)

In pulse modulation the unmodulated carrier is a periodic train of pulses as sketched in Fig. 11.2.1. The unmodulated pulse amplitude is shown as A and the pulse width as τ . The periodic time of the pulse train is shown as T_s . The reason for using subscript s will become apparent shortly. In terms of the mathematical notation introduced in Eq. (3.4.1), the pulse train may be described by

$$v_p(t) = \sum_{k=-\infty}^{\infty} A \operatorname{rect}\left(\frac{t - kT_s}{\tau}\right) \quad (11.2.1)$$

In *pulse amplitude modulation* (PAM) the amplitudes of the pulses are varied in accordance with the modulating signal. Denoting the modulating signal as $m(t)$, pulse amplitude modulation is achieved simply by multiplying the carrier with the $m(t)$ signal, as illustrated in Fig. 11.2.2. The balanced mixer/modulators described in Section 5.10 and 8.9 are frequently used as multipliers for this purpose. The output is a series of pulses, the amplitudes of which vary in proportion to the modulating signal. The particular form of pulse amplitude modulation (shown in Fig. 11.2.4) is referred to as *natural PAM*, because the tops of the pulses follow the shape of the modulating signal.

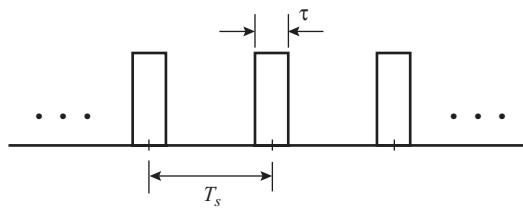


Figure 11.2.1 Periodic pulse train.

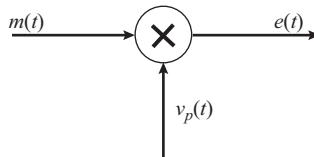


Figure 11.2.2 Product modulator used to produce natural PAM.

The pulse train acts as a periodic switching signal to the modulator, which when switched on allows *samples* of the modulating signal to pass through to the output. The periodic time of the pulse train is known as the sampling period and hence the use of the subscript s . Note that T_s is the period from the beginning of one sample to the next, not the pulse duration. The sampling frequency is

$$f_s = \frac{1}{T_s} \quad (11.2.2)$$

The equation describing natural PAM is found as follows. The Fourier series for the unmodulated pulse train is given by Eq. (2.9.1) as

$$\begin{aligned} v_p(t) &= a_0 + \sum_{n=1}^{n=\infty} a_n \cos \frac{2\pi n t}{T_s} \\ &= a_0 + a_1 \cos \frac{2\pi t}{T_s} + a_2 \cos \frac{4\pi t}{T_s} + \dots \end{aligned} \quad (11.2.3)$$

The modulated pulse train is then

$$\begin{aligned} e(t) &= m(t) \cdot v_p(t) \\ &= a_0 m(t) + a_1 m(t) \cos \frac{2\pi t}{T_s} + a_2 m(t) \cos \frac{4\pi t}{T_s} + \dots \end{aligned} \quad (11.2.4)$$

The right-hand side of this equation shows that the modulated wave consists of the modulating signal, multiplied by the dc term a_0 and a series of DSBSC-type components (see Section 8.9) resulting from the harmonics in the pulse waveform. Denoting the modulating signal spectrum by $M(f)$ and the highest-frequency component in this by W , the spectrum for the PAM signal will be as shown in Fig. 11.2.3.

To be able to transmit the higher DSBSC components, it is clear that a wide-bandwidth transmission system is required. The observant student will notice that, since all the modulating signal spectrum is contained in

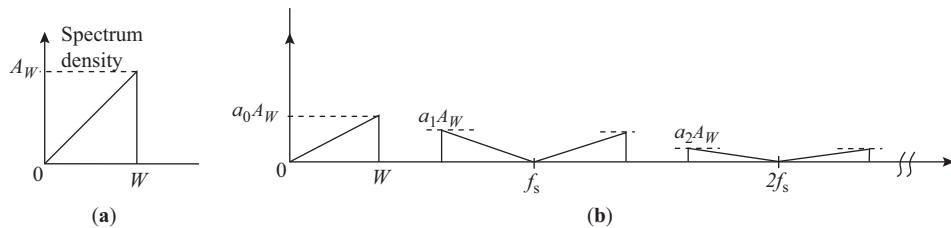


Figure 11.2.3 Spectrum $M(f)$ (a) for the modulating signal and (b) for the natural PAM wave.

the baseband part of the spectrum, it should only be necessary to transmit this, which of course is true. However, in this case there would be no point in using PAM since the original baseband signal $m(t)$ might just as well be transmitted directly, and, in addition, the factor a_0 , which is always less than unity, can severely reduce the amplitude of the baseband component of the spectrum in PAM. This therefore raises the question of why we use pulse modulation in the first place. The answer is that pulse modulation allows pulses carrying different modulating signals to be interleaved in time, forming what is known as a *time division multiplexed* (TMD) signal. Time division multiplexing is described in more detail in connection with pulse code modulation in Section 11.3. But Fig. 11.2.4 shows, in a simplified manner, how five analog signals might be time division multiplexed. This is accomplished using a synchronous analog selector switch called a multiplexer or commutator, as shown in Fig. 11.2.4. Integrated circuits such as the National MM54HC4051 eight-channel CMOS analog multiplexer can be used for this purpose. This switch sequentially connects each modulator output to the transmission line only for the short duration of the sampling pulse. At the receiving end, a similar circuit called a demultiplexer or distributor is used to separate the signals. This circuit acts to synchronously connect the transmission line output to each separate channel demodulator in the same sequence as the multiplexer. Provision must be made to transmit a synchronizing signal from the transmitter to the receiver in order to keep them in step with each other. One channel can be reserved for this purpose, or special synchronizing pulses can be periodically transmitted.

Because of its wideband nature, PAM has a very restricted range of applications for direct transmission of signals. It is used, for example, in instrumentation systems and in analog-to-digital (A/D) converters used for computer interfacing. In the next section the place of PAM as an intermediate stage in the generation of PCM signals will be studied. As a prelude to this, the spectrum of the PAM signal will be examined in more detail. The spectrum for natural PAM showing the low-frequency and the first DSBSC components is shown in Fig. 11.2.5.

To prevent the lower edge of the DSBSC spectrum from overlapping with the low-frequency spectrum, the separation Δ between these must not be less than zero. Hence, from Fig. 11.2.5(a),

$$W + \Delta = f_s - W \quad (11.2.5)$$

With $\Delta \geq 0$ it follows that

$$f_s \geq 2W \quad (11.2.6)$$

This condition imposed on the sampling frequency states that *the sampling frequency must be at least twice the highest frequency in the modulating signal*, a statement that forms part of a fundamental theorem in communications known as the *sampling theorem*. As shown in Fig. 11.2.5(b), if the sampling condition is not met, parts of the spectra overlap, and once such overlap is allowed to occur the spectra can no longer be separated by filtering. Because the high-frequency components in the DSBSC spectrum (for example, $f_s - W$) appear in the low-frequency part of the spectrum, the effect is termed *aliasing*. To avoid aliasing, the modulating signal is first passed through an *antialiasing filter*, which cuts the signal spectrum off at some value W . The system designer

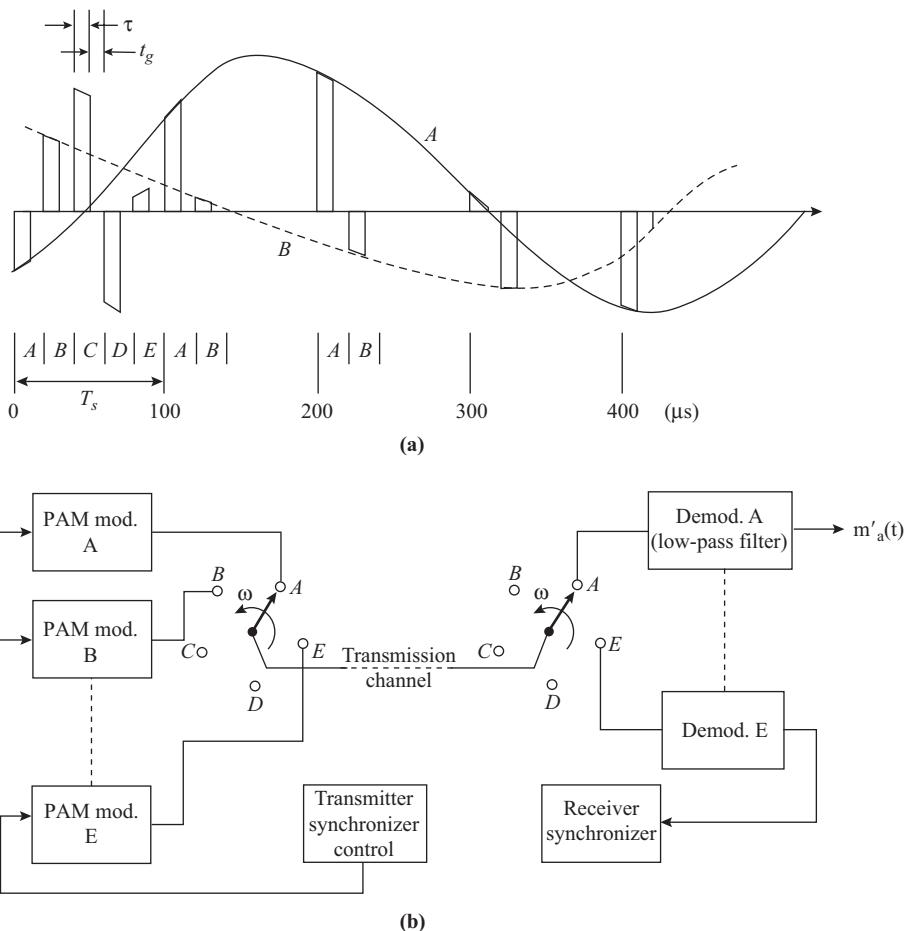


Figure 11.2.4 (a) Five-channel time division multiplexed (TDM) PAM signal, (b) System configuration.

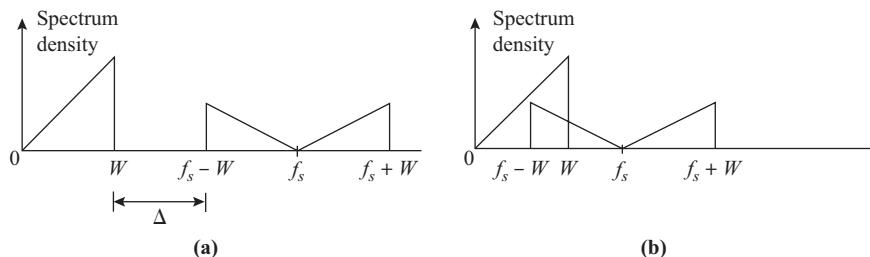


Figure 11.2.5 (a) Part of the spectrum for the natural PAM wave, showing the required separation Δ . (b) Aliasing effect.

must then ensure that the sampling frequency is at least twice this value. For example, it is standard practice to use an antialiasing filter with a cutoff frequency of 4 kHz for digital telephony, with a corresponding sampling frequency of 8 kHz, as described in Section 11.3. The sampling frequency $f_s = 2W$ is known as the *Nyquist frequency*.

If the pulse width of the carrier pulse train used in natural sampling is made very short compared to the pulse period, the natural PAM becomes what is referred to as *instantaneous PAM*. The samples in this case represent the modulating signal at the instant of sampling. Clearly, a shorter pulse width allows more time for additional channels to be included in a TDM format. However, a short pulse means less energy per sample, and as shown in Chapter 2, the magnitude of the coefficients $a_0, a_1 \dots$ are directly proportional to the pulse width. Therefore, to maintain reasonable pulse energy a *sample-and-hold* technique is employed. One way of achieving this is illustrated in Fig. 11.2.6(a). A periodic train of short clocking pulses $\phi_1(t)$ close the transistor switch Q_1 , allowing “instantaneous” samples of the analog signal to be passed on to the capacitor C . Each sample is held in the capacitor for a time T , when the delayed clocking pulses $\phi_2(t)$ operate the transistor switch Q_2 to discharge the capacitor before the next sample arrives. In this way *flat-topped samples* are formed that provide the input to the A/D converter. The A/D converter has time T in which to operate on any one pulse. As can be seen, apart from the short switching delay, $T \approx T_s$. The modulated pulses have flat tops as shown in Fig. 11.2.6(b).

Denoting the spectrum of the analog signal by $M(f)$ as before, it can be shown, by a rather advanced mathematical argument that will not be repeated here, that the spectrum of the flat-topped output pulses is described by

$$V(f) = AM(f)e^{-j\pi fT} \operatorname{sinc} fT \quad (11.2.7)$$

where A is a constant. The exponential term represents the effect of the time delay T , which does not distort the message, but the sine (fT) represents amplitude distortion, which must be corrected for. The effect that the flat-topped pulses have on the spectrum is termed the *aperture effect*. It will be recalled from Eq. (2.9.6) that

$$\operatorname{sinc} fT = \frac{\sin \pi fT}{\pi fT} \quad (11.2.8)$$

The distortion resulting from the aperture effect is compensated for in the receiver by means of an *equalizer filter*, which has an inverse characteristic proportional to $1/\operatorname{sinc} fT$.

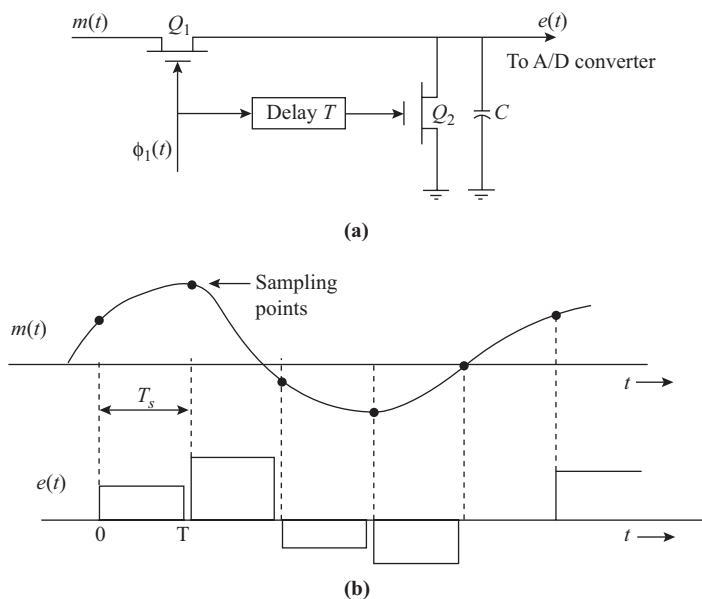


Figure 11.2.6 (a) Sample-and-hold circuit used for generating a flat-topped PAM signal, (b) Sampled waveform.

11.3 Pulse Code Modulation (PCM)

The basic elements for the generation of pulse code modulation can now be put in place. The analog signal $m(t)$ is passed through an *antialiasing filter* and sampled as described in the previous section. The flat-top samples are converted to digital numbers through an analog-to-digital converter. One important step has been added prior to the A/D conversion, that of *quantization*, as shown in Fig. 11.3.1.

Quantization

Quantization is the process of rounding off the values of the flat-top samples to certain predetermined levels in order to make a finite and manageable number of levels available to the A/D converter. Otherwise the sampling levels could take on any value within the peak-to-peak range of the analog signal, which in theory would result in an infinite number of levels.

In the quantization process, the total signal range is divided into a number of subranges as shown in Fig. 11.3.2. Each subrange has its mid-value designated as the standard or code level for that range. Comparators are used to determine which subrange a given pulse amplitude is in, and the code for that subrange is generated. To illustrate the process, a comparatively small number of levels, eight in all, are shown, along with a possible binary code for each level. It will be noticed that the level representing zero analog volts has two binary numbers, one for $0+$ and one for $0-$. In the coding scheme shown, positive values of the analog signal are signed with a binary **1** and negative values with a binary **0**. Thus the leading bit indicates the polarity of the analog signal. The remaining two bits then encode the segment that the sampled value lies in. For example, the last sample point shown is negative, and therefore the leading bit is **0**; it lies in segment L_{-3} for which the binary code is **11**; hence the binary code for this quantized sample is **011**.

In digital telephony, an 8-bit code word is generally used, which allows for 256 levels. Although the quantization stage is shown as a separate block, in practice it will usually be an integral part of the A/D conversion, and integrated circuits are readily available for this purpose.

Figure 11.3.3(a) illustrates the quantization process in a slightly different way. The straight line shows the linear input-output relationship that would exist if quantization were not employed, and the staircase function shows the quantized relationship. This particular type of function is referred to as *mid-tread* quantization, since the quantization levels correspond to the input values at the middle of the tread on the staircase function. It is also possible to have a *mid-rise* type of function, and it is left as an exercise for the student to sketch this.

Figure 11.3.3(b) shows the *quantization error* as a function of input voltage. This is the difference between the quantized level and the analog input, or $V_k - v_{ik}$. The quantization error appears as noise, referred to as *quantization noise*, on the analog signal when it is recovered at the receiver. As can be seen, the quantization error can lie between $\pm\Delta V/2$, and assuming it has a uniform probability density distribution (meaning that it is equally likely to be at any point within the range), it can be shown that the mean-square quantization error is

$$E_{nq}^2 = \frac{(\Delta V)^2}{12} \quad (11.3.1)$$

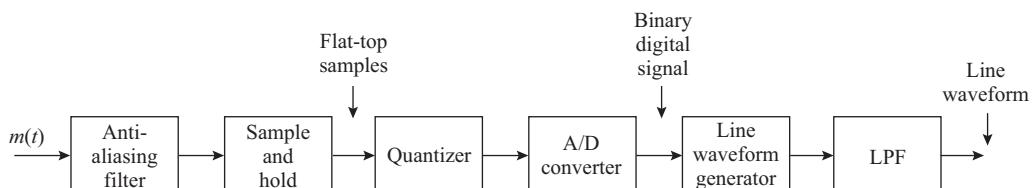


Figure 11.3.1 Basic stages in the generation of PCM.

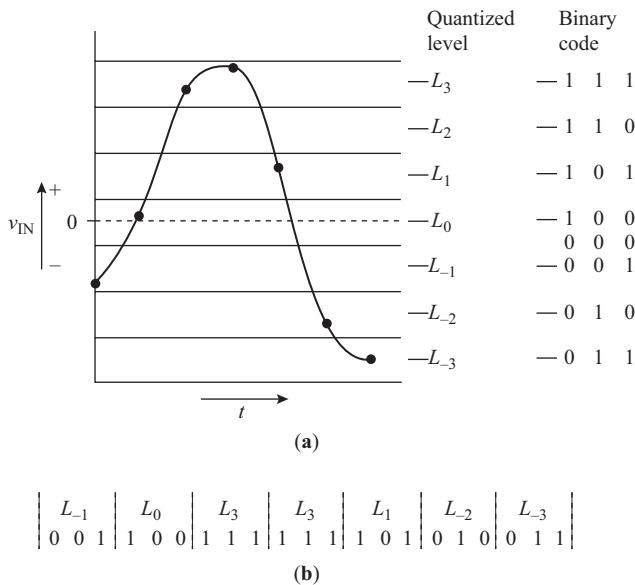


Figure 11.3.2 (a) Analog sampled levels (shown as dots) and the corresponding quantized levels, (b) Binary output for the quantized levels.

This is also the mean-square quantization noise voltage. For a total number of L levels, the peak-to-peak signal range is $\pm L\Delta V/2$, and for a signal that has a uniform probability density distribution within this range, the mean-square signal voltage is

$$E_s^2 = \frac{(L\Delta V)^2}{12} \quad (11.3.2)$$

It follows therefore that the signal-to-quantization noise ratio is

$$\left(\frac{S}{N}\right)_q = \frac{E_s^2}{E_{nq}^2} = L^2 \quad (11.3.3)$$

This shows that to maintain a high S/N_q ratio the number of steps should be high. For example, for $L = 256$, $S/N_q \cong 48$ dB. In terms of the number of bits per code word n , $L = 2^n$ and hence

$$\left(\frac{S}{N}\right)_q = 2^{2n} \quad (11.3.4)$$

It is left as an exercise for the student to show that if $m(t)$ is a sine wave that occupies the full input range then

$$\left(\frac{S}{N}\right)_q = 1.5L^2 \quad (11.3.5)$$

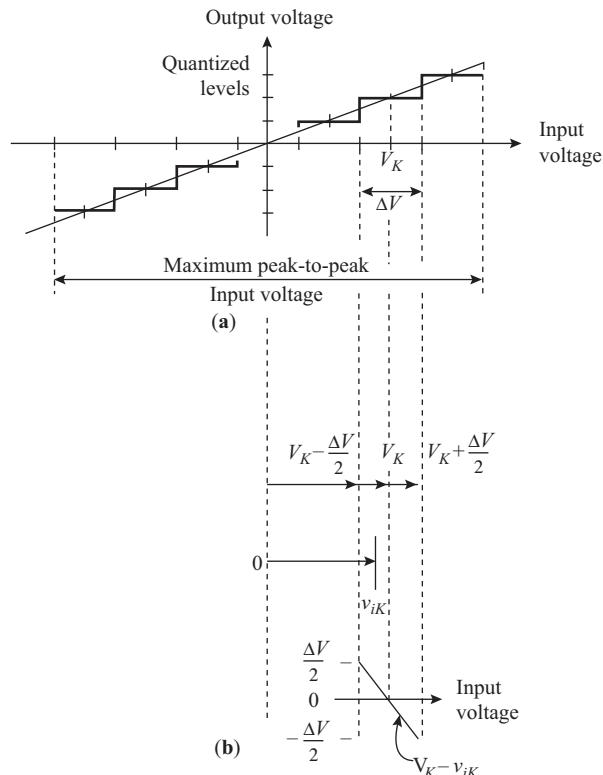


Figure 11.3.3 (a) Linear quantization, (b) Quantization error.

More generally, the ratio between the peak and rms values of the signal voltage will be some value $k = E_{\text{rms}}/E_{\text{max}}$. If distortion is to be avoided, the maximum peak signal level must not be allowed to exceed half the total input voltage range or $E_{\text{max}} = L\Delta V/2$. The signal-to-quantization noise ratio in this case is

$$\begin{aligned}
 \left(\frac{S}{N}\right)_q &= \frac{E_{\text{rms}}^2}{E_{nq}^2} \\
 &= \left(\frac{kL\Delta V}{2}\right)^2 \times \frac{12}{(\Delta V)^2} \\
 &= 3k^2 L^2
 \end{aligned} \tag{11.3.6}$$

EXAMPLE 11.3.1

A PCM system is to have a signal-to-noise ratio of 40 dB. The signals are speech, and an rms-to-peak ratio of -10 dB is allowed for. Find the number of bits per code word required.

SOLUTION Equation (11.3.6) can be written as

$$\left(\frac{S}{N}\right)_q = 3k^2 2^{2n}$$

from which

$$\begin{aligned} 10 \log \left(\frac{S}{N} \right)_q &= 10 \log 3 + 20 \log k + 20n \log 2 \\ \therefore \left(\frac{S}{N} \right)_q \text{ dB} &= 4.77 + (k) \text{ dB} + 6.02n \\ \therefore 40 &= 4.77 - 10 + 6.02n \end{aligned}$$

Therefore,

$$n = \frac{40 - 4.77 + 10}{6.02} = 7.5 \approx 8 \quad (\text{rounded up})$$

A useful relationship between the bandwidth B , required to transmit a single-channel PCM signal, and $(S/N)_q$ is readily derived. Assuming random signals such that $(S/N)_q = 2^{2n}$ applies, then in decibels this is

$$\left(\frac{S}{N} \right)_q \text{ dB} = 10 \log (2^{2n}) \cong 6n \quad (11.3.7)$$

Now with samples taken at the rate of $f_s = 2W$ (Hz), where W is the bandwidth of the signal being sampled, and with n bits per sample, the bit rate is $R_b = 2nW$ (bps). Theoretically, this signal can be transmitted on a channel with the Nyquist bandwidth or half of the bit rate. However, inter-symbol interference (ISI) increases the noise level under this condition. With raised-cosine filtering (see Section 3.4) applied to minimize the ISI, the bandwidth required is increased to $B = (1 + \rho)R_b/2 = (1 + \rho)nW$ (bps). The bandwidth expansion ratio is $B/W = (1 + \rho)n$, and this may be used to substitute for n in Eq. (11.3.7) to give

$$\left(\frac{S}{N} \right)_q \text{ dB} = \frac{6}{1 + \rho} \frac{B}{W} \quad (11.3.8)$$

This shows that the signal-to-noise ratio can be improved at the expense of bandwidth. The following example illustrates this exchange.

EXAMPLE 11.3.2

A telephone signal with a cutoff frequency of 4 kHz is digitized into 8-bit samples at the Nyquist sampling rate $f_s = 2W$. Assuming raised-cosine filtering is used with a roll-off factor of unity, calculate (a) the baseband transmission bandwidth and (b) the quantization S/N ratio.

SOLUTION (a) The transmission bandwidth is

$$B = (1 + \rho)W_n = 2 \times 4 \text{ kHz} \times 8 = \mathbf{64 \text{ kHz}}$$

The quantization signal-to-noise ratio is

$$\left(\frac{S}{N} \right)_q \text{ dB} = 6n = 6 \times 8 = \mathbf{48 \text{ dB}}$$

The example shows that the PCM signal requires a 64-kHz bandwidth to transmit a 4-kHz bandwidth analog signal, and what is gained is a high quantization signal-to-noise ratio.

Compression

With speech it is found that the peaks of the signal only infrequently extend over the full range of the input, most of the time residing within a small range about zero. In effect, the signal does not have a uniform probability density function, and as a result the $(S/N)_q$ ratio is lower than that given by Eq. (11.3.3). To compensate for this, a further stage, termed a *compressor*, is added. This is shown in Fig. 11.3.4.

In earlier designs the compressor consisted of an analog amplifier that had variable gain characteristics with a lower gain at higher input so that, in effect, the high-level signals were “compressed.” This compressed signal was passed on to a linear quantizer of the type described in the previous section. The number of quantized levels was chosen to give the required $(S/N)_q$ ratio for the low signal ranges, and the peak signal swings were compressed to fit into these. This approach gave rise to two compression characteristics, which are now fairly well standardized. In North America and Japan a characteristic known as the μ -law is used, and in Europe and many other parts of the world an A law characteristic is used. The compression functions are normally described in terms of normalized voltages. Let v_i represent the input voltage and $v_{i \text{ max}}$ its maximum value. Denoting the normalized input voltage by x , then

$$x = \frac{v_i}{v_{i \text{ max}}} \quad (11.3.9)$$

In a similar manner, the normalized output voltage is defined as

$$y = \frac{v_o}{v_{o \text{ max}}} \quad (11.3.10)$$

In terms of normalized voltages the μ -law is described by

$$y = \text{sign}(v_i) \frac{\ln(1 + \mu|x|)}{\ln(1 + \mu)} \quad (11.3.11)$$

where $\text{sign}(v_i)$ is used to indicate the sign or polarity of v_i and $|x|$ is the magnitude of x . The *compression parameter* is μ , which determines the degree of compression. With $\mu = 0$, the limiting values of the logarithmic functions must be used to show that $v_0 = v_i$ or no compression occurs (see Problem 11.12). The value $\mu = 255$ is widely used, and the characteristic for this value is shown in Fig. 11.3.5.

The A law is described by

$$y = \text{sign}(v_i) \frac{A|x|}{1 + \ln(A)}, \quad \text{for } |x| \leq \frac{1}{A}$$

or

$$y = \text{sign}(v_i) \left[\frac{1 + \ln(A|x|)}{1 + \ln(A)} \right], \quad \text{for } \frac{1}{A} \leq |x| \leq 1 \quad (11.3.12)$$

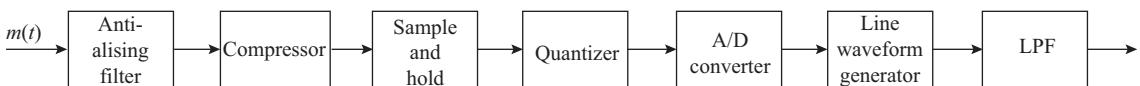


Figure 11.3.4 Addition of a compressor stage.

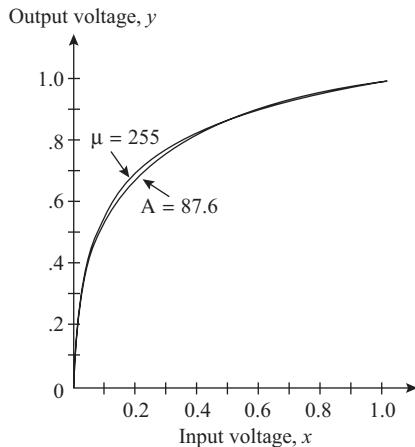


Figure 11.3.5 Compressor characteristics for $\mu = 255$ and $A = 87.6$.

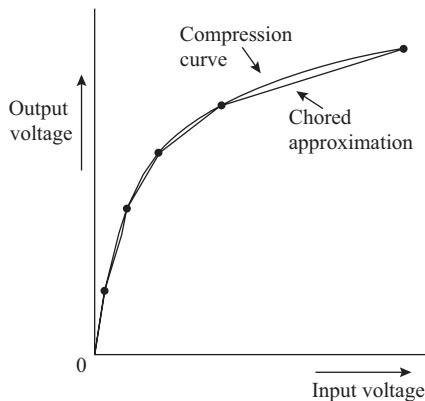


Figure 11.3.6 Chorded approximation to a compression curve.

The compression parameter in this case is A , and the value $A = 87.6$ is widely used. It will be seen that the $\mu = 255$ and $A = 87.6$ curves lie very close to one another, which means that similar “quality” is achieved. Current practice is to approximate the compression characteristics by stepwise functions that form an integral part of the analog-to-digital conversion (as described later). For this reason, the μ law and A law systems are incompatible, and special conversion units are used where interconnections are required, such as might occur on international links.

Although the compressor stage is shown as a separate block and in early PCM circuits it was implemented using analog techniques, in more recent equipment it is implemented as part of the A/D conversion. Rather than having a continuous curve, the compressor characteristic is approximated by linear segments (or chords), as sketched in Fig. 11.3.6.

Each chord is made to cover the same number of input steps, but the step size increases from chord to chord. This is equivalent to having a nonlinear quantizing stage, as illustrated in Fig. 11.3.7(a), which shows the first three chords of a hypothetical compressor. Figure 11.3.7(b) shows the quantization error, and this is seen to increase in amplitude for the larger steps.

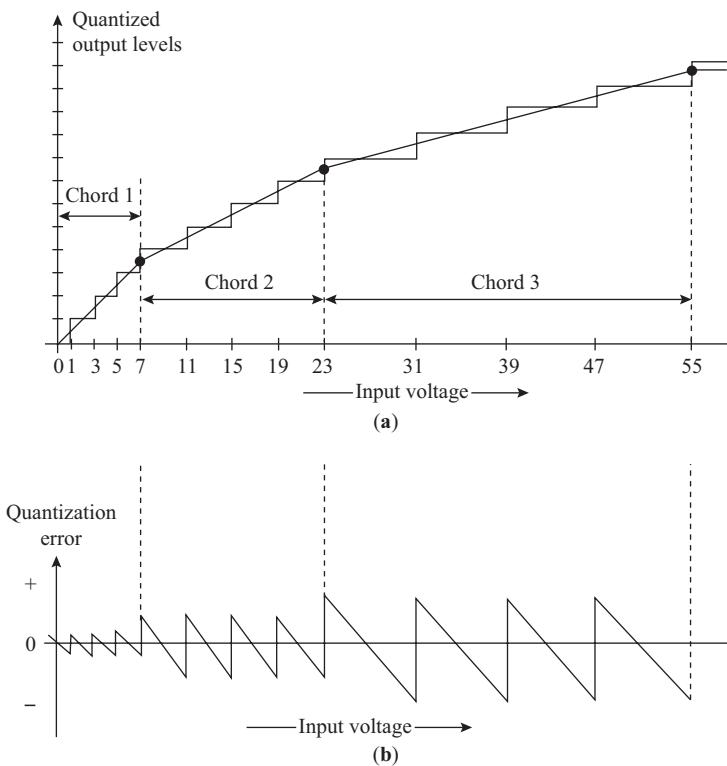


Figure 11.3.7 Nonlinear quantization.

Figure 11.3.8 shows how the analog input voltage might be quantized by such a compressor. As before, the leading bit can be used to encode the analog polarity. With four chords, 2 bits are required to encode these, for example, in ascending order as **00 01 10 11**. A further 2 bits are required to encode the step numbers within a chord, and again in ascending numbers, these might be **00 01 10 11**. Thus sample point A would be encoded as **1 11 10**, and sample point B as **0 01 11**.

Commercially available integrated circuits known as PCM *codecs* (for *coder-decoder*) incorporate all the stages necessary for the PCM conversion process, an example being the Motorola MC145500 series. The μ law or the *A* law can be selected, and Fig. 11.3.9 shows the encoding arrangement for the $\mu = 255$ approximation. As shown in the table, the leading bit in the binary code indicates the sign of the analog input, being **1** for positive and **0** for negative values. The sign bit is followed by 3 bits that indicate the chord, which in turn is followed by 4 bits indicating the step in which the analog value lies. The normalized decision levels are the analog levels at which the comparator circuits change from one chord to the next and from one step to the next. These are normalized to a value of 8159 for convenience in presentation. For example, the maximum value may be considered to be 8159 mV and then the smallest step would be 1 mV. The first step is shown as 1 (mV), but it should be kept in mind that the first quantized level spans the analog zero as shown in Fig. 11.3.3(a), and so a 0+ and 0- must be distinguished. Thus the level representing zero has in fact a step size of ± 1 mV.

As an example, suppose the sampled analog signal has a value of +500 mV. This falls within the normalized range from 479 to 511 mV and therefore the binary code is **1 011 1111**. It should be mentioned that normally the first step in a chord would be encoded **0000**, but the bits are inverted, as noted in Fig. 11.3.6. This is because low values are more likely than high values, and inversion increases the 1-bit density, which

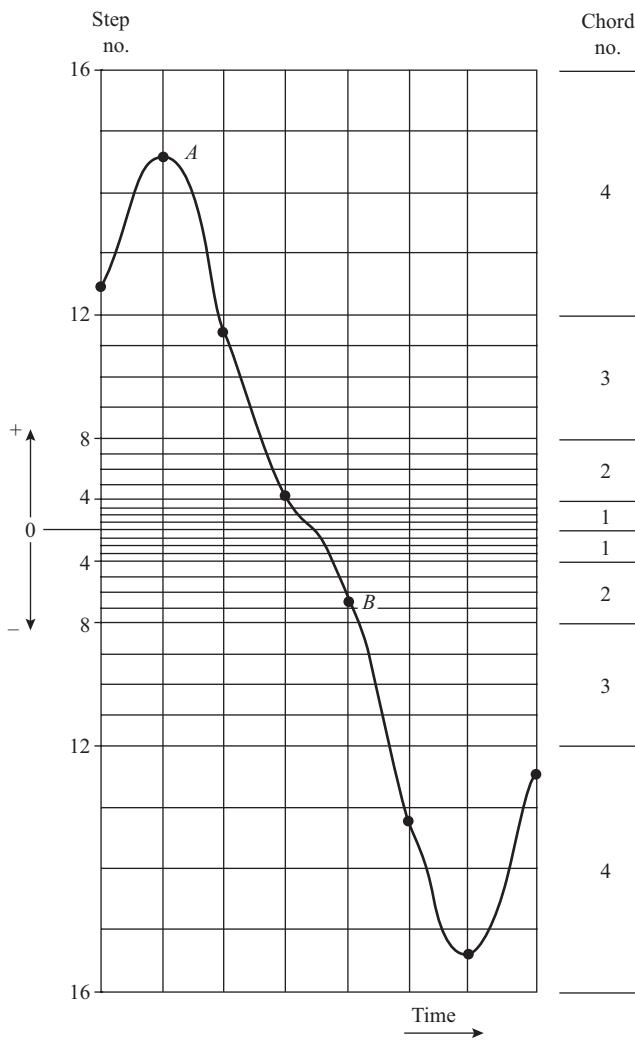


Figure 11.3.8 Nonlinear quantization of an analog signal.

helps in maintaining synchronization as described in Chapter 2. The table also shows the decoded levels corresponding to the quantized transmit levels. For example, the decoded level for the range from 479 to 511 mV is 495 mV, and thus with a +500-mV input signal the quantization error would be 5 mV.

PCM Receiver

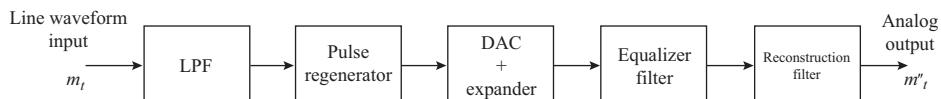
The receiving section of a codec must provide the inverse operations to those of the transmitter. Figure 11.3.10 summarizes the main receiver blocks. The function of the input filter is to limit the noise bandwidth and to complete the waveform shaping required for the avoidance of ISI. A pulse regenerator is used to generate new pulses that are free of thermal noise (but note that quantization noise is always present and cannot be removed). The digital-to-analog converter (DAC) converts the binary signal into flat-top samples and in the process provides the expansion necessary to compensate for the compression applied at the transmitter.

Mu-Law Encode characteristics

Chord Number	Number of Steps	Step Size	Normalized Encode Decision Levels	Digital Code								Normalized Decode Levels
				1	2	3	4	5	6	7	8	
Sign	Chord	Chord	Step	Step	Step	Step	Step	Step	Step	Step	Step	
8	16	256	8159		1	0	0	0	0	0	0	8031
			7903									
			:									:
			4319		1	0	0	0	1	1	1	4191
7	16	128	4063									
			:									:
			2143		1	0	0	1	1	1	1	2079
			2015									
6	16	64	:									:
			1055		1	0	1	0	1	1	1	1023
			991									
			:									:
5	16	32	551		1	0	1	1	1	1	1	495
			479									
			:									:
			239		1	1	0	0	1	1	1	231
4	16	16	223									
			:									:
			103		1	1	0	1	1	1	1	99
			95									
2	16	4	:									:
			35		1	1	1	0	1	1	1	33
			31									
			:									:
1	15	2	3									
			1		1	1	1	1	1	1	0	2
			1									
			0									

NOTES:

- Characteristics are symmetrical about analog zero with sign bit =0 for negative analog values.
- Digital code includes inversion of all magnitude bits.

Figure 11.3.9 Table showing the encoding arrangement for the Motorola MC145500 series of codecs. (Courtesy of Motorola, Inc.)**Figure 11.3.10** Basic blocks in a PCM receiver.

The *equalizer filter* following the DAC compensates for the aperture distortion introduced by flat-top sampling as described in Section 11.2. The magnitude of the equalizer filter response is given by

$$|H(f)|_{eq} = \frac{A}{\text{sinc } fT} \quad (11.3.13)$$

where A is a constant. The equalizer filter is followed by a low-pass filter, often referred to as a *reconstruction filter*, which essentially recovers the analog signal by passing only the low-frequency part of the spectrum, as shown in Fig. 11.2.3. However, quantization noise is also present on the output so that the analog output $m''(t)$ is not identical to $m(t)$. Figure 11.3.11 shows a block diagram for the National Semiconductor TP3051 codec in which the equalizing filter response is incorporated in the switched capacitor filters.

Figure 11.3.12 shows how codecs could be used in a simple TDM/PCM transmitting system. In early designs a common sampler and A/D converter would have been used for the multiplexed channels, but with advances in integrated circuits the use of an individual codec for each channel is the most economical approach for digital telephony. Thus time division multiplexing (TDM) takes place in the multiplexer following the channel codecs, as shown in Fig. 11.3.12. For illustration purposes, a common line waveform generator is shown as a separate unit, but the actual arrangement will depend on the facilities provided in the codecs and multiplexer. A waveform regenerator is also used in the receiver before the multiplexed signal is demultiplexed, following which the individual codecs recover the analog signals, as described previously.

The TDM arrangement known as the *T1 system*, introduced by Bell Telephone, is shown in Fig. 11.3.13 to illustrate some of the significant features of time division multiplexing. This is a 24-channel system in which the sampling frequency is 8 kHz, and 8 bits per sample are used. A *frame-synchronizing signal* is required, and this is achieved by inserting a *frame sync-bit* at the end of each frame. The frame-sync signal is periodic, consisting of a repetitive pattern of bits or codewords. Thus synchronization requires recovery of the sync signal over a number of frames.

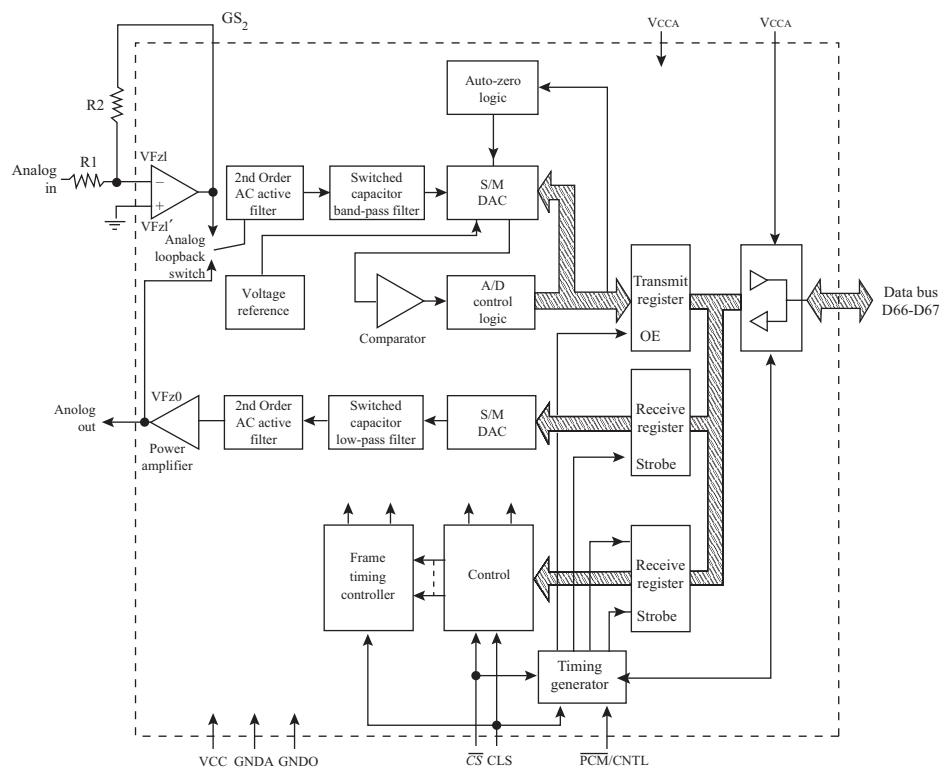


Figure 11.3.11 Block diagram of the National Semiconductor TP3051 codec. (With permission of National Semiconductor Corp.)

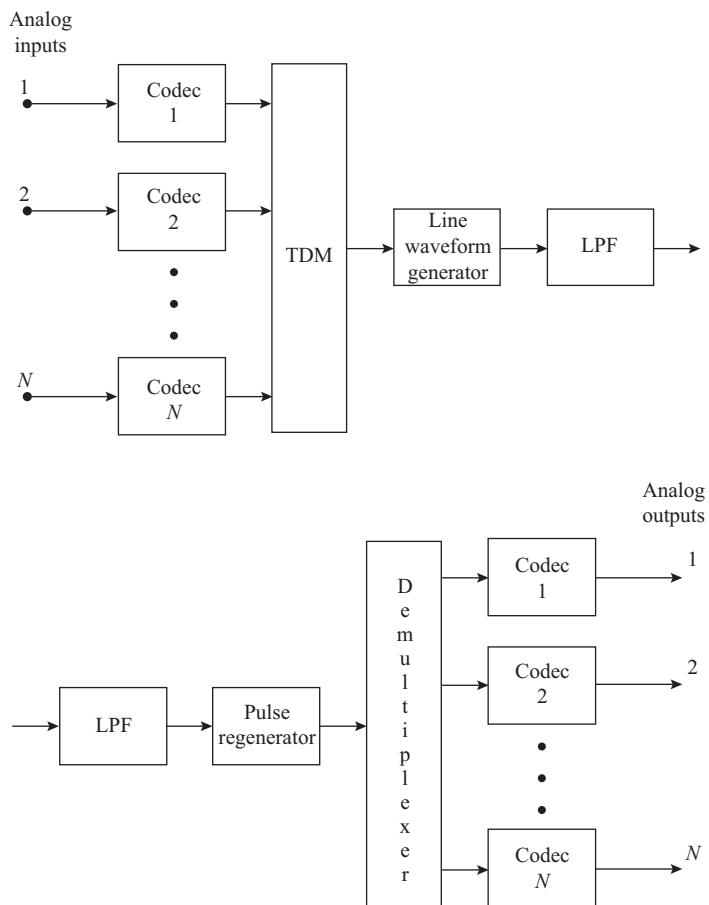


Figure 11.3.12 Basic TDM/PCM system.

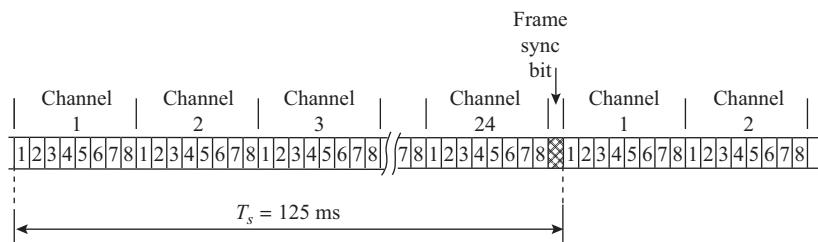


Figure 11.3.13 T1 TDM format.

Since the sampling frequency is 8 kHz, the time between samples for any given channel is $T_s = 125 \mu\text{s}$, which is also the frame period, as shown in Fig. 11.3.13. Any one frame contains $8 \times 24 + 1 = 193$ bits, and therefore the bit rate is $R_b = 193/125 \mu\text{s} = 1.544 \text{ Mbps}$. It is left as an exercise for the student to show that if raised-cosine filtering is used with a roll-off factor of unity the transmission baseband bandwidth required is 1.544 MHz.

It should also be noted that signaling information (such as busy signals and call-completed signals) are transmitted in the T1 system by replacing the eighth bit (least significant bit) in each channel by a signaling bit in every sixth frame. Thus the signaling period is $125 \mu\text{s} \times 6 = 750 \mu\text{s}$ and the signaling rate is $1/750 \mu\text{s} = 1.333 \text{ kbps}$.

Differential PCM

Differential pulse code modulation (DPCM) is a technique in which the *difference* between samples, rather than the sample values themselves, is encoded in binary. The reason for employing DPCM is that speech samples do not change drastically from sample to sample, and therefore the difference values can be encoded using fewer bits. DPCM can be achieved in a number of ways, some involving mostly analog methods, others a mixture of analog and digital circuits, and yet others mostly digital. Figure 11.3.14(a) shows a DPCM transmitter where it is assumed that flat-top sampling has been achieved by a sample-and-hold circuit, as discussed previously. The input to the quantizer is, however, the difference between the flat-top samples and a feedback signal that is an estimate of the input derived from the quantizer output.

The estimated signal is created by the *predictor* block, which is a digital-type filter. This has a shift register in which a predetermined number of samples are temporarily stored while a weighted sum is formed, this being the estimated signal. The input to the predictor at any instant consists of the sum of its own output, which is the best estimate at that instant, and the most recent difference level. Thus the estimate is continuously updated by the difference levels. The estimated flat-top sample is subtracted from the incoming flat-top sample to form the difference signal that is quantized.

At the receiver, Fig. 11.3.14(b), the incoming DPCM binary signal is decoded to form a difference PAM signal. This is fed into a feedback loop similar to that used at the transmitter. A prediction filter forms a best estimate of the flat-top PAM signal from the sum of its own output and the most recent difference PAM level. Thus the PAM output is continuously updated by the difference level. The analog output is recovered from the flat-top PAM signal by the reconstruction filter in the normal manner described for PCM.

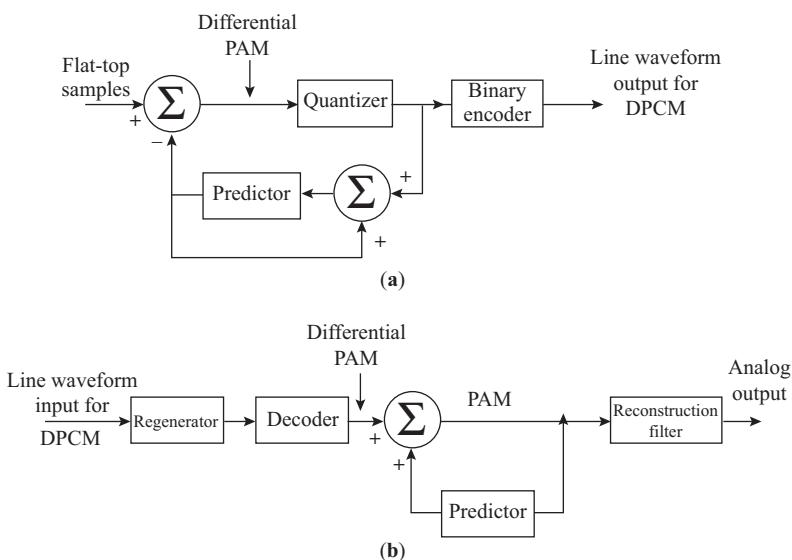


Figure 11.3.14 (a) DPCM transmitter, (b) DPCM receiver.

Delta Modulation

Delta modulation (DM) is a special case of DPCM in which only the polarity of the difference signal is encoded as output. If the difference between the analog input and the feedback signal is positive, this is encoded as a binary **1**, which is transmitted as a positive voltage pulse, and if negative, as a binary **0**, which is transmitted as a negative voltage pulse. The principle of DM is illustrated in Fig. 11.3.15. At the transmitter, one input to the multiplier is a periodic train of unipolar impulses (very short pulses) $p_i(t)$ at the sampling frequency f_s . The other input is the output from the comparator, which consists of fixed amplitude pulses whose polarity depends on the difference signal at the comparator input. The polarity is positive if the analog signal $m(t)$ is greater than and negative if less than the feedback signal $m'(t)$. The multiplier output is therefore a sequence of impulses $p_o(t)$ whose polarity depends on the difference signal. The feedback signal $m'(t)$ is the integral of the multiplier output $p_o(t)$. As shown in Fig. 11.3.15(b), the $m'(t)$ waveform is a staircase approximation to the analog signal $m(t)$. For this staircase waveform, there will be an initial transient period labeled *A*. When the modulator reaches steady state, the staircase waveform “hunts” around the analog waveform as shown at *B*. This hunting produces *granular noise*. Notice, too, that over region *B*, even though the analog waveform has a positive slope, the staircase waveform remains horizontal. At region *C*, the rate of change of the analog waveform is too great for the staircase waveform to follow, with the result that *slope overload* occurs. Slope overload is also seen at *D*.

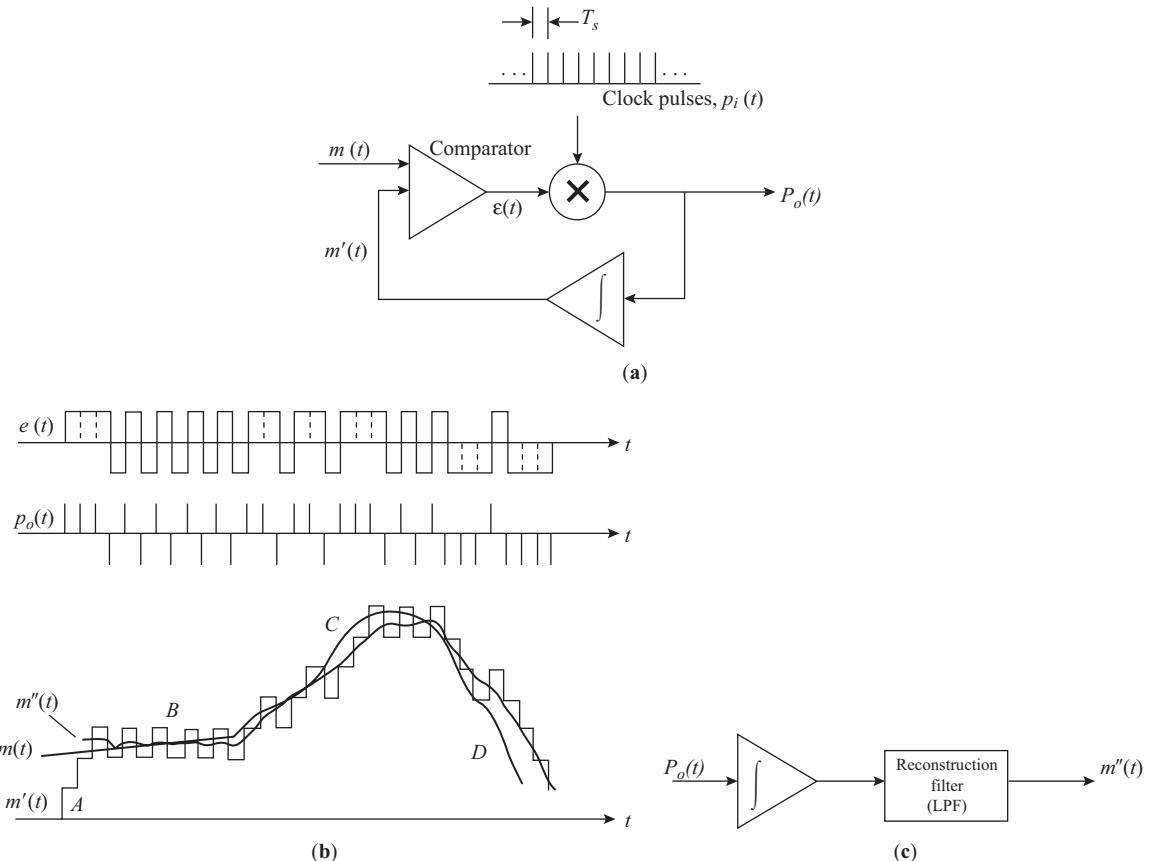


Figure 11.3.15 (a) DM transmitter, (b) DM waveforms, (c) DM receiver.

The *impulse* waveform $p_0(t)$ is converted to a line waveform (not shown) and transmitted as such. At the receiver, Fig. 11.3.15(c), a regenerator (not shown) is used to recover the $p_o(t)$ impulse waveform, which is integrated to produce the staircase approximation similar to that at the transmitter. The analog output $m''(t)$ is recovered from the integrator output through the use of a low-pass (reconstruction) filter. The recovered waveform will exhibit the effects of granular noise and slope overload and will not be an exact replica of the analog input.

Granular noise can be reduced by using a high sampling rate, well above the Nyquist ($f_s = 2W$) value. This tends to shift the granular noise into the high-frequency region of the spectrum, most of which is then removed by the reconstruction filter. A high sampling rate also means that the feedback signal changes more rapidly, reducing the danger of slope overload. It should also be noted that a high sampling rate is beneficial in that the separation Δ shown in Fig. 11.2.5 becomes large, which relaxes the design constraints on the antialiasing filter. However, a high sampling frequency means that a high bit rate is generated, which in turns means that a wider bandwidth is required compared to PCM. This is a sufficient disadvantage to make the simple delta modulation scheme unsuitable for the majority of communication applications.

Granular noise can also be kept small by using a small step value, the size of the step being fixed by the gain of the integrator. However, this leads to the disadvantage that small steps may not integrate quickly enough to follow the analog waveform when it changes rapidly, thus leading more readily to slope overload. A variation on delta modulation known as *adaptive delta modulation* can be used to counteract slope overload. In this, the step size is automatically adjusted to larger values to accommodate the larger magnitudes of rate of change of the analog signal. The *continuously variable slope delta* (CVSD) modulator is an example of such a scheme. The block diagram for the Motorola MC34115 CVSD modulator/demodulator is shown in Fig. 11.3.16(a).

In the MC34115, the contents of 3-bit shift register are monitored. If these are all **1's** or all **0's**, a condition termed *coincidence*, it is an indication that the gain of the integrator is too small. The output at the coincidence terminal is used to charge a capacitor in a low-pass filter, termed a *syllabic filter*, the output of which controls the gain of the integrator. In addition, an all **1's** coincidence means that the signal slope is positive and all **0's** that it is negative, and this information is also fed as a control signal to the integrator through the *slope polarity switch*. Figure 11.3.16(b) shows the CVSD waveforms, Fig. 11.3.16(c) the block diagram for the encoder section, and Fig. 11.3.16(d) the block diagram for the decoder.

Delta modulation is seen to be a 1-bit encoder where the quantized step size is fixed by the integrator rather than the analog signal. Thus, although the Nyquist sampling criterion $f_s \geq 2W$ may be met, the granular noise (similar to quantization noise) may be excessive if the lower limit to the sampling frequency is used. For this reason the sampling frequency is always considerably greater than the lower limit, at least two times this. As mentioned previously, a high sampling frequency eases the design of the antialiasing filter and also shifts the granular noise into the higher part of the spectrum, where most of it can be removed by the reconstruction filter in the receiver.

Sigma-Delta A/D Conversion

The sigma-delta modulator ($\Sigma - \Delta$) is a development of the delta modulator that provides high-resolution (greater than 12 bits) analog-to-digital conversion. Most of the filtering and signal processing takes place in digital circuitry, enabling advantage to be taken of very large scale integrated (VLSI) circuit design, which lowers the cost considerably compared to combined A/D (hybrid) designs. The detailed principles of operation are complex, and only a summary will be presented here, based on information provided in the Motorola publication APR8/D, which relates to the Motorola DSP56ADC16 single-chip, sigma-delta converter.

Figure 11.3.17(a) shows a basic delta modulator, in which it is seen that two integrators are required. Because the operation of integration is linear, the receiver integrator can be placed at the input to the transmitter as shown in Fig. 11.3.17(b). A further simplification can now be made. Since the output of the comparator is the difference of two integrated signals (the integrated analog signal and the integrated feedback signal), the same result can be achieved by integrating the difference of the nonintegrated signals, as shown

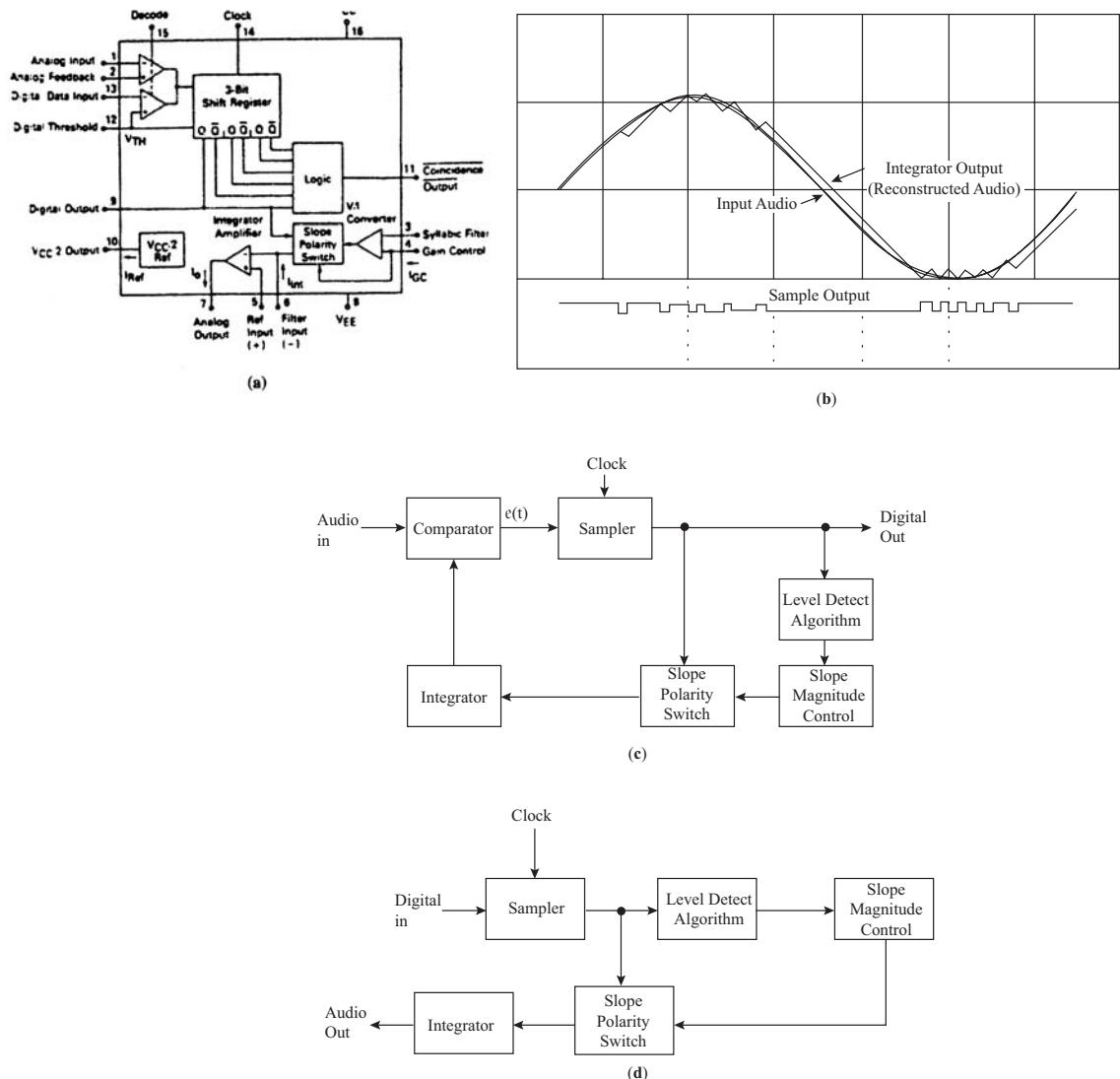


Figure 11.3.16 (a) Block diagram for the MC34115 CVSD modulator/demodulator, (b) CVDS waveforms, (c) block diagram for the encoder, and (d) block diagram for the decoder. (Courtesy of Motorola, Inc.)

in Fig. 11.3.17(c). Thus only one integrator is needed, and this is the *sigma* part of the converter. The integrator is followed immediately by a 1-bit quantizer, which is the *delta* modulator.

At the normal Nyquist sampling rate, the 1-bit resolution achieved by delta modulation would be very coarse, and to offset this a very much higher sampling rate is used. For example, instead of taking, say, one sample and converting this to 16 bits (as would be required in high-resolution PCM), 16 one-bit samples are generated in the same time. This requires a sampling rate that is 16 times the Nyquist rate. The advantage of this approach is that a very simple antialiasing filter can be used, as the separation Δ shown in Fig. 11.2.5 becomes large. Once the signal is in digital form, it can be processed by digital circuitry (for example, digital filtering can be used), and what is known as a *digital decimation filter* is used to reduce the sampling rate to the normal Nyquist value. Noise spectrum shaping also takes place in which most of the granular noise is

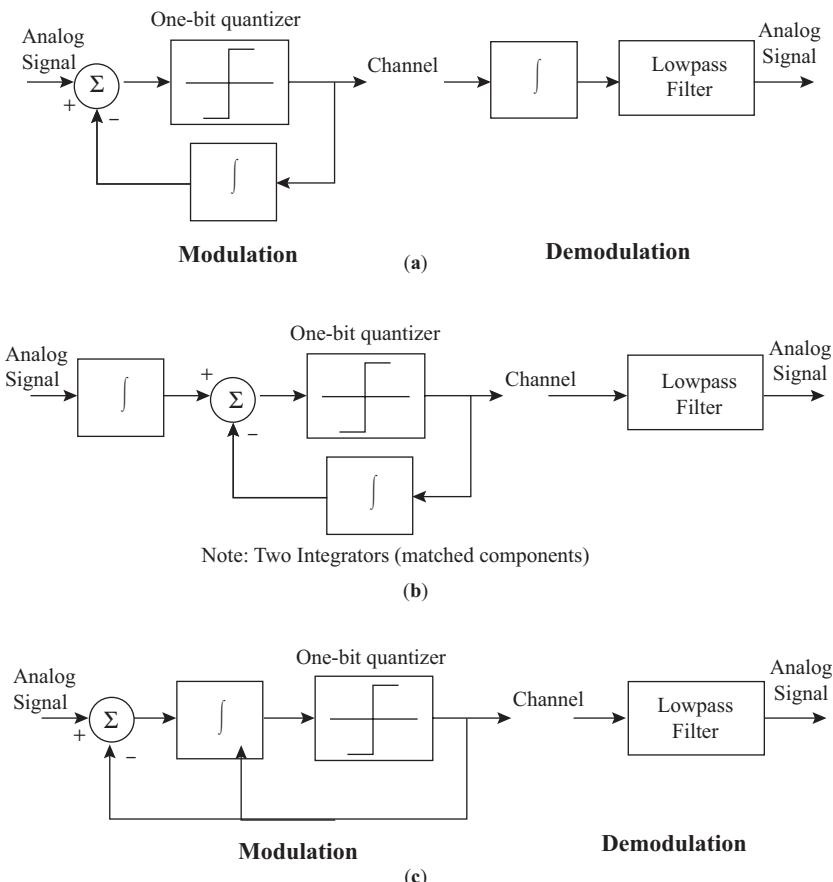


Figure 11.3.17 (a) A basic delta modulation system, (b) Shifting the position of the receiver integrator, (c) Combining the two integrators into one to create a basic sigma-delta modulation system.

shifted to a high part of the spectrum, which is then removed by the digital filter. The quantization signal-to-noise ratio can be as high as 90 dB or better for the Motorola DSP56ADC16.

11.4 Pulse Frequency Modulation (PFM)

A train of rectangular pulses is frequency modulated if (1) their amplitude is kept constant, and (2) the pulse period T_c and pulse duration τ are both made proportional to the modulating signal so that the duty cycle (τ/T_c) of the pulse train remains constant. The modulating wave is not sampled at fixed intervals, as was the case for PAM, but is sampled at the time of occurrence of the modulated pulses, as shown in Fig. 11.4.1(a). The sample taken is used to adjust the following pulse period.

The resulting frequency modulated pulse train is illustrated in Fig. 11.4.1(b). It produces a spectrum as shown in Fig. 11.4.1(c), which contains a fixed dc level and a frequency-modulated carrier and sidebands at each harmonic of the unmodulated carrier frequency $f_c = 1/T_c$. The amplitudes of the harmonics of the unmodulated carrier are constrained by the envelope of $(\sin x)/x$, where $x = n\pi\tau/T = n\pi f\tau$.

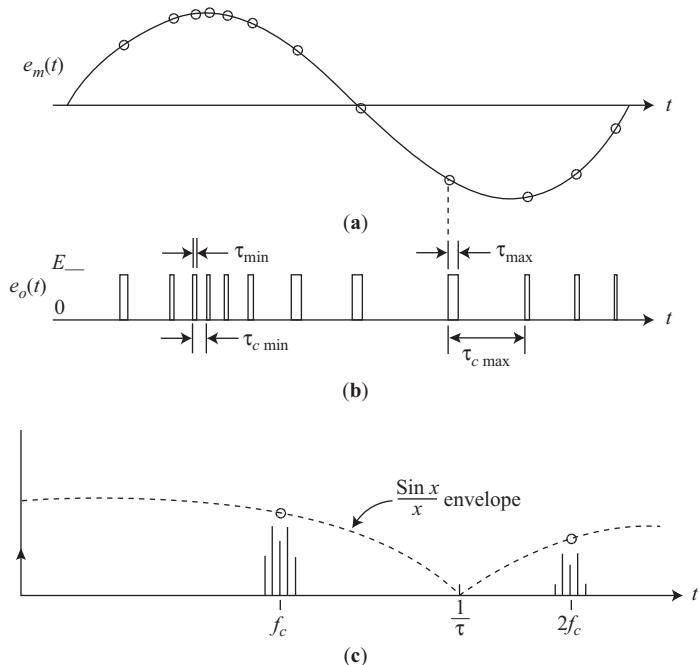


Figure 11.4.1 Pulse frequency modulation (PFM). (a) Modulating signal, (b) Modulated pulse train, (c) Spectrum of the PFM signal.

The spectrum does not contain any of the baseband frequency components, so the modulating signal cannot be recovered by simple low-pass filtering. A frequency demodulator must be used. However, the signal is easy to generate using a mixture of digital and analog components, which has made PFM popular for some types of analog instrumentation systems.

11.5 Pulse Time Modulation (PTM)

Pulse time modulation includes pulse position modulation (PPM) and pulse width modulation (PWM). Both of these produce a form of pulse phase modulation and are sometimes called by that name. Pulse frequency modulation (PFM) is also included, although it is not strictly a time modulation.

11.6 Pulse Position Modulation (PPM)

An unmodulated fixed frequency pulse train produces a recurring time window that is T_c seconds wide, with a pulse of fixed width τ appearing at its center. If the position of the fixed width pulse (and the sampling point) within the fixed width window is varied with a modulating signal, as shown in Fig. 11.6.1(b), the pulse train is pulse position modulated. The effect is that of pure phase modulation of the pulse train.

The spectrum of a PPM signal is shown in Fig. 11.6.1(c). It contains a fixed dc component and a set of carrier and phase modulated sidebands at each harmonic of the carrier frequency. This spectrum is very similar to that for the PFM signal, where the difference exists in the magnitudes and phase shifts that occur in the side frequencies about the carrier harmonics.

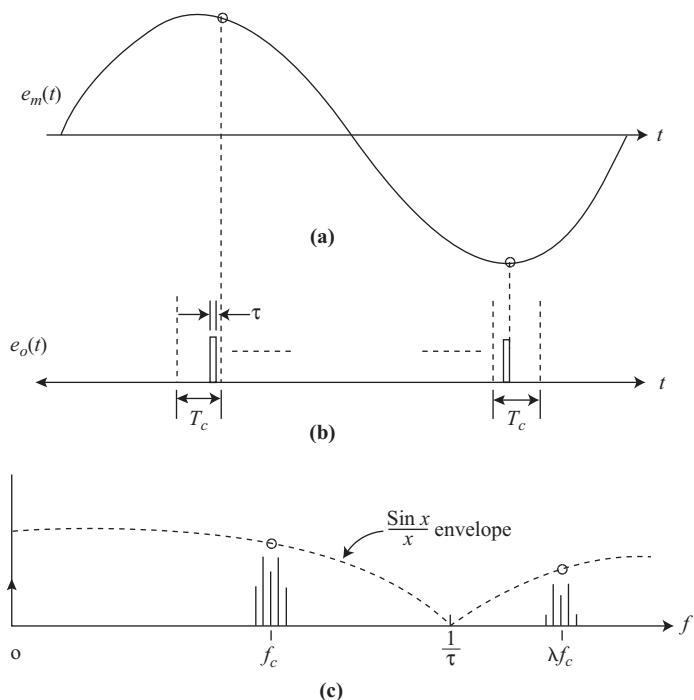


Figure 11.6.1 Pulse position modulated (PPM) signal, (a) Modulating signal, (b) PPM waveform, (c) Spectrum of a PPM signal.

The carrier harmonics are constrained within a $(\sin(x))/x$ envelope and extend indefinitely in frequency. The spectrum does not contain any baseband components, so the modulating signal cannot be recovered using a low-pass filter. This fact is obvious if one notes that the duty cycle of the modulated signal remains constant, so the average signal level (dc) is constant. A phase detector is needed for demodulation, but again generation of this type of signal is relatively easy.

11.7 Pulse Width Modulation (PWM)

If the frequency and amplitude of a pulse train are kept constant and the width of the pulses is varied with a modulating signal, then the result is a pulse width modulated (PWM) signal. Three variations are possible, as shown in Fig. 11.7.1(b), (c), and (d). First, the pulse center may be fixed in the center of the repeating time window T_c and both edges of the pulse moved to compress or expand the width τ . Second, the lead edge can be held at the lead edge of the window and the tail edge modulated. Third, the tail edge can be fixed and the lead edge modulated. The resulting spectra are similar, and, as shown in Fig. 11.7.1(e), they each contain a dc component and a base sideband containing the modulating signal, as well as phase modulated carriers at each harmonic of the pulse frequency. The amplitudes of the harmonic groups are constrained by a $(\sin(x))/x$ envelope and extend to infinity.

Since the baseband information appears in the signal and is not distorted by any modulation effects, it may be recovered using a simple low-pass filter to remove the carrier and its harmonics and a high-pass filter to remove the dc component.

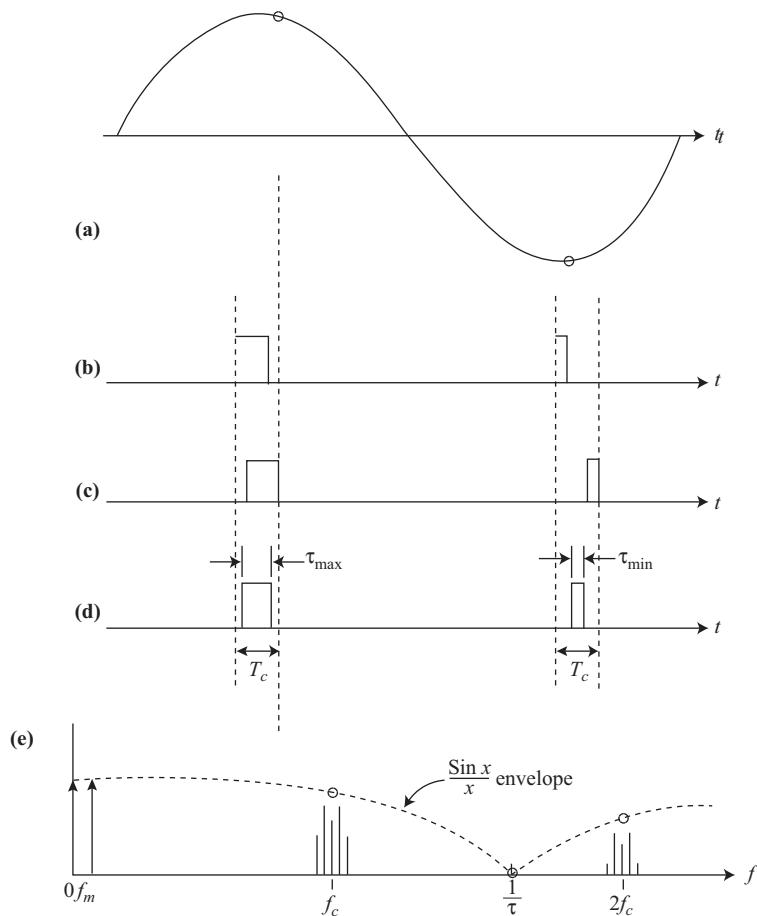


Figure 11.7.1 Pulse width modulated (PWM) signal, (a) Modulating signal, (b), (c), and (d) PWM signals with lead edge, tail edge, or center fixed in the T_c window, (d) Spectrum of a PWM signal showing the baseband component.

PROBLEMS

- 11.1. What sampling rate would be appropriate for each of the following? (a) A 4-kHz telephone channel. (b) A music channel with a maximum signal frequency of 20 kHz. (c) A video channel with a band width of 4.5 MHz.
- 11.2. Given a cosinusoidal modulating signal $m(t) = V_m \cos \omega_m t$ applied to a PAM modulator whose output $e(t)$ is given by Eq. (11.2.4), with a carrier pulse train of magnitude V_p , and duty cycle DC = T_b/T_s , expand Eq. (11.2.4) to the form $e(t) = b_o \cos \omega_m t + b_n [\cos(n\omega_s - \omega_m)t + \cos(n\omega_s + \omega_m)t]$. Write expressions for b_o and b_n in terms of V_m , V_p , T_s , and T_b .
- 11.3. An acosinusoidal signal to be sampled has a magnitude of 0.2 V peak at 3 kHz. A natural PAM generator uses a 1-V carrier pulse train for which the sampling frequency is 8 kHz and the duty cycle is 20%. Calculate the magnitude and frequency for the five components in the spectrum up to the second harmonic of the carrier. Calculate each side frequency as a separate component, and sketch to

scale the portion of the spectrum in question. (Use the results of Problem 11.2 and note that the arguments of the cosinusoids are in radians.)

- 11.4. The sampler in Problem 11.3 is operated at a sampling frequency of 5 kHz instead of 8 kHz. Calculate the frequency of all components in the recovered signal after a low-pass filter with a cutoff frequency of 3.5 kHz. What effect is this result evidence of?
- 11.5. Derive Eq. (11.3.5) for the situation of a maximum-sized sinusoidal signal input.
- 11.6. A sinusoidal signal with a maximum peak input voltage of 5 V is applied to a PCM channel using a 10-bit code word. Find (a) the number of quantization levels used, (b) the RMS quantization noise level in volts, and (c) the maximum sinusoidal signal to quantization noise ratio in decibels.
- 11.7. Given Eq. (11.3.4), derive Eq. (11.3.7) for the decibel $(S/N)_q$ of a PCM system.
- 11.8. A PCM system is to carry a 20-kHz music channel. It is to have a signal-to-noise ratio of 80 dB and the peak maximum signal is 15 dB over its rms value, (a) What sampling rate should be used? (b) How many bits should be used in the sample code word?
- 11.9. Show that, if the output bit rate of a PCM system is 1.544 Mbps and raised-cosine filtering to protect against ISI with a roll-off factor of unity is used, the required transmission bandwidth is 1.544 MHz (the same number as the bit rate).
- 11.10. A 15-kHz signal is transmitted on an 8-bit PCM channel. Raised- cosine filtering with $\rho = 0.5$ is applied to the transmission facility to reduce ISI noise. Find (a) the bandwidth required on the transmission facility and (b) the $(S/N)_q$.
- 11.11. Explain how companding (compression before transmission and expansion after) reduces the amount of noise introduced by the transmission channel.
- 11.12. Show that a μ law compressor has no compression if $\mu = 0$, so that $V_0 = V_i$.
- 11.13. Use Mathcad or a similar plotter to plot on the same graph, for values of $(0 \leq x \leq 0.1)$ (a) the μ law characteristic for $\mu = 255$, and (b) the A law characteristic for $A = 87.6$. (Use the equations for $x > 1/A$.)
- 11.14. Repeat Problem 11.13 for $\mu = 100$ and $A = 50$.
- 11.15. For the μ law encoder of Fig. 11.3.9, determine the code words for the following input voltages: (a) +857 mV, (b) -136 mV, (c) +4.125 V, and (d) -0.015 V. Assume a maximum voltage range of 5.000 V.
- 11.16. For the μ law decoder of Fig. 11.3.9, find the output voltage for the following code words, given that $V_o(\max) = 8159$ mV: (a) 1101 0110, (b) 0000 0010, (c) 1000 0000, and (d) 7E (hex).
- 11.17. Generate PWM using the *modulate()* command in MATLAB.
- 11.18. Generate PPM using the *modulate()* command in MATLAB.
- 11.19. Plot the *A-law* using MATLAB.
- 11.20. Plot the μ -*law* using MATLAB.
- 11.21. Consider a chopper sampled waveform given by the equation: $x_s(t) = C_0x(t) + 2C_1x(t)\cos\omega_s(t) + 2C_2x(t)\cos2\omega_s(t) + \dots$, with $\tau = \frac{T_s}{2}$, $f_s = 100\text{Hz}$ and $x(t) = 2 + 2\cos60\pi t + 2\cos160\pi t$. Draw and label the one-sided spectrum of $x_s(t)$ for $0 \geq f \geq 300\text{Hz}$. Then find the output waveform when $x_s(t)$ is applied to an ideal low pass filter with $B = 75\text{Hz}$.
- 11.22. The signal $x(t) = \text{sinc}^2 5t$ is ideally sampled at $t = 0, \pm 0.1, \pm 0.2, \dots$ and reconstructed by an ideal low pass filter with bandwidth, $B = 5$, unit gain, and zero time delay. Carry out the sampling and re-construction graphically, using MATLAB.
- 11.23. What is the *Nyquist rate* to adequately sample the following signals: (a) $\text{sinc}(100t)$ and (b) $\text{sinc}^2(100t)$.
- 11.24. Show that a PAM signal can be demodulated using a product detector. Specify the frequency parameters for the local oscillator (LO) and the low pass filter.

Digital Communications

12.1 Introduction

As mentioned in the introduction to Chapter 3, digital signals are coded representations of information to be transmitted. The block diagram of a basic digital communications system is shown in Fig. 12.1.1. The diagram shows a number of encoding functions. The source encoder converts the information into binary code; each source codeword then represents one of the discrete levels at the input. If the input is an analog waveform, part of the job of the source encoder is to first digitize this. Source encoding for speech waveforms is described in detail in Chapter 11.

The next function shown is that of channel encoding, which generates channel codewords from the source codewords. Certain bit patterns in the source codewords may be troublesome to transmit, and one function of channel encoding is to change these. The high-density bipolar codes described in Chapter 3 are examples of this aspect of channel encoding. Another function of channel encoding is to alter the source codewords in a controlled manner in order to combat noise-induced errors at the receiving end. This is channel encoding for error control and is described more fully in Section 12.13.

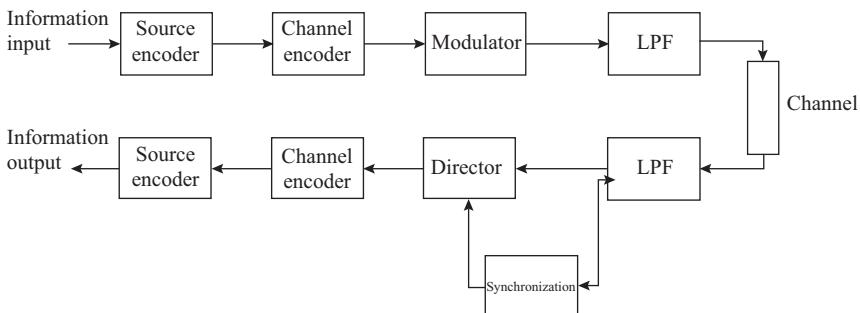


Figure 12.1.1 Basic digital communications system.

For transmission, a continuous waveform is modulated by the channel codewords, this being the function of the modulator. The continuous waveform may be a carrier wave as described in Chapters 8 and 10, and the modulated waves in such cases are referred to as *band-pass waveforms*. Band-pass waveforms are required for radio transmission. The modulated continuous wave may also be simply a dc level change, for example the unipolar and polar waveforms described in Chapter 3. These are referred to as *baseband waveforms*, and they are suitable for transmission over lines. Hence they are also known as line waveforms.

The waveforms, whether baseband or band pass, are referred to as digital waveforms or digital signals, although in fact they are continuous waveforms that are analog representations of the channel codewords. Referring to them as digital waveforms does make clear the distinction between this mode of analog transmission and *analog communications*, in which information such as speech is transmitted in analog rather than digitally encoded form.

To recover the information contained in the transmitted waveform, the receiver must have decoding functions corresponding to the encoding functions at the transmitter, as shown in Fig. 12.1.1. In addition, in what are termed *self-synchronizing systems*, the receiver requires circuitry to recover correct bit timing and, in the case of band-pass systems, the carrier frequency. These functions are in the block labeled *synchronization*.

12.2 Synchronization

Synchronization is required at a number of different levels in digital communications systems, these being classified as follows:

1. *Network synchronization*: required so that stations sharing a network can transmit and receive in an orderly fashion.
2. *Frame synchronization*: required to keep track of the individual channels in a time division multiplexed system as described in Chapter 11.
3. *Codeword and node synchronization*: required to keep track of blocks of bits in a bit stream, where each block forms a codeword, usually designed for the purpose of error control. This is discussed further in Section 12.13.
4. *Symbol synchronization*: required in order that symbols, which maybe hidden in a noisy waveform, are sampled at the optimum time. This is discussed further in Section 12.7 under the heading of bit-timing recovery.
5. *Carrier synchronization*: required in order to demodulate a carrier modulated wave in the most efficient manner. This is discussed further in Section 12.9.

Network synchronization is a specialized topic, which will not be covered in this book. In this chapter the emphasis will be on the *self-clocking* or *self-synchronizing* systems, in which the synchronization information is obtained from the transmitted waveform itself, rather than utilizing special synchronizing bits. This is often referred to simply as *synchronous transmission*. A characteristic of synchronous transmission is that each bit in the message suffers the same transmission delay. In the next section, very brief mention is made of another form of transmission known as *asynchronous transmission*, which utilizes special codes for synchronization.

12.3 Asynchronous Transmission

Asynchronous means, rather generally, *not synchronous* and is used to denote digital systems where there is no synchronization between transmit and receive clocks. In asynchronous transmission, codewords are preceded by a start symbol and followed by a stop symbol. These frame individual codewords, which in turn

represent individual discrete characters. Examples of two binary codes used for asynchronous transmission are shown in Fig. 12.3.1.

Figure 12.3.2 shows the structure for these 5-bit and 8-bit codes. The idle condition on the line is always high, so that the start of a codeword transmission is always indicated by a waveform transition from high to

Character	Bit position	Bit position				
		1	2	3	4	5
	12345	76	76	8	76	54321
1 A -	11000					00000
2 B ?	10011					00001
3 C :	01110					00010
4 D WRU	00010					00011
5 E 3	10000					00100
6 F %	10110					00101
7 G @	01011					00110
8 H £	00100					00111
9 I 8	01100					01000
10 J BELL	11010					01001
11 K (11110					01010
12 L)	01001					01011
13 M .	00111					01100
14 N ,	00110					01101
15 O 9	00011					01110
16 P Ø	01101					01111
17 Q 1	00111					
18 R 4	01010					
19 S ,	10100					
20 T 5	00001					
21 U 7	11100					
22 V =	01111					
23 W 2	11001					
24 X /	10111					
25 Y 6	10101					
26 Z +	10001					
27 CARR. RETURN	00010					
28 LINE FEED	01000					
29 FIGS. SHIFT	11011					
30 LETT. SHIFT	11111					
31 SPACE	00100					
32 BLANK	00000					

76 = 00 group-control codes
 76 = 11 group-lower case alpha
 and special marks

(a)
(b)

Figure 12.3.1 (a) CCITT-2 and (b) ASCII asynchronous codes. (Courtesy Howard W. Sams and Company, Inc.)

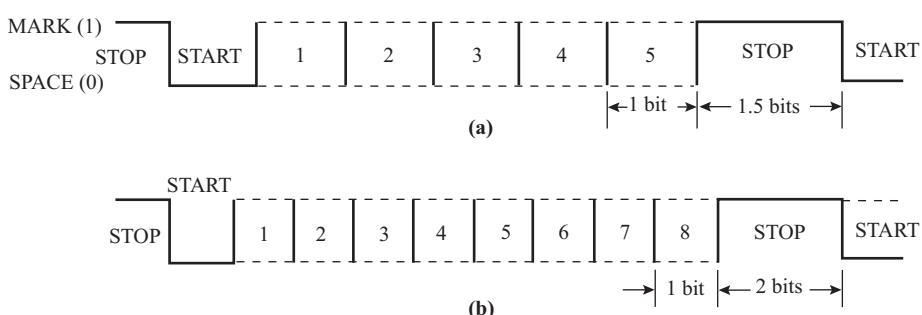


Figure 12.3.2 Character structure for (a) the CCITT-2 code and (b) the ASCII code.

low, and the end of a codeword transmission by a transition from low to high. It is necessary for the receiver to be able to keep track of the number of bits between start and stop symbols, so local synchronization is necessary. Note, however, that the receiver timing does not have to be synchronized to the transmitter timing. Asynchronous transmission finds greatest use where data are generated sporadically, an example being the output from a manually operated keyboard.

Another level of asynchronous transmission can be found in *integrated services digital networks* (ISDN), which provide end-to-end digital transmission. The European authorities presently have the objective of providing broadband data channels for ISDN using the *asynchronous transfer mode* (ATM). In this mode, *virtual channels* are created by grouping the high-speed data into packets which are then transmitted asynchronously over a number of lower speed channels, and reassembled at the receiving end.

12.4 Probability of Bit Error in Baseband Transmission

Thermal noise, discussed in Chapter 4, tends to degrade a communications system, and in digital communications, the degradation takes the form of errors introduced into a bit stream. Although the noise may originate at a number of points in the system, it may be represented as a noise voltage at the receiver input, as was done for AM and FM receivers. Figure 12.4.1(a) shows the receiver for a baseband digital system including noise.

The received signal is shown as $v(t)$, the noise voltage referred to the receiver input as $n(t)$, and the noisy signal is the sum of these. The noisy signal is passed through a low-pass filter to limit the noise. The filter has a transfer function $H_R(f)$, which may also form part of the overall filtering needed to eliminate intersymbol

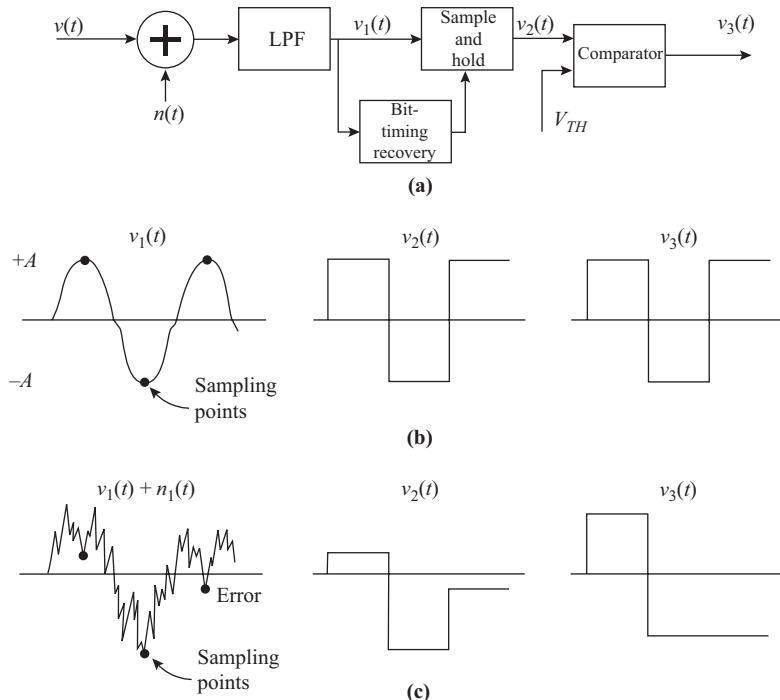


Figure 12.4.1 (a) Digital baseband receiver including noise. (b) Signal conditions without noise. (c) Signal conditions with noise, showing how an error can occur.

interference (see Section 3.6). The received signal after filtering is shown as $v_1(t)$. This is sampled at the bit rate by means of the sample-and-hold circuit. The timing of the sampling circuit is critical and will be discussed shortly. For the present it will be assumed that the sampling is synchronized for optimum reception. This means that if no noise was present the received pulses would be sampled at or near their peak values, as shown in Fig. 12.4.1(b). To illustrate the process, a polar signal is shown where the peak values $\pm A$ represent binary **1** and **0**, respectively. The sample-and-hold circuit produces output levels, shown as $v_2(t)$ and these in turn are compared to a threshold level. Received levels above the threshold result in a $+A$ output pulse and below the threshold a $-A$ output pulse, shown as $v_3(t)$.

There is a finite *probability* that noise will introduce an error, as shown in Fig. 12.4.1(c), and for digital systems, an important design parameter is the *probability of bit error*. How good or bad a system is in relation to the bit-error probability can only be established by experience. For example, for acceptable-quality speech signals the bit-error probability should be no more than $P_{be} \cong 10^{-5}$, while some data transmission systems may require values of $P_{be} \cong 10^{-8}$ or less. A probability of bit error of 10^{-5} means that, on average, 1 bit in every 100,000 will be in error. The probability of bit error is also referred to as the *bit-error rate*, denoted by BER.

Calculation of the bit-error probability requires a knowledge of the statistics of the noise and some details of the detection process. As shown, the sampled voltage is compared to a threshold voltage V_{TH} . If the sampled voltage is less than the threshold, the output from the comparator is low, indicating a binary **0**. If it is above the threshold, the output is high, indicating a binary **1**. Choice of the threshold is important, and where the transmitted bits are equiprobable, the threshold is set at the average voltage of the received baseband signal. For the polar waveform this is simply zero, and for a unipolar waveform of peak value A , it is $A/2$. Thus, in either case, positive-going pulses result in a high output and negative-going pulses in a low output. The bit error probability becomes a function of the *peak swing* about the threshold level. Denoting this by V_s , then for the binary polar waveform

$$\begin{aligned} V_s &= V_{\text{peak}} - V_{TH} \\ &= A - 0 \\ &= A \end{aligned} \quad (12.4.1)$$

and for the binary unipolar waveform

$$\begin{aligned} V_s &= V_{\text{peak}} - V_{TH} \\ &= A - \frac{A}{2} \\ &= \frac{A}{2} \end{aligned} \quad (12.4.2)$$

The noise statistics are determined by the manner in which the noise is generated, and a widely used model is Gaussian noise. For example, the thermal noise described in Chapter 4 has Gaussian statistics. The statistical details will not be given here, but the noise can be assumed to have a zero-mean value and a root-mean-square voltage V_n . The average probability of a bit error takes into account the probability that a transmitted **1** will be erroneously detected as a **0**, and a transmitted **0** as a **1**. The probability is given by a statistical function known as the *complementary error function*, denoted by $\text{erfc}(\cdot)$.

$$P_{be} = \frac{1}{2} \text{erfc} \left(\frac{V_s}{\sqrt{2}V_n} \right) \quad (12.4.3)$$

This function is generally available in tabular and graphical form as shown in Fig. 12.4.2.

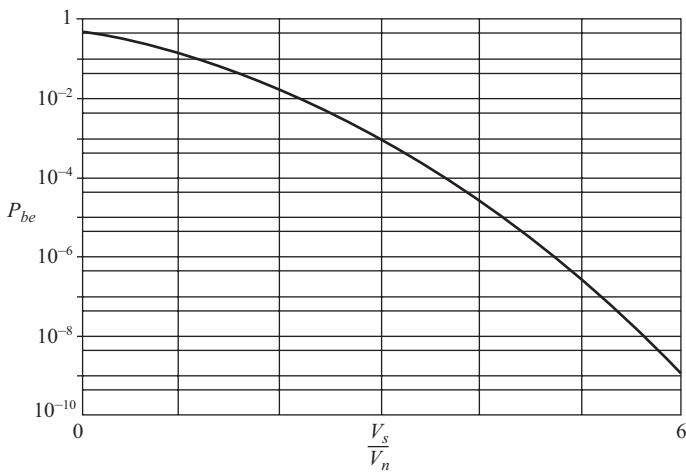


Figure 12.4.2 Bit-error probability function.

Mathematical computer programs such as Mathcad include some form of error function as one of the built-in functions. In some instances the built-in function is a standard *error function* $\text{erf}(\cdot)$, related to the complementary error function as

$$\text{erfc}(x) = 1 - \text{erf}(x) \quad (12.4.4)$$

EXAMPLE 12.4.1

For a binary polar waveform, the received signal-to-noise voltage ratio is $V_s/V_n = 4$. Determine values of (a) $\text{erf}(4/\sqrt{2})$, (b) $\text{erfc}(4/\sqrt{2})$, and (c) the bit error probability.

SOLUTION (a) Using Mathcad, the value of $\text{erf}(4/\sqrt{2}) \approx 0.99994$

$$(b) \therefore \text{erfc}(4/\sqrt{2}) = 1 - 0.99994 = 6 \times 10^{-5}$$

$$(c) P_{be} = \frac{1}{2} \times 6 \times 10^{-5} = 3 \times 10^{-5}$$

At the risk of confusing matters, yet a third function must be mentioned that is often used to specify bit-error probability. This is known as the *Q-function*, and it is related to the complementary error function as

$$Q(x) = \frac{1}{2} \text{erfc} \left(\frac{x}{\sqrt{2}} \right) \quad (12.4.5)$$

It follows therefore, that in terms of the Q-function

$$P_{be} = Q \left(\frac{V_s}{V_n} \right) \quad (12.4.6)$$

It will be seen that the functions $\text{erf}(x)$, $\text{erfc}(x)$, and $Q(x)$ are interrelated and are simply different ways of expressing the same information. In this text, to avoid confusion, the $\text{erfc}(x)$ function will be used.

EXAMPLE 12.4.2

For a binary unipolar signal, the received signal voltage has a maximum value of 4 mV. The rms noise voltage is 0.5 mV. Determine the bit-error probability.

SOLUTION $A = 4$ mV and therefore the threshold voltage for the comparator should be $V_{TH} = A/2 = 2$ mV. The voltage swing about threshold is $V_s = 2$ mV, from which

$$P_{be} = \frac{1}{2} \operatorname{erfc} \left(\frac{2}{\sqrt{2} \times 0.5} \right) = 3.17 \times 10^{-5}$$

To bring out more clearly the advantage of polar transmission over unipolar transmission, a comparison of the bit-error probabilities can be made for the same received signal-to-noise ratio in each case. As shown in Chapter 4, the signal-to-noise ratio can be expressed in terms of voltages as

$$\frac{S}{N} = \left(\frac{V_{rms}}{V_n} \right)^2 \quad (12.4.7)$$

where V_{rms} is the root-mean-square signal voltage. For the random polar waveform of rectangular pulses of amplitude A_{pol} , $V_{rms} = A_{pol}$, and as shown previously, the swing about the threshold for the polar waveform is $V_s = A_{pol}$. Hence

$$\begin{aligned} \frac{V_s}{V_n} &= \frac{V_{rms}}{V_n} \\ &= \sqrt{\frac{S}{N}} \end{aligned} \quad (12.4.8)$$

Thus Eq. (12.4.3) for the probability of bit error for polar transmission becomes

$$P_{be} = \frac{1}{2} \operatorname{erfc} \left(\frac{1}{\sqrt{2}} \sqrt{\frac{S}{N}} \right) \quad (12.4.9)$$

For the unipolar waveform of pulse height A_{uni} , $V_{rms} = A_{uni}/\sqrt{2}$ and the swing about threshold is $V_s = A_{uni}/2 = V_{rms}/\sqrt{2}$. Hence

$$\begin{aligned} \frac{V_s}{V_n} &= \frac{V_{rms}}{\sqrt{2}V_n} \\ &= \sqrt{\frac{1}{2} \frac{S}{N}} \end{aligned} \quad (12.4.10)$$

and the corresponding probability of bit error, Eq. (12.4.3), for the unipolar transmission is

$$P_{be} = \frac{1}{2} \operatorname{erfc} \left(\frac{1}{2} \sqrt{\frac{S}{N}} \right) \quad (12.4.11)$$

The following example illustrates the difference between these two expressions for bit-error probability.

EXAMPLE 12.4.3

Compare the bit-error probabilities for polar and unipolar transmissions for which the received S/N ratio is 9 dB in each case.

SOLUTION First convert the decibel value to a power ratio: 9 dB = 7.94.

$$\text{Polar: } P_{be} = \frac{1}{2} \operatorname{erfc} \left(\frac{1}{\sqrt{2}} \sqrt{7.94} \right) \cong 2.5 \times 10^{-3}$$

$$\text{Unipolar: } P_{be} = \frac{1}{2} \operatorname{erfc} \left(\frac{1}{2} \sqrt{7.94} \right) \cong 2.5 \times 10^{-2}$$

AMI encoding is like unipolar encoding but with two thresholds, one at $+A/2$ and the other at $-A/2$. A signal level falling between these will be interpreted as a binary **0**, and if falling outside these, whether on the positive or negative side, as a binary **1**. The resulting bit-error probability is given by

$$P_{be} = \frac{3}{4} \operatorname{erfc} \left(\frac{1}{2} \sqrt{\frac{S}{N}} \right) \quad (12.4.12)$$

12.5 Matched Filter

The low-pass filter shown in Fig. 12.4.1 can be designed to maximize the received ratio V_s/V_n , it being known as a *matched filter* under these conditions. Consider first the situation where the received pulses are *time limited*. This means that each pulse has a definite cutoff point beyond which it is zero, and overlap of pulses transmitted in sequence can be avoided. In other words, by time-limiting the pulses intersymbol interference (ISI) is avoided, and the only design criterion for the matched filter is maximizing the signal-to-noise voltage ratio. To do this, the transfer function of the matched filter is designed to be inversely proportional to the noise spectral density, with the result that the noise spectrum at the output of the filter tends to be flat. The transfer function is also made proportional to the voltage spectrum density of the signal pulse, which results in an output pulse that is more peaked than that at the input, and sampling takes place at the peak. The filter in effect distorts the received pulse such that it has a well-defined maximum at the sampling instant, while reducing the noise through shaping the spectrum.

The detailed analysis of the matched filter is too advanced to be included here, but the results can be stated. Before doing so, two important parameters, the average bit energy and the noise spectrum density for white noise, need to be defined. Let T_b represent the bit duration, and P_R be the average power in the received signal; then the average bit energy is

$$\begin{aligned} E_b &= P_R T_b \\ &= \frac{P_R}{R_b} \end{aligned} \quad (12.5.1)$$

If the binary **1**'s and **0**'s are generated with equal probability by the source, then the average bit energy is E_b for each type of bit. As shown in Chapter 3, a higher density of **1**'s is sometimes introduced into the

binary waveform to assist in bit-timing recovery, in which case the energy for each type of bit will differ. However, to illustrate the basic principles of matched filtering, equiprobable bits will be assumed.

By white noise is meant noise that has a flat frequency spectrum (see Chapter 4), the one-sided noise spectral density being given by

$$N_o = kT_s \text{ joules} \quad (12.5.2)$$

Here, T_s is the noise temperature referred to the receiver input, and k is Boltzmann's constant (J_s is used for temperature rather than T_s to avoid confusion with T used for time). Under these conditions, the matched filter yields a maximum value of the signal-to-noise voltage ratio given by

$$\text{Unipolar: } \left(\frac{V_s}{V_n} \right)_{\max} = \sqrt{\frac{E_b}{N_o}} \quad (12.5.3)$$

$$\text{Polar: } \left(\frac{V_s}{V_n} \right)_{\max} = \sqrt{\frac{2E_b}{N_o}} \quad (12.5.4)$$

It is important to note that this maximum value depends on the energy in the pulse and not the pulse shape. The minimum bit-error probability, obtained when the signal-to-noise voltage ratio is maximized, is

$$\text{Unipolar: } P_{be} = \frac{1}{2} \operatorname{erfc} \sqrt{\frac{E_b}{2N_o}} \quad (12.5.5)$$

$$\text{Polar: } P_{be} = \frac{1}{2} \operatorname{erfc} \sqrt{\frac{E_b}{N_o}} \quad (12.5.6)$$

The ratio E_b/N_o is a key parameter in the design of digital communications systems. Note that it is a dimensionless ratio since both quantities have dimensions of joules, and it is normally expressed in decibels, where

$$\frac{E_b}{N_o} \text{ dB} = 10 \log \frac{E_b}{N_o} \quad (12.5.7)$$

EXAMPLE 12.5.1

A binary unipolar signal has an average power of 6 pW, and the pulse duration is 0.02 μs. The equivalent noise temperature at the receiver input is 550 K. Determine the bit-error probability.

SOLUTION The average energy per pulse is $E_b = 6 \times 10^{-12} \times 0.02 \times 10^{-6} = 1.2 \times 10^{-19}$ J. The noise spectral density is $N_o = 1.38 \times 10^{-23} \times 550 = 7.59 \times 10^{-21}$ J. Hence $E_b/N_o = 15.81$ or ≈ 24 dB.

$$P_{be} = \frac{1}{2} \operatorname{erfc} \sqrt{\frac{15.81}{2}} = 3.5 \times 10^{-5}$$

Because of the importance of the relationship between P_{be} and E_b/N_o , the graph of Fig. 12.4.2 is replotted in Fig. 12.5.1 with E_b/N_o as abscissa.

The matched filter can be built in a number of ways, but each design is quite specific to the digital waveform being transmitted. A popular form of matched filter for binary waveforms where the received pulse is close to rectangular is shown in Fig. 12.5.2. The low-pass filter limits the noise bandwidth without significantly

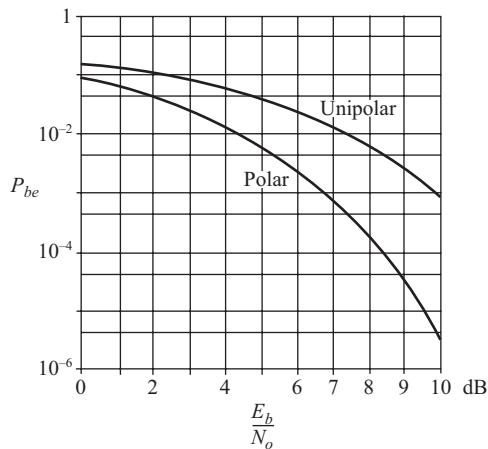


Figure 12.5.1 Bit-error probability as a function of E_b/N_o in decibels for unipolar and polar waveforms. Equal received powers are assumed in each case.

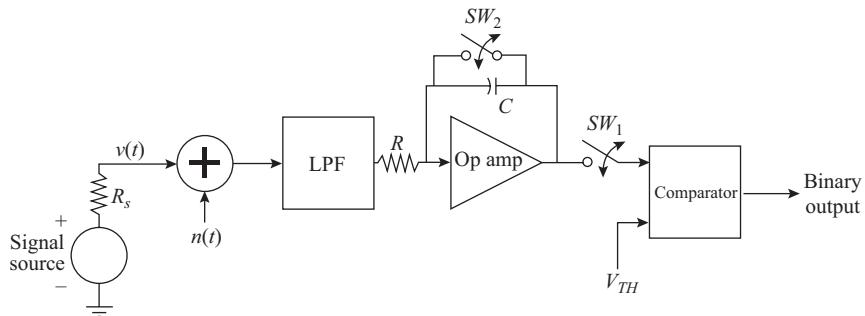


Figure 12.5.2 Matched filter for rectangular pulses.

affecting the signal pulse shape. The assumption of rectangular pulses means that a comparatively wide bandwidth transmission system is being used and intersymbol interference is avoided (see Section 3.6).

The matched filter shown in Fig. 12.5.2 is known as an *integrate and dump* circuit. The integrator integrates the incoming signal over each bit period. Accurate synchronization of switches SW_1 and SW_2 is required to ensure that integration starts at the beginning of a bit period and finishes at the end of the bit period. Switch SW_1 is momentarily closed as the integration is completed to allow the output to be sampled while the capacitor is fully charged. In this way, the maximum signal level is reached. Immediately after sampling, switch SW_2 is closed momentarily to discharge the capacitor (the charge is dumped), which resets the initial conditions for the next bit. For a rectangular pulse of level A, the sampled signal voltage (excluding the noise) is

$$V_{\text{out}} = A \frac{T_b}{\tau} \quad (12.5.8)$$

where $\tau = RC$ is the integrator time constant. To obtain the sampled output signal in terms of the input signal power, consider the input to the receiver to be a voltage source of internal resistance R_s providing an available input power P_R . For a unipolar stream of equiprobable bits, where 1's are a rectangular pulse of +A volts, and

0's are zero volts, the rms voltage is $A / \sqrt{2}$) (*coincidentally* the same as for a sine wave). Hence the maximum power transfer theorem yields.

$$P_R = \frac{A^2}{8R_s} \quad (12.5.9)$$

But, as known, $E_b = P_R T_b$, and combining these results gives, for the sampled output voltage,

$$\text{Unipolar: } V_{\text{out}} = \frac{\sqrt{8R_s E_b T_b}}{\tau} \quad (12.5.10)$$

For the polar waveform, the rms voltage is A , and hence

$$\text{Polar: } V_{\text{out}} = \frac{\sqrt{4R_s E_b T_b}}{\tau} \quad (12.5.11)$$

The *swing* about the threshold depends on the nature of the input signal, whether it is unipolar or polar. As shown previously, for unipolar signals the threshold is set at one-half the output voltage, and hence the swing is

$$\begin{aligned} \text{Unipolar: } V_s &= V_{\text{out}} - V_{TH} \\ &= \frac{\sqrt{2R_s E_b T_b}}{\tau} \end{aligned} \quad (12.5.12)$$

For polar waveforms, the threshold is zero, and therefore

$$\text{Polar: } V_s = \frac{\sqrt{4R_s E_b T_b}}{\tau} \quad (12.5.13)$$

Finding the mean-square noise output voltage is more difficult. The transfer function for the integrator that integrates over a period T_b is

$$H(f) = \frac{1 - e^{-j\omega T_b}}{j\omega\tau} \quad (12.5.14)$$

This rather formidable expression can be used to find the equivalent noise bandwidth as described in Section 4.2. The integration will not be carried out here, but the result is that $B_N = T_b/(2\tau^2)$, and the noise voltage output originating from the source of internal resistance R_s is

$$\begin{aligned} V_n^2 \text{ out} &= 4R_s k T_s B_N \\ &= 4R_s N_o B_N \end{aligned} \quad (12.5.15)$$

Substituting for B_N and simplifying yields

$$V_n \text{ out} = \frac{\sqrt{2R_s N_o T_b}}{\tau} \quad (12.5.16)$$

Hence the maximum signal-to-noise voltage ratio is

$$\text{Unipolar: } \left(\frac{V_s}{V_n} \right)_{\text{max}} = \sqrt{\frac{E_b}{N_o}} \quad (12.5.17)$$

$$\text{Polar: } \left(\frac{V_s}{V_n} \right)_{\text{max}} = \sqrt{\frac{2E_b}{N_o}} \quad (12.5.18)$$

Thus, the integrate and dump circuit yields the maximum ratios as given by Eqs. (12.5.3) and (12.5.4).

12.6 Optimum Terminal Filters

The matched filter described in the previous section is optimum in the sense that it maximizes the signal-to-noise ratio at the input to the decision detector, but it requires an input signal for which ISI is absent: that is, the pulses do not overlap to any significant extent. However, as shown in Section 3.6, where the bandwidth of the transmission system is limited, pulses may overlap, and the receive filter must shape the incoming pulses so that ISI is avoided. For a given channel response and given input pulse shape at the transmitter, the frequency responses of the transmit and receive filters together determine the final pulse shape at the receiver, as shown by Eq. (3.7.1).

Where these filters are designed to maximize the received signal-to-noise ratio and eliminate ISI, they are known as *optimum terminal filters*. Under certain conditions it is possible to use identical filters at the transmitter and receiver, which affords considerable savings in production and maintenance costs. In general, the maximum signal-to-noise ratio obtained with optimum filters will be less than that achieved using a matched filter at the receiver. In practice, the receive filter may be designed in two sections. One section meets the optimum terminal requirements under normal (or so-called standard) operating conditions, and a second section, known as an *equalizer filter*, is used to fine-tune the overall response, to take into account unpredictable variations in the communications link.

Where the channel introduces a constant attenuation and constant time delay, as for example in digital radio systems, the optimum terminal filter is in fact a matched filter, and the minimum bit-error rate is achieved.

12.7 Bit-timing Recovery

It is seen that accurate bit timing is needed at the receiver in order to be able to sample the received waveform at the optimum points. Although in some systems a clocking signal is transmitted as a separate component along with the information signal, the most common arrangement is where the clocking signal is extracted from the information signal itself. These are known as *self-clocking* or *self-synchronizing* systems. As shown in Chapter 3, line waveforms that have a high density of zero crossings can be devised, and a zero-crossing detector can be used at the receiver to recover the clocking signal. In practice, the received waveform is often badly distorted by the frequency response of the transmission link and by noise, and the design of the bit-timing recovery is quite complicated. In most instances the spectrum of the received waveform will not contain a discrete component at the clock frequency. However, it can be shown that a periodic component at the clocking frequency is present in the squared waveform for digital signals (unless the received pulses are exactly rectangular, in which case squaring simply produces a dc level for a binary waveform). A commonly used baseband scheme is shown in block schematic form in Fig. 12.7.1.

As mentioned in the previous section, the optimum terminal filter is in two sections, shown as *A* and *B* in Fig. 12.7.1. The signal for the bit-timing recovery is tapped from the junction between *A* and *B* and passed along a separate branch, which consists of a filter, a squaring circuit, and a band-pass filter that is sharply tuned to the clock frequency component present in the spectrum of the squared signal. This is then used to synchronize the clocking circuit, the output of which clocks the sampler in the detector branch. (Further details of this arrangement will be found in “Carrier and Bit Synchronization in Data Communication—A Tutorial Review,” by L. E. Franks, *IEEE Transactions on Communications*, Vol. COM28, no. 8, August 1980.)

Another arrangement referred to as the *early-late gate* is shown in Fig. 12.7.2. This method does not rely on there being a clocking component in the spectrum of the received waveform; rather the circuit utilizes a feedback loop in which magnitude changes in the outputs from matched filters control the frequency of a local clocking circuit.

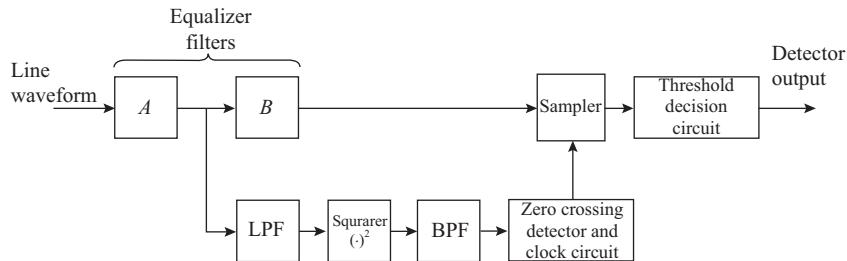


Figure 12.7.1 Scheme for bit-timing recovery.

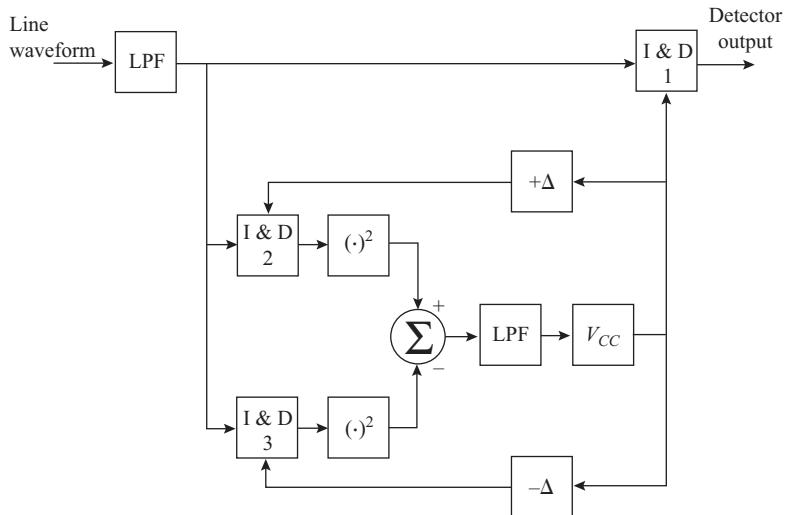


Figure 12.7.2 Early-late gate circuit for bit-timing recovery.

With a rectangular pulse waveform, the matched filters take the form of integrate-and-dump filters, indicated by the I&D blocks in Fig. 12.7.2, although the circuit is readily modified to suit other pulse shapes. VCC is a voltage-controlled clock, which is similar in operation to a voltage-controlled oscillator (VCO), except that the output is a clocking signal rather than a sinusoid. Assuming for the moment that the VCC is correct, the timing pulses going to I&D 2 and 3 circuits are symmetrically advanced and retarded by an amount Δ , which is small compared to the bit period. This means that the integrate-sample-dump operations of I&D 2 start and finish early compared to the correct times, while the corresponding operations of I&D 3 start and finish late by an equal amount. The squaring circuits ensure that the output waveform from each I&D circuit consists of positive-going impulses of varying magnitudes. The output from one of the I&D circuits is inverted so that the output from the summing junction is a waveform consisting of alternating positive and negative impulses. The average of this waveform, obtained by low-pass filtering as shown, is applied as a control voltage to the VCC, and, for example, if the amplitudes of the impulses are equal, zero control voltage results.

It is a property of integrate-and-dump filters that the magnitude of the output has the same value for a symmetrical displacement of $\pm\Delta$ about the correct timing, and hence the average output from the summing junction will be zero when the clock is correct. If however, the clock itself is off by an amount δ , then the displacement is asymmetrical. For example, if the clock is in advance by δ , then the timing of I&D 2 will be

advanced by $\Delta + \delta$, and that of I&D 3 will be retarded by $\Delta - \delta$. This asymmetry results in the outputs no longer being equal on average, and so a finite control voltage is applied to the VCC to reduce the error δ .

Bit-timing recovery is one aspect of synchronization. Where the information is transmitted in blocks or frames, *framing codewords* are interleaved with the data stream so that the blocks can be identified.

12.8 Eye Diagrams

An *eye diagram* provides a useful way of examining a digital signal for the effects of noise, ISI, and timing jitter. It consists of an oscilloscope display in which the time base is synchronized to the bit rate, and the digital waveform provides the vertical deflection signal. The time base is usually limited to a width of two symbol periods (bit periods if a binary signal is being displayed) and is triggered to start at the center of peaks (the sampling points).

Figure 12.8.1(a) shows part of a random binary waveform and Fig. 12.8.1(b) the corresponding eye diagram. The letters on the waveform are shown for the purpose of illustrating how the eye diagram is traced out. Starting at A, the two periods covered by A–B–C–D–E are traced out. At E, the oscilloscope is retriggered, so that the E–F–G pattern starts at the same point as A. At G, retriggering again occurs, and the next two periods are traced out. Over many bit periods, many of the traces coincide, and this, along with the normal persistence of the tube phosphor, creates the appearance of a stationary pattern similar to that shown in Fig. 12.8.1(b).

The eye diagram shown in Fig. 12.8.1(b) illustrates the ideal condition where the waveforms of overlapping bits are identical so that no blurring of the traces occurs. In practice, ISI if present will alter the pulse shapes from bit to bit, and noise will also prevent overlapping pulses from having identical shapes. The result is that the traces become smeared out, as shown in Fig. 12.8.2.

As Fig. 12.8.2 shows, the effect of noise and ISI is to “close the eye,” which reduces the noise margin. The noise margin is the difference between the 0-V level and the lower limit at the peaks, where the signal would normally be sampled. The slope of the waveform around the edges of the eye indicates the sensitivity of the system to timing errors. If timing errors move the sample point away from center, but the slope is small, there will be little change in the sampled value. If, however, the slope is large, a small timing error can result in a large change of level, so the sampled value may be in error.

Pulse distortion also blurs the zero crossover points. This reduces the range over which sampling may take place. In addition, the detection of zero crossover points is used to generate the clocking signal for bit timing at

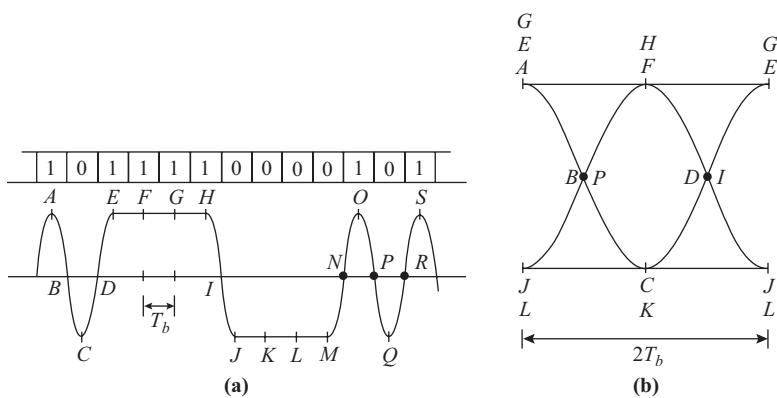


Figure 12.8.1 (a) Part of a random binary waveform. (b) Corresponding eye diagram.

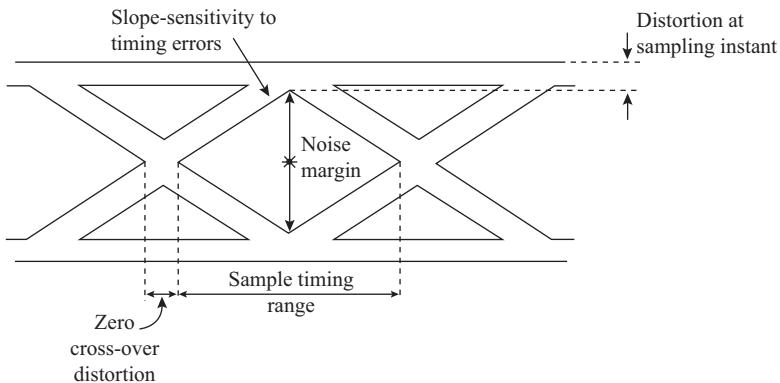


Figure 12.8.2 Showing how the impairments are determined from an eye diagram.

the receiver. The distortion occurring at the crossover points introduces uncertainty in the bit timing, giving rise to the *timing jitter* mentioned previously.

12.9 Digital Carrier Systems

Digital data may be modulated onto a carrier wave. In general, the function of the carrier is to shift the digital data from the baseband region into a pass-band region of the frequency spectrum. Passband signals are needed, for example, in radio transmission and also for frequency multiplexing on lines. The modulation methods of amplitude, frequency, and phase, described in previous chapters, are all feasible for digital signals and in fact are usually easier to implement than is the case with analog signals. As with analog modulation, the carrier is a sinusoidal (or cosinusoidal) wave, and the waveforms for three common types of binary modulated carriers are shown in Fig. 12.9.1.

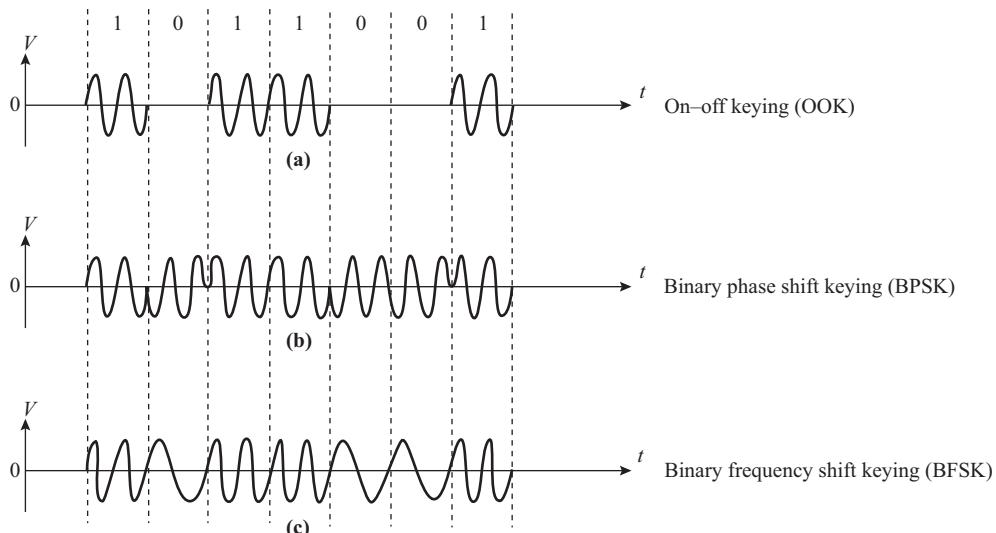


Figure 12.9.1 Binary modulated carriers: (a) binary amplitude shift keying (BASK), also known as on-off keying (OOK); (b) binary phase shift keying (BPSK), also known as phase reversal keying (PRK); (c) binary frequency shift keying (BFSK).

As shown in Section 12.5, a key parameter in digital systems is the E_b/N_0 ratio, where E_b is the average bit energy and N_0 the thermal noise power density. The average bit energy can be found from the average received power and the bit rate, as shown by Eq. (12.5.1). In the case of the carrier systems shown, assuming a carrier of rms value E_c and a load resistance of 1Ω , the average received power for the BPSK and BFSK waves is simply E_c^2 . For the OOK wave the average received power is $E_c^2/2$, because it is off half the time on average. This must be kept in mind when comparing binary modulated carriers. If the average received power is the same in each case, the bit energy will also be the same, but this requires that during the on periods for OOK the transmitted power must be doubled or the power amplifiers operated at a peak voltage level that is $\sqrt{2}$ times that for continuous carrier systems. If the comparison is made on the basis of the same value of received voltage, then the average power, and hence the bit energy, of the OOK modulation is half that of the other modulation methods.

Amplitude Shift Keying

With *amplitude* modulation the digital signal is used to switch the carrier between amplitude levels, and hence it is referred to as *amplitude shift keying* (ASK). The particular case of binary modulation is illustrated in Fig. 12.9.2, where the modulating waveform consists of unipolar pulses. Because in this particular case the carrier is switched on and off, the method is known as *on-off keying* (OOK), or sometimes as *interrupted continuous wave* (ICW) transmission.

Amplitude shift keying is fairly simple to implement in practice, but it is less efficient than angle-modulation methods, to be described shortly, and is not as widely used in practice. Applications do arise, however, in such diverse areas as emergency radio transmissions and fiber-optic communications (described in chapter 20). On-off keying of a radio transmitter may be achieved as shown in the block diagram of Fig. 12.9.3(a).

In Fig. 12.9.3(a), the carrier frequency is generated by a crystal oscillator, which is followed by a buffer amplifier to maintain good frequency stability. Again in the interests of maintaining good frequency stability, the oscillator frequency is usually lower than the required carrier frequency, and one or more frequency multiplier stages are necessary. The driver amplifier is a power amplifier that provides the required drive for the final RF amplifier, which is a class C stage. This is similar to the circuits described in Section 8.10. Although the keying circuit could be used to simply interrupt the current in the final amplifier by means of a “make-break” contact,

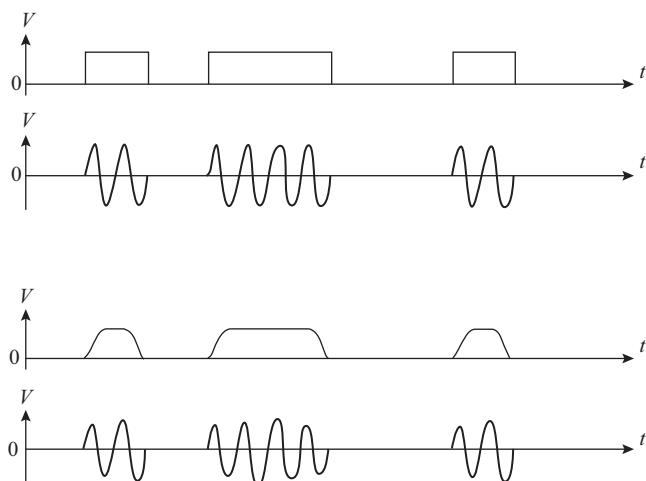


Figure 12.9.2 Amplitude shift keying (ASK) (a) with unipolar rectangular pulses and (b) with filtered pulses.

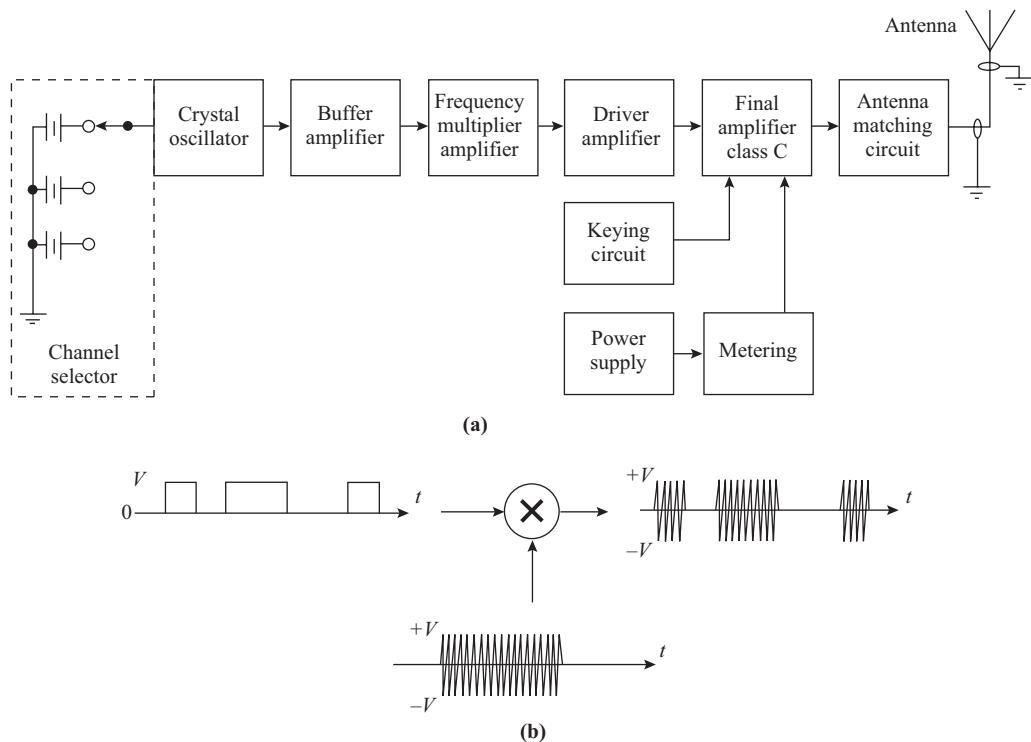


Figure 12.9.3 On-off keying achieved by (a) controlling the bias of a radio telegraph transmitter and (b) use of a multiplier circuit.

this could give rise to undesirable transients and would be avoided in high-power circuits. The more usual method is to use the keying signal to bias the class C into cutoff for the off binary bits. For radio transmission it is undesirable to have rapid changes in amplitude because these give rise to *sideband spatter*; and the digital modulating waveform is filtered to remove the sharp transitions so that the modulated waveform appears more as shown in Fig. 12.9.1(b). A simple *RC* filter may be used, and in practice this may have a time constant on the order of 2 ms.

Although keying could take place at a lower power stage in the transmitter, this is not usually done since the class C bias on the final amplifier is derived from the drive signal, and removal of this could result in excessive current rise in the final amplifier. Where only low power stages are involved, a product modulator may be used, as shown in Fig. 12.9.3(b). Denoting the unmodulated carrier by

$$e_c(t) = E_{c \max} \cos(2\pi f_c t + \phi_c) \quad (12.9.1)$$

and the binary modulating waveform as $e_m(t)$, then the modulated waveform is

$$e(t) = k e_m(t) \cos(2\pi f_c t + \phi_c) \quad (12.9.2)$$

Although this appears identical to the DSBSC expression given by Eq. (8.9.1), the difference is that a dc component is present in the unipolar waveform, and this results in a carrier component being present in the spectrum. For example, if the unipolar waveform consists of an alternating series ...1 0 1 0 1 0..., it appears just as the square wave of Fig. 2.7.2, and hence the modulated spectrum has side frequencies extending indefinitely on either side of the carrier, as shown in Fig. 12.9.4(a).

More generally, when the binary waveform is random, the baseband spectrum for the ac component is as shown in Fig. 3.4.3, and the modulated spectrum (in this case a spectrum density) is as shown in Fig. 12.9.4(b). Again, the dc component gives rise to a spike at the carrier frequency in the spectrum density plot. The presence of a component at the carrier frequency is important in the coherent detection of OOK waveforms, to be discussed shortly.

If B is the overall system bandwidth for the binary signal, the bandwidth for the modulated wave, B_T , is

$$B_T = 2B \quad (12.9.3)$$

In the particular case where raised-cosine filtering is used on the baseband signal, then from Eq. (3.7.3), the modulated bandwidth becomes

$$\begin{aligned} B_T &= 2 \frac{1 + \rho}{2T_b} \\ &= (1 + \rho)R_b \end{aligned} \quad (12.9.4)$$

where $R_b = 1/T_b$ is the transmitted bit rate. The ratio of bit rate to system bandwidth, the parameter defined by Eq. (3.7.11), becomes, for the OOK modulated wave

$$\begin{aligned} \alpha &= \frac{R_b}{B_T} \\ &= \frac{1}{1 + \rho} \end{aligned} \quad (12.9.5)$$

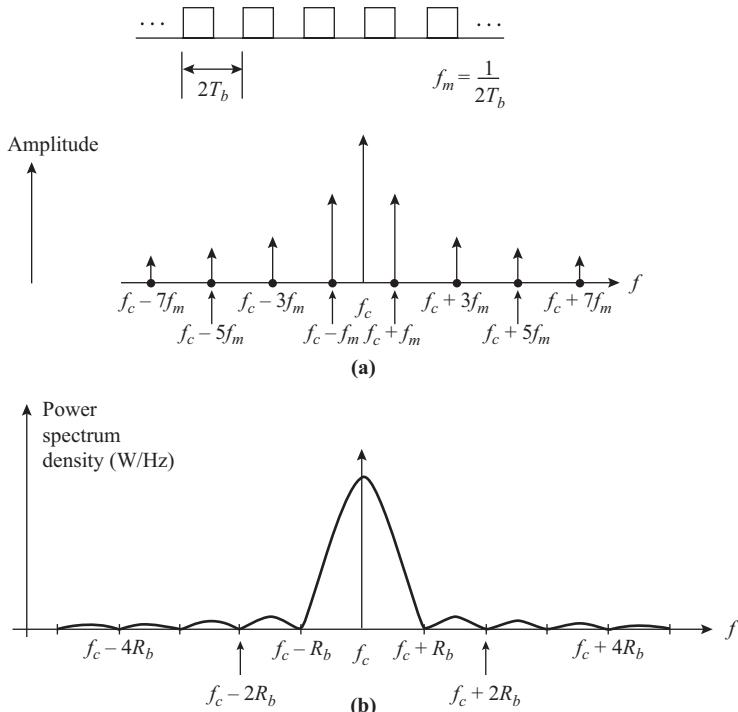


Figure 12.9.4 (a) Spectrum for OOK squarewave modulation. (b) More general picture of the OOK spectrum.

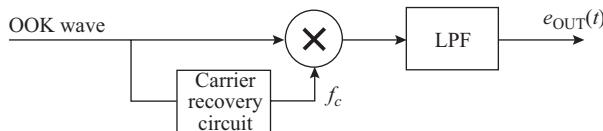


Figure 12.9.5 Synchronous demodulation of an OOK wave.

Thus, for the ideal bandwidth $\rho = 0$, OOK provides a rate of 1 bps/Hz, and for $\rho = 1$, the rate is 0.5 bps/Hz. Demodulation of the OOK waveform may take place using a simple envelope detector as described in Section 8.11. A more efficient method is to use *synchronous* detection. This method is illustrated in block schematic form in Fig. 12.9.5.

Synchronous detection requires a *carrier recovery* circuit, which is used to generate a local carrier component exactly synchronized to the transmitted carrier. As shown, the spectrum contains a component at the carrier frequency that can be used to phase lock the VCO in a PLL. Applying the locally generated carrier and the received signal to the multiplier results in an output

$$\begin{aligned} e_{out}(t) &= Ae(t) \cos(2\pi f_c t + \phi_c) \\ &= Ae_m(t) [\cos(2\pi f_c t + \phi_c)]^2 \end{aligned} \quad (12.9.6)$$

where A is an amplitude constant. Expanding the cosine squared term,

$$e_{out}(t) = Ae_m(t) \left[\frac{1}{2} + \frac{1}{2} \cos(4\pi f_c t + 2\phi_c) \right] \quad (12.9.7)$$

The second harmonic carrier term is easily removed by filtering, leaving as the output

$$e_{out}(t) = \frac{A}{2} e_m(t) \quad (12.9.8)$$

Thus the baseband signal is recovered. The constant term $A/2$ is easily allowed for by adjustment of gain. The synchronous detection just described is also referred to as *coherent detection*. The demodulation of the OOK wave can also be carried out by using an envelope detector, as described in Section 8.11 (this also being known as noncoherent or nonsynchronous detection). Once the baseband signal is recovered, it can be used to regenerate new pulses, or the digital information can be recovered, as shown in Fig. 12.4.1(a).

The coherent detector is more complicated than the envelope detector, but it results in a lower probability of error for a given signal-to-noise input. The analysis will not be presented here, but the results are that the coherent OOK detection has a probability of error identical to that for the baseband system, which for optimum detection is given by Eq. (12.5.5) as

$$P_{be} = \frac{1}{2} \operatorname{erfc} \sqrt{\frac{E_b}{2N_o}} \quad (12.9.9)$$

The *optimum* noncoherent detector requires that $E_b/N_o \gg 1$, and for this condition

$$P_{be} \approx \frac{1}{2} e^{-E_b/2N_o} \quad (12.9.10)$$

EXAMPLE 12.9.1

Calculate the bit-error probability for OOK using (a) synchronous carrier demodulation and (b) nonsynchronous carrier demodulation when the bit energy to noise density ratio is 10 dB.

SOLUTION $10 \text{ dB} = 10 : 1$ energy ratio, and therefore:

$$(a) P_{be} = \frac{1}{2} \operatorname{erfc} \sqrt{\frac{10}{2}} = 7.83 \times 10^{-4}$$

$$(b) P_{be} \approx \frac{1}{2} e^{-10/2} = 0.00337$$

In both cases, the bit energy is determined from the average carrier power. Let P_R represent the average received signal power; then for binary transmission

$$\begin{aligned} E_b &= P_R T_b \\ &= \frac{P_R}{R_b} \end{aligned} \quad (12.9.11)$$

A useful relationship can be derived between the E_b/N_0 ratio and the S/N ratio. The noise power at the receiver input is $P_N = N_o B_N$, where B_N is the noise bandwidth. Hence

$$\begin{aligned} \frac{E_b}{N_o} &= \frac{P_R}{R_b} \cdot \frac{B_N}{P_N} \\ &= \frac{S}{N} \cdot \frac{B_N}{R_b} \end{aligned} \quad (12.9.12)$$

For the situation where $B_N \approx B_T$, where B_T is the system bandwidth, then using Eq. (3.7.11), which defines $\alpha = R_b/B_T$, gives

$$\alpha \frac{E_b}{N_o} = \frac{S}{N} \quad (12.9.13)$$

Thus the product of the two significant parameters for digital transmission, R_b/B_T and E_b/N_o , is equal to the received signal-to-noise ratio.

Where the transmission channel may be assumed distortionless and where raised-cosine filtering is used such that the receiver filter is the square root of the raised-cosine characteristic (matched by a similar filter at the transmitter, as is often the case with radio transmission), it can be shown that the noise bandwidth at the receiver is $B_N = R_b$. In this case, Eq. (12.9.13) becomes

$$\frac{E_b}{N_o} = \frac{S}{N} \quad (12.9.14)$$

The usefulness of Eqs. (12.9.13) and (12.9.14) is that often the signal-to-noise ratio is the known quantity, while to calculate the bit-error probability, the E_b/N_o is the ratio required.

Frequency Shift Keying

With frequency modulation, usually referred to as *frequency shift keying* (FSK), the carrier frequency is shifted in steps or levels corresponding to the levels of the digital modulating signal. In the case of a binary signal, two carrier frequencies are used, one corresponding to the binary 0 and the other to a binary 1. In general, for binary modulation the carriers can be represented by

$$\begin{aligned} \text{Binary 0: } e_0(t) &= A_0 \cos(2\pi f_0 t + \alpha_0) \\ \text{Binary 1: } e_1(t) &= A_1 \cos(2\pi f_1 t + \alpha_1) \end{aligned} \quad (12.9.15)$$

The two carriers may be generated from separate oscillators, independent of one another, and this is indicated by separate subscripts for the amplitudes and fixed phase angles. The combined signal can therefore have discontinuities in amplitude and phase, which are undesirable. Alternatively, the modulation can be achieved by frequency modulating a common carrier, which prevents the discontinuities from occurring. The block schematics for the two methods are shown in Fig. 12.9.6.

Denoting the mean carrier frequency by f_c , then a binary 1 results in $f_1 = f_c + \delta f$, and a binary 0 in $f_0 = f_c - \delta f$, where $2\delta f$ is the difference between the two signaling frequencies. The modulated signal is given by

$$\text{Binary 0: } e_0(t) = A \cos 2\pi f_0 t$$

$$\text{Binary 1: } e_1(t) = A \cos 2\pi f_1 t \quad (12.9.16)$$

where without loss of generality the fixed phase angle has been set equal to zero for each signal. Where a single oscillator is frequency modulated by the digital signal, the method is referred to as *continuous phase frequency shift keying* (CPFSK). The modulation could be represented more concisely by Eq. (10.2.3) for FM, but a detailed analysis of such a waveform is very complicated. A reasonable picture of the spectrum can be obtained by utilizing a different approach. Figure 12.9.7 shows the waveform for the CPFSK wave where, for convenience, an integer number of carrier cycles per bit period is shown.

By treating the CPFSK wave as two OOK waves as shown in the figure, the spectrum for the OOK wave shown in Fig. 12.9.4 can be used for each, and the resultant spectrum is as sketched in Fig. 12.9.8(a). A special and important case of CPFSK known as *minimum shift keying* (MSK) occurs when $\delta f = R_b/4$. This is the minimum separation for which correlation between the two signaling waveforms is zero. It can be shown that for any closer spacing the correlation between the waveforms results in an increased probability of bit error.

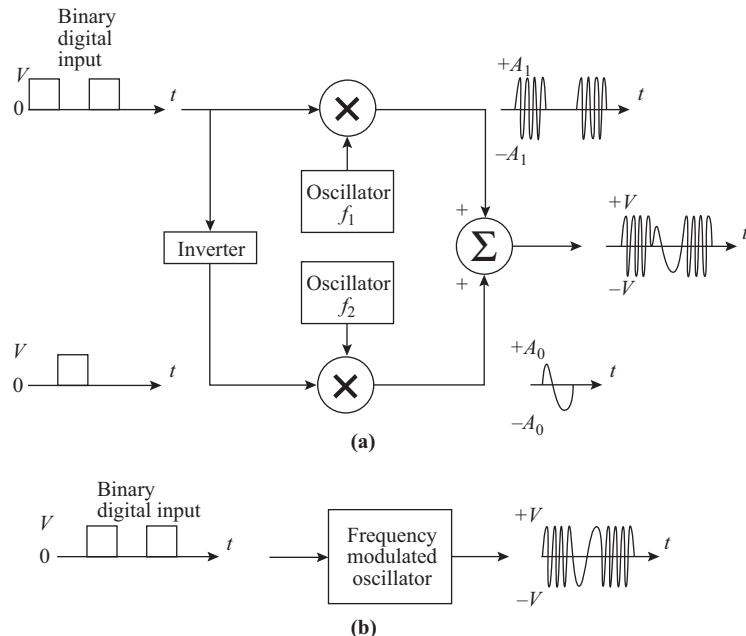


Figure 12.9.6 (a) Separate oscillator method of realizing FSK. (b) Single oscillator method, or CPFSK.

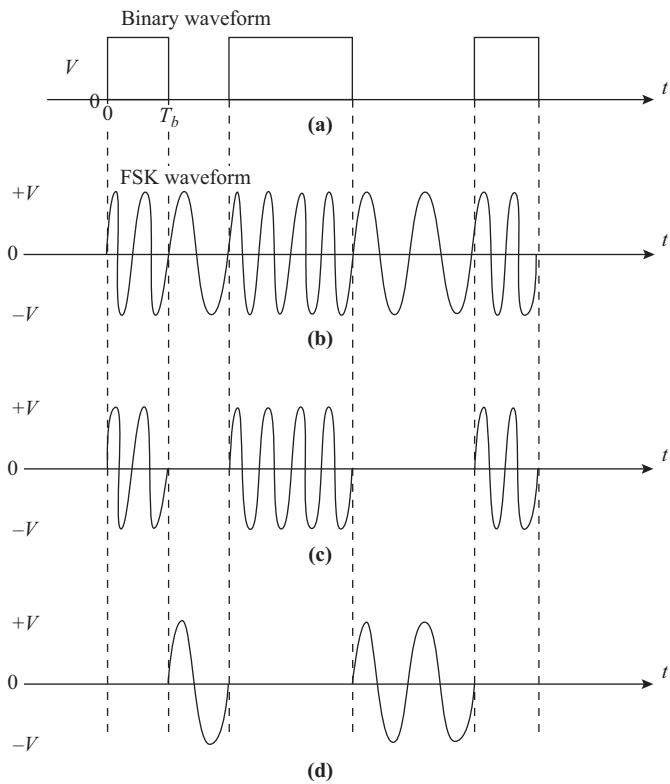


Figure 12.9.7 (a) Binary waveform used in FSK. (b) FSK waveform. (c) and (d) Resolution of (b) into two OOK waves.

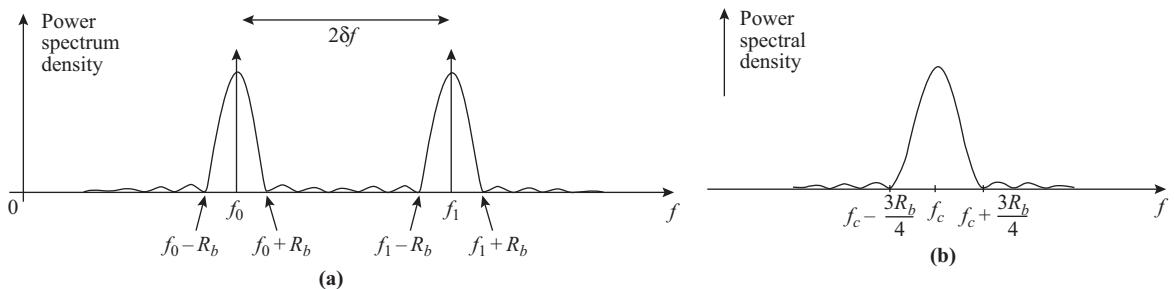


Figure 12.9.8 Power spectra for (a) CPFSK with widely spaced signaling frequencies and (b) MSK.

For MSK, if it is assumed that the spectrum beyond the first nulls can be ignored, the overall spectrum bandwidth is seen to be

$$B_T = -R_b \quad (12.9.17)$$

This gives a bps/Hz figure of merit of $\frac{1}{2}$. In practice, it is found that the rate of decay of the spectrum outside a bandwidth given by $B_T = R_b/2$ is very rapid, with the result that the bps/Hz figure can be improved to approximately unity.

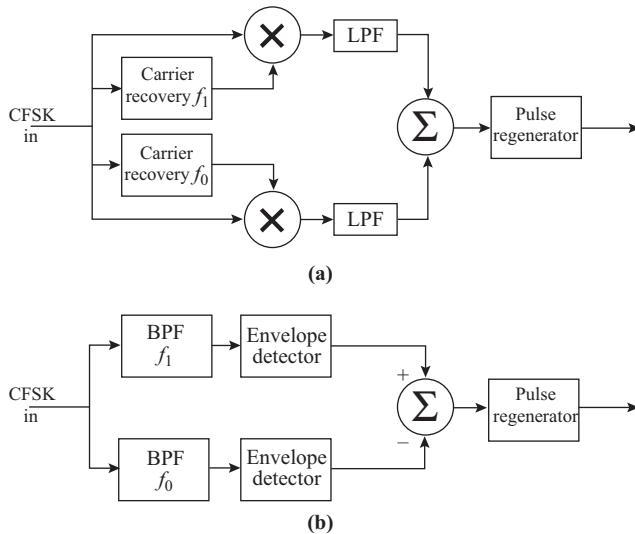


Figure 12.9.9 (a) Coherent and (b) noncoherent detection for CPFSK.

Because FSK appears as two OOK waves, the coherent receiver can be constructed by using two separate OOK coherent detectors, as shown in Fig. 12.9.9(a). The outputs are combined to form a polar binary signal, which, for optimum detection, is then passed to a matched filter. Correlation between the two signaling frequencies results in general in an increased probability of bit error, but with MSK the correlation is zero and the expression for bit-error probability is the same as that for OOK, which is repeated here for convenience:

$$P_{be} = \frac{1}{2} \operatorname{erfc} \sqrt{\frac{E_b}{2N_o}} \quad (12.9.18)$$

As pointed out in the introduction, when comparing OOK with CPFSK, it should be kept in mind that the bit energy with CPFSK is double that of OOK for the same carrier voltage levels at the receiver.

Noncoherent detection can also be used with FSK signals. Again, because FSK appears as two OOK waves, the noncoherent receiver need consist only of two separate paths with band-pass filters tuned to the individual frequencies, as shown in Fig. 12.9.9(b). Each filter is followed by an envelope detector, as described in Section 8.11. The outputs are combined to form a polar waveform, which is then passed as input to the pulse regenerator operating at zero voltage threshold. When properly adjusted for optimum performance, the probability of bit error for the noncoherent FSK detector is given by

$$P_{be} \cong \frac{1}{2} e^{-E_b/2N_o} \quad (12.9.19)$$

The noncoherent receiver is much simpler to build than the coherent receiver, and for many applications the degradation in bit-error probability is acceptable.

Phase Shift Keying

With *phase* modulation, usually referred to as *phase shift keying* (PSK), the binary signal is used to switch the phase between \$0^\circ\$ and \$180^\circ\$. It is also known as *phase reversal keying* (PRK). The modulated carrier is described by

$$e(t) = \begin{cases} E_c \max \cos(2\pi f_c t + \phi_c), & \text{binary 1} \\ E_c \max \cos(2\pi f_c t + \phi_c + 180^\circ), & \text{binary 0} \end{cases} \quad (12.9.20)$$

The circuit for implementing BPSK is a balanced modulator, as shown in Fig. 12.9.10(a).

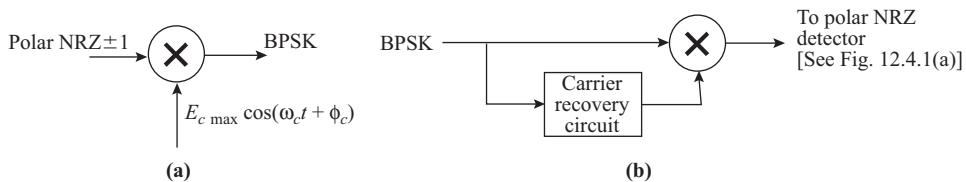


Figure 12.9.10 (a) Balanced modulator producing BPSK. (b) Detection of BPSK.

The modulating signal is polar NRZ, and when this has value $+1$, the modulated output is $+1 \times E_c \max \cos(2\pi f_c t + \phi_c) = E_c \max \cos(2\pi f_c t + \phi_c)$, and when it is -1 , the modulated output is $-1 \times E_c \max \cos(2\pi f_c t + \phi_c) = E_c \max \cos(2\pi f_c t + \phi_c + 180^\circ)$. Coherent detection must be used with BPSK since the envelope does not contain the modulating information. The coherent BPSK receiver is shown in Fig. 12.9.10(b).

The BPSK modulator is similar to the OOK modulator, the difference being that no dc component is present in the modulating waveform and therefore no carrier component is transmitted. The spectrum is shown in Fig. 12.9.11. This is similar to that shown in Fig. 12.9.4, but with no spike at the carrier frequency.

The BPSK wave has in effect a DSBSC spectrum. With raised-cosine filtering on the baseband signal, the bps/Hz figure of merit is also given by Eq. (12.9.5) as $1/(1 + \rho)$. Coherent detection of BPSK followed by matched filter detection results in a bit-error probability given by

$$P_{be} = \frac{1}{2} \operatorname{erfc} \sqrt{\frac{E_b}{N_o}} \quad (12.9.21)$$

Just as binary baseband signals can be reformatted as M -ary signals with a consequent reduction in transmitted bandwidth, so M -ary level modulation can be used to similar effect. *Quadrature phase shift keying* (QPSK) utilizes four distinct levels of phase shift and is a widely used form of multilevel modulation. In this method a serial-to-parallel converter is used to convert the binary signal $p(t)$ into two separate binary signals in which the bit period is doubled, as shown in Fig. 12.9.12(a) and (b). These two binary signals are labeled $p_i(t)$ for *in-phase* and $p_q(t)$ for *quadrature-phase* components, respectively. The in-phase component modulates a carrier to produce a BPSK signal, while the quadrature component modulates a carrier component shifted by 90° (hence the label *quadrature*), also to produce a BPSK signal. The two BPSK signals are added to produce the QPSK signal, the modulator states being as shown in Fig. 12.9.12(c). Thus the QPSK signal is equivalent to two BPSK signals, but with the carriers 90° out of phase with one another. Each BPSK waveform has a DSBSC spectrum, as described previously, and these spectra overlap. However, they do not interfere with one another because of the phase difference between the carriers. Thus QPSK signaling requires one-half the bandwidth of BPSK signaling for the same input bit rate in both cases, and the bps/Hz figure of merit is $2/(1 + \rho)$, where raised-cosine filtering is used.

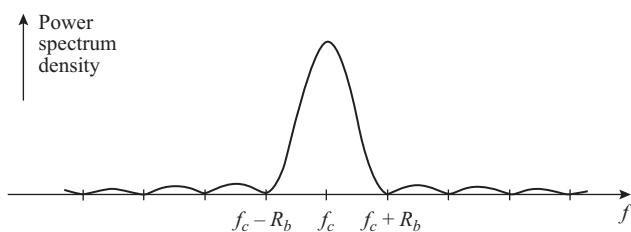


Figure 12.9.11 Spectrum for the BPSK wave.

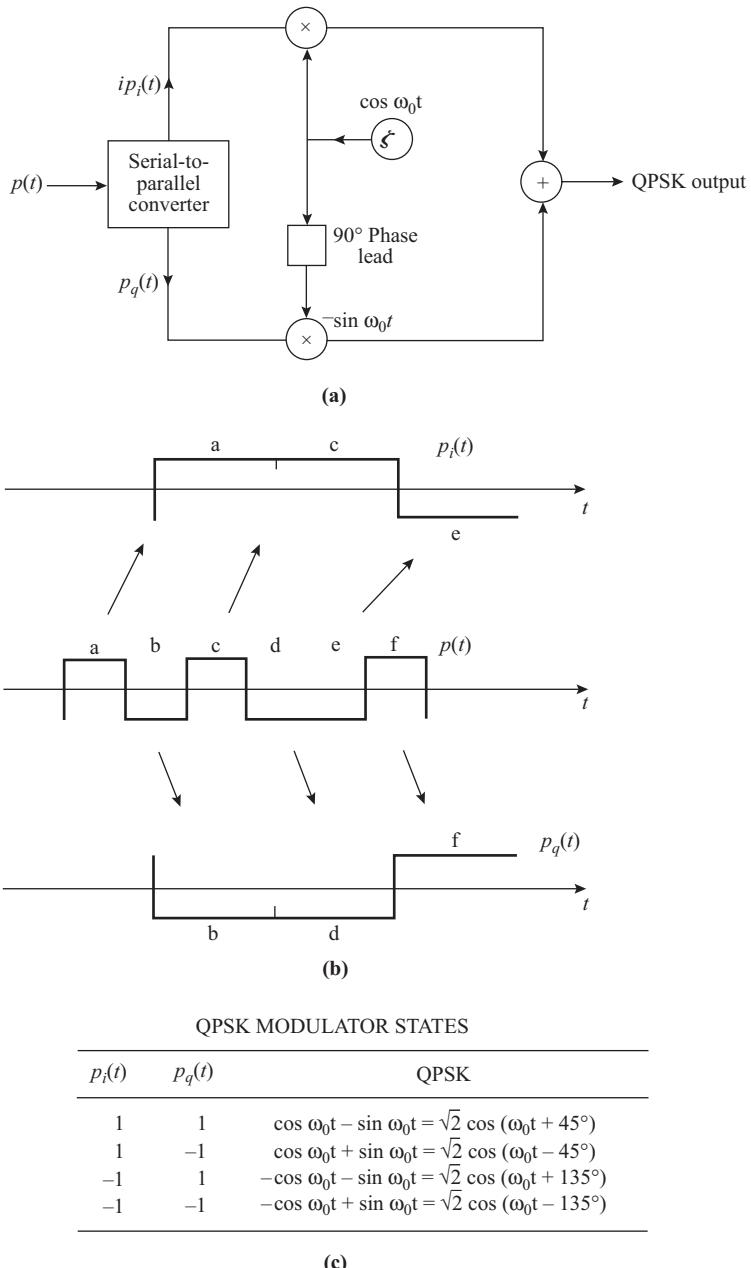


Figure 12.9.12 (a) QPSK modulator. (b) Waveforms for (a). (c) Modulator states.

Detection of QPSK is similar to that for BPSK with the difference that the recovered carriers must also have the 90° phase difference. A demodulator circuit is shown in Fig. 12.9.13. Assuming that the demodulated output is followed by a matched filter detector, the bit-error probability for QPSK is the same as that for BPSK as given by Eq. (12.9.21).

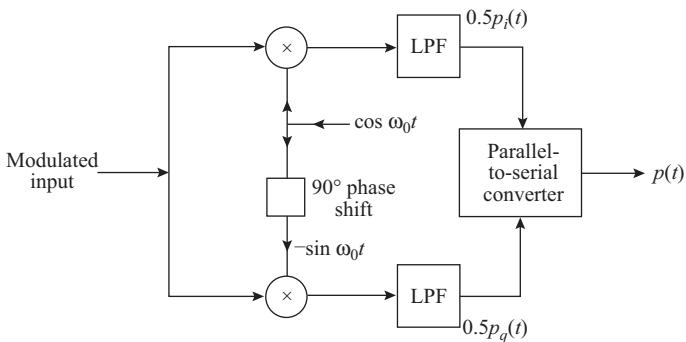


Figure 12.9.13 QPSK demodulator circuit.

TABLE 12.9.1

Bit Error Probability	Signal Type	BPS/Hz
$P_{be} = \frac{1}{2} \operatorname{erfc} \sqrt{\frac{E_b}{N_o}}$	Polar NRZ	2
	BPSK	1
	QPSK	2
$P_{be} = \frac{1}{2} \operatorname{erfc} \sqrt{\frac{E_b}{2N_o}}$	Unipolar NRZ	2
	OOK	1
	MSK	1

At this point it may be useful to summarize some of the results obtained for ideal systems, that is, where $\rho = 0$ for the raised-cosine spectrum, and the transmission channel is distortionless so that only a constant attenuation and constant delay are introduced. Under these conditions the comparisons shown in Table 12.9.1 may be made, assuming coherent demodulation for the modulated signals and optimum detection for the pulses.

It will be recalled that the average bit energy is obtained by multiplying the average transmitted power by the bit period, and that in the case of unipolar (including OOK) transmission, the voltage levels have to be increased by a factor of $\sqrt{2}$ compared to continuous wave transmissions.

12.10 Carrier Recovery Circuits

To implement coherent detection, a local oscillator must be provided at the receiver that is exactly synchronized to the carrier. In the case of the OOK signal, a carrier component is present in the spectrum of the received signal as shown in Fig. 12.9.4(b), which can be used to synchronize a local oscillator. In general, however, the spectrum will not have such a carrier component; for example the BPSK signal is a DSBSC-type signal that has no carrier component directly available. A pilot carrier could be transmitted, but this is an inefficient use of power, and the more common solution is to use nonlinear circuits to recover the carrier from the received signal. The two circuits that are most widely used are the *squaring loop* shown in Fig. 12.10.1(a) and the *Costa loop*, shown in Fig. 12.10.1(b).

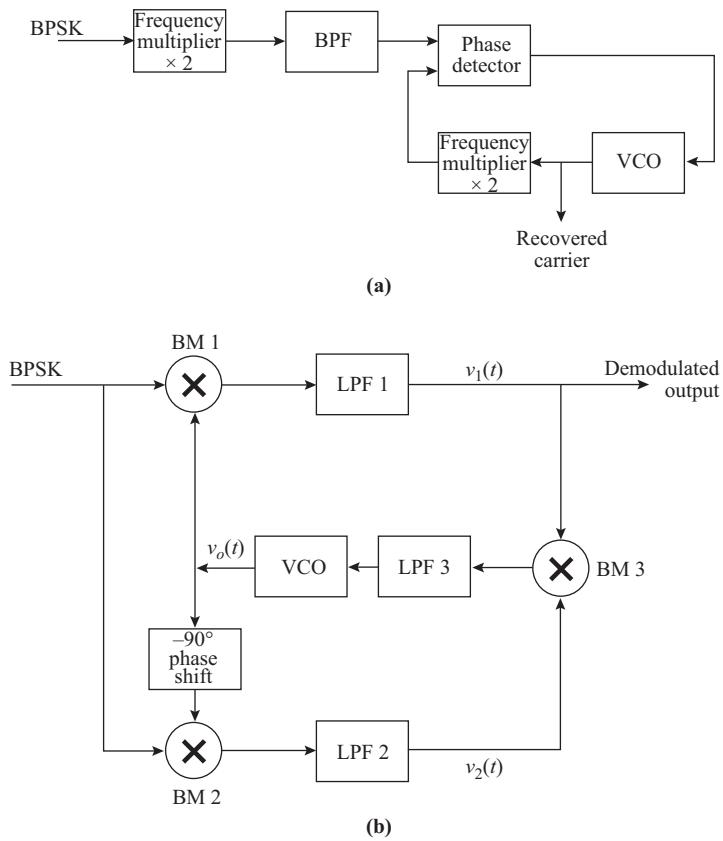


Figure 12.10.1 Carrier recovery circuits. (a) The squaring loop and (b) the Costa loop.

Consider first the operation of the squaring loop. Let the incoming signal be represented as $p(t)E_{c \max} \cos(2\pi f_c t + \phi_c)$, where the amplitude $p(t) = \pm 1$ depending on the polarity of the transmitted bit. The output from the squaring circuit is

$$E_{c \max}^2 \cos^2(2\pi f_c t + \phi_c) = \frac{E_{c \max}^2}{2}(1 + \cos 2(2\pi f_c t + \phi_c))$$

The band-pass filter is tuned to the second harmonic of the carrier, which eliminates the dc term. The phase locked loop, formed by the phase detector and the voltage-controlled oscillator (VCO), is locked onto the carrier frequency, with the phase detector functioning at the second harmonic frequencies. When properly adjusted, the output from the VCO, which provides the local oscillator signal, is synchronized in frequency and phase with the carrier of the incoming signal. A similar arrangement may be used with QPSK, except that frequency quadruplers rather than frequency doublers must be used. The use of frequency multipliers can be avoided by using an arrangement known as the Costa loop.

With the Costa loop, again assume the same input signal $p(t)E_{c \max} \cos(2\pi f_c t + \phi_c)$ and assume the VCO has acquired lock with the carrier frequency, but with some small phase error so that the VCO output is $v_o(t) = E_{o \max} \cos(2\pi f_c t + \phi_o)$. After multiplication and low-pass filtering in branch 1, the output from balanced modulator BM1 is proportional to

$$p(t)E_{c \max} \cos(2\pi f_c t + \phi_c) \times E_{o \max} \cos(2\pi f_c t + \phi_o)$$

Multiplying this out, and selecting the low-frequency component that is passed by low-pass filter 1 gives, for the branch 1 output,

$$v_1(t) = p(t) \frac{E_o \max E_c \max}{2} \cos \phi_e \quad (12.10.1)$$

where $\phi_e = \phi_c - \phi_o$. Similar reasoning for branch 2 results in the output from low-pass filter 2 being

$$v_2(t) = p(t) \frac{E_o \max E_c \max}{2} \sin \phi_e \quad (12.10.2)$$

These two signals are fed as inputs to balanced modulator 3, the output of which is proportional to

$$\begin{aligned} v_1(t)v_2(t) &= \frac{(p(t)E_o \max E_c \max)^2}{4} \cos \phi_e \sin \phi_e \\ &= \frac{(p(t)E_o \max E_c \max)^2}{8} \sin 2\phi_e \end{aligned} \quad (12.10.3)$$

This is essentially the dc control voltage for the VCO, and with $\sin \phi_e \approx \phi_e$ for small phase error, the control voltage is proportional to the phase error. Low-pass filter 3 eliminates any variations that might result from the $p(t)^2$ term and ensures that only the slowly varying dc component gets through as control voltage, which keeps the VCO locked onto the carrier frequency. Also, with small phase error, $\cos \phi_e \approx 1$ and the output from low-pass filter 1 is

$$\begin{aligned} v_1(t) &= p(t) \frac{E_o \max E_c \max}{2} \cos \phi_e \\ &\approx p(t) \frac{E_o \max E_c \max}{2} \\ &= K_P(t) \end{aligned} \quad (12.10.4)$$

where K is a constant. This shows that the output from branch 1 is the demodulated output.

One problem that arises with the carrier recovery schemes is that there exists the possibility of a 180° phase ambiguity in the recovered carrier. The analyses carried out for both loops apply even when the output from the VCO contains a 180° phase shift. Such a 180° phase shift results in the binary output being complemented; that is, **1**'s become **0**'s and **0**'s become **1**'s. To overcome this, a short binary test sequence may be transmitted at the beginning of each message, which allows the receiver to be adjusted to give the correct output. Another popular method is the use of differential encoding, as described in Section 3.4. Differential encoding may be applied directly to the modulation scheme as described in the next section.

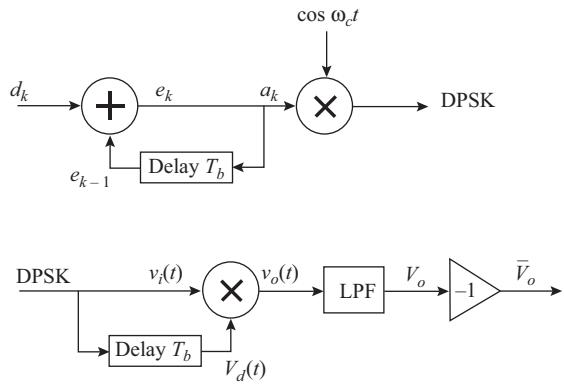
12.11 Differential Phase Shift Keying (DPSK)

Differential encoding of baseband signals is described in Section 3.4, where a binary message sequence d_k is shown differentially encoded as sequence e_k . For ease of comparison, this same sequence is used here and is shown in Fig. 12.11.1; the e_k sequence phase modulates a carrier $\cos \omega_c t$ as shown. The output from the modulator is a BPSK signal, but because it is modulated by a differentially encoded signal, it will be referred to as a differentially phase shift keyed or DPSK signal.

By assigning $a_k = +1$ V to a binary **1** and $a_k = -1$ V to a binary **0**, the DPSK signal can be represented by

$$\text{DPSK} = a_k \cos \omega_c t \quad (12.11.1)$$

It will be seen that a binary **1** corresponds to no phase shift and a binary **0** to a 180° phase shift in the carrier. At the receiving end, the decoder consists of a multiplier, the inputs to which are the DPSK signal



d_k (logic levels)	1	1	0	1	0	0	1
e_k (logic levels)	1	0	0	1	1	1	0
a_k (volts)	+1	-1	-1	+1	+1	+1	-1

a_k (volts)	+1	-1	-1	+1	+1	+1	-1
a_{k-1} (volts)	-1*	+1	-1	-1	+1	+1	+1
*Initial state							
V_o (volts)	-1	-1	+1	-1	+1	+1	-1
V_o (logic levels)	0	0	1	0	1	1	0
\bar{V}_o (logic levels)	1	1	0	1	0	0	1

Figure 12.11.1 Differential phase-shift keying (DPSK).

and a 1-bit delayed version of this. Ignoring the attenuation and delay produced in the transmission path (since these are common to both inputs), the output from the multiplier is

$$\begin{aligned} v_o(t) &= kv_l(t)v_d(t) \\ &= k(a_k \cos \omega_c t) \times (a_{k-1} \cos \omega_c(t - T_b)) \end{aligned} \quad (12.1.2)$$

where k is the multiplier constant. By making $f_c = N/T_b$, where N is an integer (and usually $N \gg 1$), $\cos \omega_c(t - T_b) \equiv \cos \omega_c t$ and

$$\begin{aligned} v_o(t) &= ka_k a_{k-1} \cos^2 \omega_c t \\ &= ka_k a_{k-1} \left(\frac{1}{2} + \frac{1}{2} \cos 2\omega_c t \right) \end{aligned} \quad (12.11.3)$$

The low-pass filter at the output of the multiplier removes the second harmonic term of the carrier, leaving only the baseband component. Denoting the constants by A , the low-pass output is

$$\begin{aligned} V_o &= A a_k a_{k-1} \\ &= +A, \quad \text{for } a_k = a_{k-1} \\ &= -A, \quad \text{for } a_k \neq a_{k-1} \end{aligned} \quad (12.11.4)$$

The logic levels corresponding to V_o are shown in the lower table of Fig. 12.11.1. These are the message levels inverted, and an inverter is required to restore the message sequence at the output. Comparison of the inverted output \bar{V}_o shown in Fig. 12.11.1 with the binary input d_k shows these to be the same. More generally, it can be shown that the multiplication followed by inversion is equivalent to the exclusive-OR decoder operation shown in Fig. 3.4.9 for the differentially encoded baseband signal.

A disadvantage with DPSK is that bit errors tend to occur in pairs, because the polarity of a given bit depends on the polarity of the preceding bit. The average bit-error probability for DPSK can be shown to be given by

$$P_{be} = \frac{1}{2} e^{-(E_b/N_o)} \quad (12.11.5)$$

12.12 Hard and Soft Decision Decoders

The process of decoding where the received bit is compared with a threshold level and a binary decision made is referred to as *hard-decision decoding*. In other words, the output is either right or wrong. Furthermore, where encoding for the purpose of error correction is employed, as described in the following section, the bit stream is partitioned into codewords. Again, with hard decision decoding, a firm right or wrong decision is made for each received codeword. An alternative to hard decision decoding is *soft decision decoding*. In soft decision decoding, a received codeword is compared with a (possibly large) number of alternatives, and the one that shows the closest, but not necessarily exact, correlation is assumed to be correct. Soft decision decoding results in about a 2-dB improvement in performance over hard decision decoding, meaning that the same bit-error probability can be achieved with a 2-dB reduction in the received signal-to-noise ratio. However, the cost and complexity of the additional equipment needed mean that it is only used where absolutely necessary. In the next section, only hard decision decoding will be considered.

12.13 Error Control Coding

The output from the binary source could be applied directly to the modulator shown in Fig. 12.1.1 and transmitted as a channel waveform. However, as explained in Section 12.4, noise introduces the probability of errors in the detected signal, this being given by P_{be} . One function of the *channel encoder* is to modify the binary stream in such a way that errors in the received signal can be detected and possibly corrected. Where an error can be corrected at a receiver without the need to request a retransmission of a message, it is referred to as *forward error correction* (FEC).

Figure 12.13.1 shows the notation that will be used to distinguish between coded and uncoded transmissions. The subscript U is used for the uncoded and the subscript C for the coded quantities. In Fig. 12.13.1(a), which applies to the uncoded system, the input bit rate to the channel is denoted by R_{bU} and the output bit-error probability as P_{beU} . For example, with polar NRZ transmission with matched filtering

$$P_{beU} = \frac{1}{2} \operatorname{erfc} \sqrt{\frac{E_b U}{N_0}} \quad (12.13.1)$$

In the case of the uncoded transmission, the output *bit-error rate*, denoted by BER_U , is equal to the bit-error probability. For example, if $P_{beU} = 0.0001$, the bit-error rate is 1 bit in every 1000 on average.

Figure 12.13.1(b) applies where error-control coding is used. The same uncoded bit rate is present at the input, but following the encoding stage a new bit rate, denoted by R_{bC} , is generated, which is transmitted

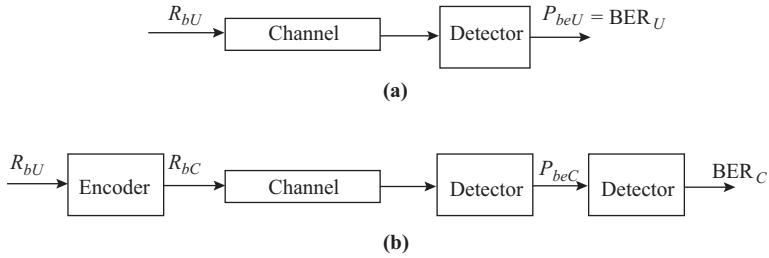


Figure 12.13.1 (a) Uncoded and (b) coded transmission system.

along the channel. At the receiver, the probability of bit error following detection is denoted by P_{beC} and the bit-error rate at the output by BER_C . If the error control coding is doing its job properly, the bit-error rate at the output should be less than the bit-error probability at the decoder input; that is, $\text{BER}_C < P_{beC}$. These ideas are examined in more detail in the following sections.

Block Codes

With block encoding, the bit stream is partitioned into binary words (or blocks of bits). It may be that the binary bit stream from the source is generated as a sequence of well-defined binary words, each representing a discrete value of input, which may be used as the partitioning, but in any case, partitioning is necessary. To distinguish the binary words before and after block encoding, the binary words from the source will be referred to as *datawords* and those from the channel encoder as *codewords*. The datawords each contain k bits and the codewords n bits, where $n > k$. The manner in which the extra $n - k$ bits per word are used to provide error control is described in the following sections. The code is referred to as a (n, k) block code, and the code rate is defined as

$$r = \frac{k}{n} \quad (12.13.2)$$

Although this is called a code rate, it is not a rate in bits per second, but rather a measure of dataword bits per codeword bit. Let the time required for the transmission of the complete message be T_{MU} for an uncoded transmission and T_{MC} for a coded transmission. Let the message consist of W datawords. The uncoded bit rate is therefore $R_{bU} = Wk/T_{MU}$, and the coded bit rate is $R_{bC} = Wn/T_{MC}$. It follows that

$$\begin{aligned} \frac{R_{bU}}{R_{bC}} &= \frac{k}{n} \frac{T_{MC}}{T_{MU}} \\ &= r \frac{T_{MC}}{T_{MU}} \end{aligned} \quad (12.13.3)$$

The relationship in general between average transmitted power and bit rate is

$$E_b = \frac{P_R}{R_b} \quad (12.13.4)$$

For the same average power transmitted for coded and uncoded signals carrying the same message, it follows that

$$\frac{E_{bC}}{E_{bU}} = \frac{R_{bU}}{R_{bC}} \quad (12.13.5)$$

and therefore

$$r \frac{T_{MC}}{T_{MU}} = \frac{E_{bC}}{E_{bU}} \quad (12.13.6)$$

Two simple but important conclusions can be drawn from this. With matched filtering, the bit-error probability depends on the bit energy, as shown in Section 12.5. If the bit energy and hence the probability of bit error are to remain the same for coded and uncoded signals, the transmission time must be increased by a factor $1/r$ (recall that $r < 1$). Alternatively, if the transmission times are to be the same, the bit energy for the coded signal must be reduced by a factor r compared to the uncoded signal. This means that the probability of bit error P_{beC} is increased (made worse). For example, the equation given in Section 12.5 for polar transmission would have to be modified to

$$\begin{aligned} P_{beC} &= \frac{1}{2} \operatorname{erfc} \sqrt{\frac{E_{bc}}{N_o}} \\ &= \frac{1}{2} \operatorname{erfc} \sqrt{\frac{rE_b}{N_o}} \end{aligned} \quad (12.13.7)$$

but, as mentioned previously, good coding ensures that $\text{BER}_C < P_{beC}$.

A general relationship can be developed for the probability of i errors occurring in an n -bit codeword. The probability of getting any one arrangement of exactly i errors, which also implies getting exactly $(n - i)$ correct bits, is

$$P_i = P_{beC}^i (1 - P_{beC})^{n-i} \quad (12.13.8)$$

But the number of combinations of n bits taken i at a time is

$$C_{i,n} = \frac{n!}{i!(n-i)!} \quad (12.13.9)$$

Each combination represents a codeword with i errors, and hence the probability of receiving any codeword with i errors is

$$P_{i,n} = C_{i,n} P_i \quad (12.13.10)$$

Normally, $P_{beC} \ll 1$ and, as a result, $P_i \approx P_{beC}^i$. The probability of receiving a codeword with i errors can therefore be simplified to

$$P_{i,n} \approx \frac{n!}{i!(n-i)!} P_{beC}^i \quad (12.13.11)$$

The extent of this simplification is illustrated in the following example.

EXAMPLE 12.13.1

A digital signal utilizes 8-bit codewords. Calculate the probability of a received codeword containing three errors, given that the bit-error probability in transmission is 0.01.

SOLUTION Given data:

$$P_{beC} := .01, \quad n := 8, \quad i := 3$$

Computations:

$$P_i := P_{beC}^i [1 - P_{beC}]^{n-i} \quad \text{Eq. (12.13.8)}$$

$$C_{in} := \frac{n!}{i!(n-i)!} \quad \text{Eq. (12.13.9)}$$

$$P_{in} := C_{in} P_i \quad \text{Eq. (12.13.10)}$$

$$P'_{in} := C_{in} P_{beC}^i \quad \text{Eq. (12.13.11) (approximate)}$$

$$P_{in} := 5.326 \times 10^{-5}, \quad P'_{in} := 5.6 \times 10^{-5}$$

Often it is necessary to know the probability of a codeword containing *at least i* errors. This will be termed the probability of codeword error and will be denoted by P_{we} . Thus

$$P_{we} = P_{i,n} + P_{i+1,n} + P_{i+2,n} + \dots + P_{n,n} \quad (12.13.12)$$

With $P_{beC} \ll 1$, which is usually the case, the probability of getting *more than i* errors is negligible compared to the probability of getting *i* errors. It is then permissible to approximate

$$P_{we} \approx P_{i,n} \quad (12.13.13)$$

This is illustrated in the following example.

EXAMPLE 12.13.2

A digital signal utilizes 8-bit codewords. Calculate the probability of a received codeword containing (a) exactly three errors and (b) at least three errors. The bit-error probability in transmission is 0.01.

SOLUTION Given data:

$$P_{beC} := .01 \quad n := 8, \quad i := 3 \dots n$$

Computations: For purposes of comparison, the exact equations leading to Eq.(12.13.12) will be used, followed by the approximate approach given by Eq.(12.13.13).

$$P_{i,n} := P_{beC}^i (1 - P_{beC})^{n-i} \quad \text{Eq. (12.13.8)}$$

$$C_{i,n} := \frac{n!}{i!(n-i)!} \quad \text{Eq. (12.13.9)}$$

$$P_{i,n} := C_{i,n} P_{i,n} \quad \text{Eq. (12.13.10)}$$

$$P_{we} := \sum_i P_{i,n} \quad \text{Eq. (12.13.12)}$$

$$(a) P_{we} = 5.393 \times 10^{-5}$$

$$(b) P_{3,n} = 5.326 \times 10^{-5} \quad \text{Eq. (12.13.13)}$$

The function of error control may be broadly classified into *error detection* and *error correction*. An error detection system is designed to detect when errors occur, but has no means of locating which bits or symbols are in error. A repeat transmission is then necessary if the errors are to be removed. The request for repeat transmission is usually made automatically, this being known as *automatic repeat request* (ARQ). A positive acknowledgment signal, ACK, is returned for correctly received codewords, and a negative acknowledgment,

or NAK, for codewords in error. These signals are returned over a low-bit-rate (and hence narrow bandwidth) channel in which the probability of bit errors is negligibly small.

An error correction system can detect and correct errors and is referred to as *forward error correction* (FEC) encoding. It is also possible for an encoding system to provide error correction up to some limit and additional error detection for errors it cannot correct. Some of these ideas will be illustrated with reference to specific encoding schemes in the following sections.

Repetition Encoding

Suppose, for example, that the datawords consist of single **1**'s or **0**'s and that *triple repetition* is used to encode these so that the corresponding codewords are **111** and **000**. At the receiver, it is assumed that synchronization enables the individual codewords to be recognized. For *error detection*, any received word other than a **111** or **000** will be interpreted as an error. The probability of an error going undetected can be calculated knowing the probability of bit-error P_{beC} . An undetected codeword occurs only when all three bits are in error. The probability of this occurring is simply

$$P_{we} = P_{beC}^3 \quad (12.13.14)$$

Each time a word error occurs, a wrong output bit is generated, and hence the output bit-error rate is

$$\text{BER}_C = P_{beC}^3 \quad (12.13.15)$$

As an example, if $P_{beC} = 0.01$, then $\text{BER}_C = 10^{-6}$. For the uncoded condition described previously in which the transmission time is decreased by a factor r , the bit energy remains unchanged and the probability of bit error remains at 0.01, so that $\text{BER}_U = P_{beU} = 0.01$. This shows that a significant improvement in the average bit error can be achieved, but at the expense of increasing the transmission time.

With error correction, a decision is made at the receiver based on a simple “majority vote.” Suppose, for example, the codeword **111** is transmitted. A single bit error in the received codeword would change it to one of **110**, **101**, or **011**. A simple majority decision would allow any of these to be interpreted as a **1**, and so in this case no error would occur in the detection. If, however, 2 bits in the codeword were in error, the received codeword would be one of **001**, **010**, or **100**, which simple majority voting would interpret as a **0**. Three errors would result in **000**, which would also be interpreted as a **0**. In these situations the simple majority vote results in an error. Thus with triple repetition encoding, the word-error probability when error correction is employed is given by Eq. (12.13.12) as

$$\begin{aligned} P_{we} &= P_{2,3} + P_{3,3} \\ &\cong P_{2,3} \end{aligned} \quad (12.13.16)$$

In this situation also, each time a word error occurs a wrong bit is generated, and therefore

$$\begin{aligned} \text{BER}_C &= P_{we} \\ &\cong P_{2,3} \end{aligned} \quad (12.13.17)$$

Again, with $P_{beC} = 0.01$, evaluation of the probability equations (which is left as an exercise for the student) gives $\text{BER}_C \cong 0.0003$ compared to $\text{BER}_U = 0.01$. To avoid the possibility of a tie in the simple majority vote, the number of repetitions n is chosen to be an odd number. With triple repetition encoding the code rate is seen to be $r = 1/3$.

Parity Encoding

Parity is a method of encoding such that the number of **1**'s in a codeword is either *even* or *odd*. Single parity is established as follows. Each dataword is examined to determine whether it contains an odd or even number

of 1 bits. If even parity is to be established, a 1 bit is added to each odd dataword, and a 0 bit is added to each even dataword. The result is that all the codewords contain an even number of 1 bits after parity is added. It follows that $k - n = 1$ and the code rate is $r = k/n = (n - 1)/n$, where n is the number of bits in a codeword. Single parity is capable of detecting single errors only and it cannot provide error correction.

Continuing with the example of even parity, after transmission, each codeword is examined to see if it contains an even number of 1 bits. If it does not, the presence of an error is indicated. If it does, the parity bit is removed and the data passed to the user. This form of parity will detect errors only if an odd number of bits is disturbed. An even number of errors within the same codeword will be self-compensating and go undetected. The probability of a codeword containing an undetected error is equal to the probability of a codeword containing an even number of errors, which from Eq. (12.13.13) is

$$\begin{aligned} P_{we} &\cong P_{2,n} \\ &= \frac{n!}{2!(n-2)!} P_{beC}^2 \\ &= \frac{n(n-1)}{2} P_{beC}^2 \end{aligned} \quad (12.13.18)$$

For a long message of W words, where $W \gg 1$, the number of words in error that go undetected is $W_E = WP_{we}$. Since each word error contains 2 bits in error, the number of undetected bit errors in the message is $2WP_{we}$. Each codeword contains a total of k data bits. Assuming a uniform distribution of errors among the n codeword bits, the fraction of errors expected in data bits is $r = k/n$. Thus the number of data bits in error in the message, on average, is $2rWP_{we}$. But the total number of data bits in the message is kW , and hence the bit-error rate, based on the relative frequency definition of probability, is

$$\begin{aligned} \text{BER}_C &= \frac{2rWP_{we}}{kW} \\ &= \frac{2}{n} \frac{n(n-1)}{2} P_{beC}^2 \\ &= (n-1)P_{beC}^2 \end{aligned} \quad (12.13.19)$$

As before, without coding the bit-error rate is $\text{BER}_U = P_{beU}$. The improvement in bit-error rate with single parity is illustrated in the following example.

EXAMPLE 12.13.3

For a binary transmission link, the uncoded E_b/N_o ratio is 9 dB, and matched filtering with polar transmission is used. (a) Calculate the bit-error rate. (b) Calculate the new bit-error rate obtainable with single parity coding and $n = 10$, given that the transmit power and message time remain unchanged.

SOLUTION

$$\text{SN dB} := 9$$

$$\text{SNR} := 10^{\left(\frac{\text{SN dB}}{10}\right)}$$

- (a) $P_{beU} := 0.5 \left(1 - \text{erf} \sqrt{\text{SNR}}\right)$
 $\text{BER}_U := P_{beU}$
 $\text{BER}_U = 3.363 \times 10^{-5}$

$$(b) \quad n := 10, \quad k := n - 1, \quad r := \frac{k}{n}$$

$$\text{SNR} := r \times \text{SNR}$$

$$P_{beC} := 0.5 \left(1 - \operatorname{erf} \sqrt{\text{SNR}} \right)$$

$$\text{BER}_C := (n - 1) P_{beC}^2,$$

Eq. (12.13.19)

$$\text{BER} = 5.478 \times 10^{-8}$$

A code with natural built-in parity was developed in the 1940s for radio telegraphy (Fig. 12.13.2), which also incorporated ARQ facilities. Each data-word is encoded into a 7-bit codeword that contains exactly three **1**'s distributed throughout the codeword. A codeword with more or less than three **1**'s will be known to contain an error, and a repeat transmission can be requested.

If errors should occur in the same word in such a manner as to keep the number of **1**'s at three, these errors will go undetected. On the assumption that only two errors at most per word need to be considered,

Character		Bit Position					
		1	2	3	4	5	6
1	A	—			0011010		
2	B	?			0011001		
3	C	:			0001100		
4	D	WRU			0011100		
5	E	3			0111000		
6	F	%			0010011		
7	G	(a)			1100001		
8	H	£			1010010		
9	I	8			1110000		
10	J	BELL			0100011		
11	K	(0001011		
12	L)			1100010		
13	M	.			1010001		
14	N	,			1010100		
15	O	9			1000110		
16	P	Ø			1001010		
17	Q	1			0001101		
18	R	4			1100100		
19	S	,			0101010		
20	T	5			1000101		
21	U	7			0110010		
22	V	=			1001001		
23	W	2			0100101		
24	X	/			0010110		
25	Y	6			0010101		
26	Z	+			0110001		
27	CARR. RETURN				1000011		
28	LINE FEED				1011000		
29	FIGS. SHIFT				0100110		
30	LETT. SHIFT				0001110		
31	SPACE				1101000		
32	BLANK				0000111		

Figure 12.13.2 ARQ teleprinter code. (Courtesy Howard W. Sams and Company, Inc.)

		Column		
		1	2	3
Row	1	b_1	b_2	b_3
	2	b_4	b_5	b_6
	3	b_7	b_8	b_9
		c_4	c_5	c_6

Figure 12.13.3 Error correction by means of parity.

the word error probability is given by the product of the probability that one error will occur in the three **1's** at the same time as one error occurs in the four **0's**. This is

$$\begin{aligned} P_{we} &= P_{1,3} P_{1,4} \\ &= 12P_{beC}^2 \end{aligned} \quad (12.13.20)$$

For a message of W words, where $W \geq 1$, the number of words in error that go undetected is $W_E = WP_{we}$. Since each word error contains 2 bits in error, the number of undetected bit errors in the message is $2WP_{we}$. The total number of bits transmitted is $7W$, and hence the bit-error rate is

$$\begin{aligned} \text{BER}_C &= \frac{2WP_{beC}}{7W} \\ &= \frac{24}{7} P_{beC}^2 \end{aligned} \quad (12.13.21)$$

The code rate for the teleprinters ARQ code is 35/128. This is arrived at by noting that a 7-bit word is capable of transmitting 128 different words, and by restricting the words to those containing exactly three **1's**, the number of valid words is $7!/(3!4!) = 35$.

Simple parity can detect, but not correct, errors. By using *multiple parity*, certain classes of errors can be corrected. Several datawords are combined to form a block of information. Within a block, each dataword forms a row and has simple parity added. In addition, each column in the block may have parity established for it by adding a parity word to the end of the block. This is illustrated in Fig. 12.13.3, where the parity or *check* bits are denoted by c and the data bits by b .

If a single error occurs within a block, a parity error will be indicated for the word in which the error occurs and also for the column in which it occurs. The intersection of the row and column is the location of the faulted bit, and *correction* can be achieved merely by inverting that bit. For example, if row 2 and column 3 show parity violation, then message bit 6 is in error. The system will also detect multiple errors, but if these occur in more than one word or in more than one row of the block, correction cannot be accomplished. The coordinates for two such errors will intersect in four locations, and with the simple form of two-dimensional parity described, it is not possible to distinguish which of these contain the errors.

Bit-error Probability with Forward Error Correction

The efficiency of error control coding is measured by the reduction in the bit-error probability achieved. It must be understood that the error control coding does not alter the probability of bit error in transmission,

which is determined by the signal-to-noise ratio (these are the P_{beC} and P_{beU}); but what is altered is the output bit-error rate, denoted above by BER_C when coding is employed. Certain classes of code used for forward error correction (FEC) can correct all errors up to some maximum number, denoted here by t . The probability of receiving a codeword with more than this number of errors is given by Eq. (12.13.13).

$$P_{we} \equiv P_{i,n} \quad (12.13.22)$$

where $i = t + 1$. For a long message of W words, where $W \gg 1$, the number of words in error that go undetected is $W_E = WP_{we}$. Since each word error contains i bits in error, the number of undetected bit errors in the message is iWP_{we} . Each codeword contains a total of k data bits. Assuming a uniform distribution of errors among the n codeword bits, the fraction of errors expected in data bits is $r = k/n$. Thus the number of data bits in error in the message, on average, is $irWP_{we}$. But the total number of data bits in the message is kW , and hence the bit-error rate, based on the relative frequency definition of probability, is

$$\begin{aligned} BER_C &= \frac{irWP_{we}}{kW} \\ &= \frac{i}{n} \left(\frac{n!}{i!(n-i)!} P_{beC}^i \right) \\ &= \frac{(n-1)!}{t!(n-1-t)!} P_{beC}^{t+1} \end{aligned} \quad (12.13.23)$$

As before, without coding the bit-error rate is $BER_U = P_{beU}$. The improvement in bit-error rate with FEC is illustrated in the following example.

EXAMPLE 12.13.4

For a binary transmission link the uncoded E_b/N_o ratio is 8 dB, and matched filtering with polar transmission is used. (a) Calculate the bit-error rate, (b) Calculate the new bit-error rate obtainable with block encoding for which $n = 15$, $k = 11$, and $t = 1$, given that the transmit power and message time remain unchanged.

SOLUTION

$$SN \text{ dB} := 8$$

$$SNR := 10^{\left(\frac{SN \text{ dB}}{10}\right)}$$

$$\begin{aligned} (a) \quad P_{beU} &:= 0.5 \left(1 - \operatorname{erf} \sqrt{SNR} \right) \\ BER_U &:= P_{beU} \\ BER_U &= 1.909 \times 10^{-4} \end{aligned}$$

$$(b) \quad n := 15, \quad k := 11, \quad t := 1, \quad r := \frac{k}{n}$$

$$SNR := r \times SNR$$

$$P_{beC} := 0.5 \left(1 - \operatorname{erf} \sqrt{SNR} \right)$$

$$BER_C := \frac{(n-1)!}{t! \times (n-1-t)!} P_{beC}^{t+1} \quad \text{Eq. (12.13.19)}$$

$$BER_C = 1.932 \times 10^{-5}$$

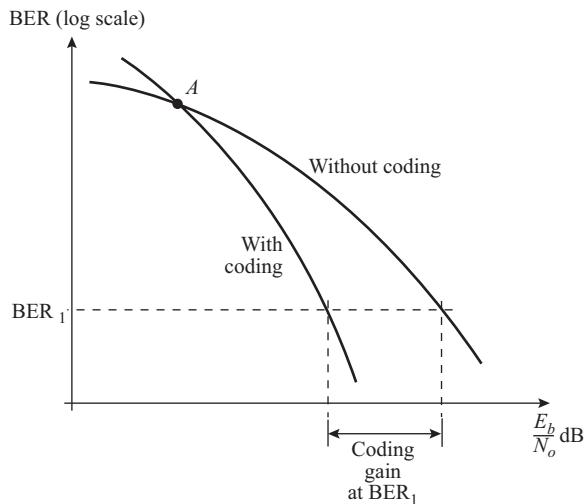


Figure 12.13.4 The meaning of coding gain.

Coding Gain

Plotting the uncoded and coded bit-error rates as functions of the *uncoded* value of E_b/N_o in decibels results in curves similar to those shown in Fig. 12.13.4. The *coding gain* is the difference, in decibels between the E_b/N_o values obtained for a given bit-error rate, for example BER_1 , shown in Fig. 12.13.4. It will be noticed that the curves cross over at point A. For E_b/N_o ratios less than the value at A, the situation is made worse by coding, this resulting from the reduced bit energy as described previously. A numerical example is given in Problem 12.44.

Some Block Codes

Many classes of block codes exist, some of which are designed for special purposes. These codes are named after their discoverers, and the main parameters of some of these are summarized in Table 12.13.1.

Another group of block codes known as *cyclical redundancy codes* (CRC) is frequently used for error checking. These codes rely on the fact that Modulo 2 multiplication and Modulo 2 division (done without carries) yield identical results. The M-bit long data block to be encoded is EXORed with the output of an R-bit long register (and simultaneously transmitted). The output of the EXOR gate becomes a feedback bit that is connected to several EXOR gates placed at predetermined positions within the register chain, which effectively Modulo 2 divides the data block by a predetermined R-bit long check function number. The R-bit long dividend number is then shifted out of the register and transmitted as the added block check number.

TABLE 12.13.1 Some Block Codes

Code Name	n	k	t	Notes
Hamming	$2^m - 1$	$n - m$	1	m an integer
Golay	23	12	3	
Bose–Chaudhuri–Hocquenghem (BCH)	$2^m - 1$	$\geq n - mt$	t	Multiple error correction
Reed–Solomon (RS)	$2^m - 1$	$n - 2t$	t	$k, n, m,$ and t symbols: designed to cope with burst errors

At the receive end, the multiplication process is repeated. If no errors occur during transmission, the contents in the receiving register at the end will be zero (or the initial binary number in the sending register before entering the block data). If not, then ARQ can be invoked to request a repeat transmission. The check bits are discarded after a successful receipt.

Convolution Encoding

A second form of encoding known as *convolution encoding* works on a continuous stream of data. The source data appear as a continuous stream of bits at a given bit rate (bps), which is passed through a shift register. As the bits are temporarily stored in the shift register, they are combined in a known manner using modulo 2 adders to form the encoded output. A convolution encoder utilizing a three-stage shift register is shown in Fig. 12.13.5.

Data are shifted in k data bits at a time, and the combinatorial logic generates an output of n encoded bits at each shift operation. The combinatorial circuits will be described shortly, but for the moment, only the *states* of the encoder will be examined. One way of showing the states of the encoder is by means of a *tree diagram*, as shown in Fig. 12.13.6. This particular diagram is for the encoder taking in 1 bit ($k = 1$) at a time.

The upward arrows on the branches indicate the change in the shift register when the incoming bit is a **0**, and downward arrows when it is a **1**. It must be kept in mind that as a new bit enters location $S1$ in the shift register, the stored bits all move one place to the right, so that the old $S3$ bit is completely displaced and drops out. In its initial (or cleared) state, the shift register is assumed to store **000**. The encoder can be in one of four different states labeled a , b , c , and d on the tree diagram. In state a the branches lead to the shift register storing either **000** or **100**; in state b , either **010** or **110**; in state c , either **001** or **101**; and in state d , either **011** or **111**. It will be seen that these states repeat, and a more compact form of diagram, known as a *trellis diagram*, is normally used. The development of the trellis diagram is shown in Fig 12.13.7.

The encoder states are shown as rows of dots on the diagram. From the tree diagram it is seen that state a can lead to state a or state b , and state b can lead to state c or state d . These conditions are shown in Fig. 12.13.7(a) by the two branches leaving each dot, the uppermost branch in each case showing the change when a **0** enters the shift register, and the lowermost branch the change when a **1** enters. Also from the tree diagram it is seen that state c leads to state a or state b , and state d to state c or state d . These conditions are shown in Fig. 12.13.7(b). These two diagrams can be superimposed to give the trellis diagram of Fig. 12.13.7(c).

To be of more use, the trellis should also show the encoded output as the input changes, but this depends on the combinatorial circuits used in Fig. 12.13.5. As an example, consider the encoder of Fig. 12.13.8(a). Here the $S1$ and $S3$ bits are Modulo 2 added to form the $U1$ output, and the $S1$, $S2$, and $S3$ bits are similarly added to form the $U2$ output. These are converted to a serial output through the switch, which is synchronized to the clocking signal for the shift register, $U1$ and $U2$ must be read out during one input bit period, when the shift register is in a steady-state condition. This means that if the input bit rate is r bps, the output bit rate is $2r$ bps. The $U1$, $U2$ output bits for the encoder states are shown in Fig. 12.13.8(b), and the corresponding branches of the trellis are labeled with these, as shown in Fig. 12.13.8(c).

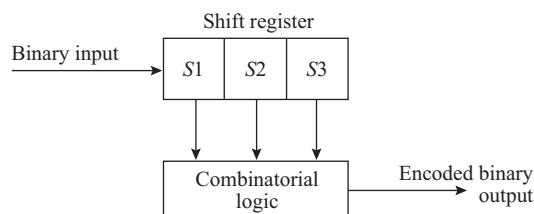


Figure 12.13.5 Convolution encoder.

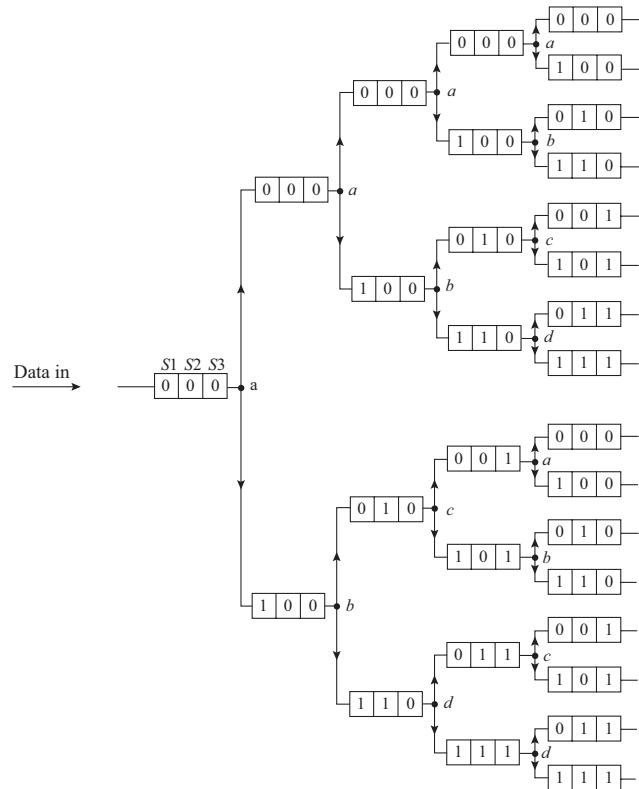


Figure 12.13.6 Tree diagram for the convolution encoder of Fig. 12.13.5.

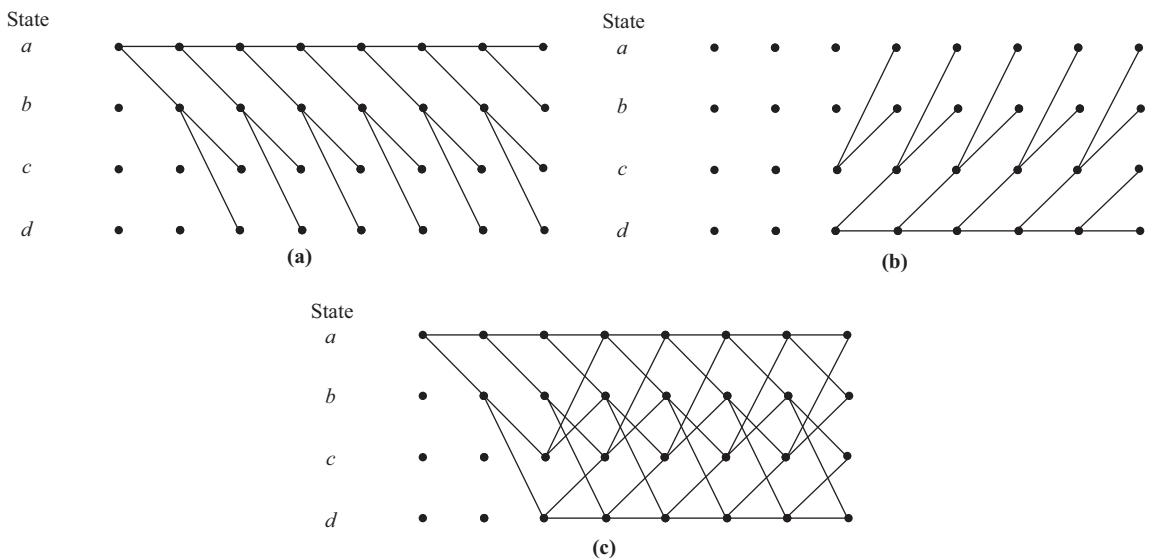


Figure 12.13.7 Trellis diagram corresponding to the tree diagram of Fig. 12.13.6.

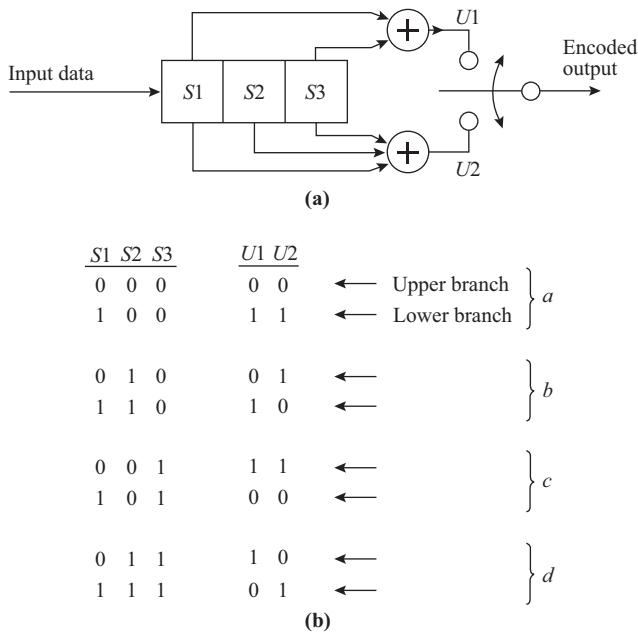


Figure 12.13.8 (a) Example of a convolution encoder. (b) Output bits corresponding to the different encoder states. (c) Trellis diagram labeled with the possible outputs.

The principle of error correction is best illustrated by means of an example. Consider the input sequence **10011** shown in Fig. 12.13.9(a). The trellis path for this is shown in Fig. 12.13.9(b), from which the encoded output sequence **11|01|11|11|10** is obtained [Fig. 12.13.9(c)]. Now suppose that at the receiver an error occurs in one of the bits in the third group of two so that the received sequence is **11|01|10|11|10**, as shown in Fig. 12.13.9(d). The trellis path for the received sequence is shown in the second trellis diagram [Fig. 12.13.9(e)]. In fact, a number of paths, labeled *p*, *q*, *r*, and *s*, are likely. Keeping in mind that the branch labels for the trellis are known at the receiver (since these are fixed by the particular encoder used), it is possible to compare the received branch sequences with those expected for a particular branch. This is done in Fig. 12.13.9(e), where solid lines are used to show where the received sequences agree with the branch labels, and dashed lines show where there is a conflict. Although the paths are described as being “likely,” they are not equally likely. Paths *p* and *q* each

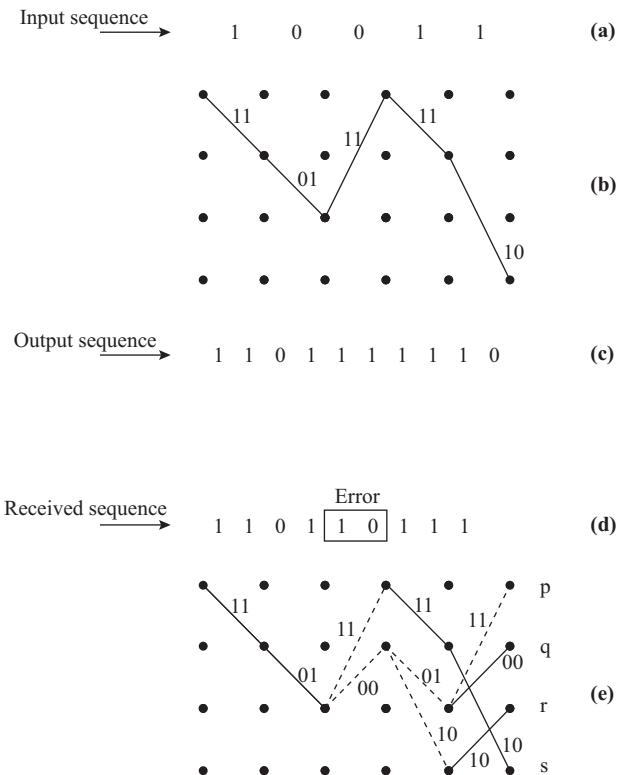


Figure 12.13.9 (a) Input sequence, (b) trellis path, (c) output sequence, (d) received sequence containing an error, and (e) possible trellis paths at the receiver, with s being the maximum-likelihood path.

have three conflicts, path r has two conflicts, and path t has one conflict. Path s therefore would be chosen as being the most likely, this being referred to as a *maximum likelihood* decision. This is seen to be the correct decision, because the path agrees with the correct path shown in Fig. 12.13.9(b). A number of computational strategies are available for implementing the maximum likelihood decoding. These are too complex to be gone into here, but details will be found in most advanced books on digital transmission.

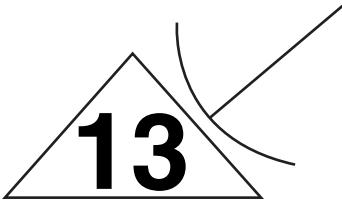
PROBLEMS

- 12.1. Briefly describe the functional blocks in a digital communications system.
- 12.2. Explain what is meant by *synchronization* in connection with digital transmission.
- 12.3. Using a rectangular pulse of amplitude A to represent a binary 1 and the zero level for a binary 0, draw the waveform for the message *you are* sent in CCITT-2 teleprinter code.
- 12.4. Repeat Problem 12.3 for the ASCII code.
- 12.5. For a signal amplitude of 0.3 V and an rms noise voltage of 90 mV, calculate the probability of bit error for (a) unipolar and (b) polar digital transmission.

- 12.6.** Given that the rms noise voltage in a digital system is 90 mV, on a common set of axes plot the probability of bit error for unipolar and polar transmission for a signal voltage amplitude in the range from 0.1 to 0.3 V.
- 12.7.** A polar digital signal consists of rectangular pulses of amplitude 500 mV. Calculate (a) the rms signal voltage, (b) the S/N ratio, and (c) the probability of bit error, given that the rms noise voltage is 120 mV.
- 12.8.** Repeat Problem 12.7 for a unipolar signal.
- 12.9.** For a digital transmission, the signal-to-noise ratio is 9 dB. Calculate the probability of bit error for (a) unipolar, (b) polar, and (c) AMI transmissions.
- 12.10.** Explain what is meant by a *matched filter*.
- 12.11.** On a common set of axes, plot the probability of bit error for unipolar and polar transmissions utilizing matched filter detection for E_b/N_o ratios in the range from 1 to 11 dB.
- 12.12.** For a polar transmission, the bit rate is 1.544 Mbps, the received power is 1 pW, and the noise figure of the receiver is 12 dB. Calculate the probability of bit error for matched filter detection.
- 12.13.** An integrate-and-dump detector is used for the detection of unipolar pulses in noise. The pulse period is 2 ms and the integrator time constant is 30 ms. The bit energy is 50 pJ and the source feeding the detector has an internal resistance of 50 Ω . Calculate (a) the output voltage and (b) the noise bandwidth.
- 12.14.** For the integrate-and-dump detector of Problem 12.13, plot the magnitude squared of the transfer function over a frequency range from 0 to 1000 Hz. Assuming that the curve can be approximated by a triangular shape to the first null, estimate the noise bandwidth using Eq. (4.2.12). How does this compare with the value obtained in Problem 12.13?
- 12.15.** Explain how bit timing may be recovered from a digital signal by (a) use of nonlinear circuits and (b) through the use of the early-late gate. Why is the latter so called?
- 12.16.** By tracing the lettered points on the waveform of Fig. 12.8.1, complete the lettering on the eye diagram in the figure.
- 12.17.** The binary stream shown in Fig 12.8.1(a) is encoded as an AMI waveform. Draw the waveform and the corresponding eye diagram.
- 12.18.** A sinusoidal carrier is modulated by a digital signal. Calculate the average received power and hence the average bit energy for (a) OOK, (b) BPSK, and (c) BFSK. The modulating pulses may be assumed rectangular, and the peak carrier voltage is 3 V. A 1- Ω load resistance may be assumed, and the bit rate is 25 Mbps.
- 12.19.** A sinusoidal carrier has a peak value of 3 V and frequency of 100 MHz. It is modulated with a binary digital message **101**, the bit rate being 25 Mbps. Draw accurately to scale the modulated waveform for (a) OOK and (b) BPSK.
- 12.20.** A binary digital signal can be represented as a periodic square wave at a bit rate of 1.5 Mbps. It is used to produce an OOK wave, the carrier amplitude being 2 V and the frequency 10 MHz. Sketch the spectrum of the OOK wave up to the 9th side frequency. Compare the bandwidth with that required using raised-cosine filtering with $\rho = 0.25$.
- 12.21.** Explain the difference between coherent and noncoherent detection of OOK signals. An OOK signal produces an $E_b/N_o = 12$ dB at the receiver. Calculate the probability of bit error for (a) coherent and (b) optimum noncoherent detection.
- 12.22.** The average power in a received signal is 3 pW. The noise bandwidth of the receiver is 100 kHz, and the noise figure is 15 dB. Given that the bit rate of the digital modulation is 48 kbps, calculate (a) the

- noise power at the receiver, (b) the signal-to-noise ratio in decibels, and (c) the bit energy-to-noise density ratio in decibels.
- 12.23. The signal-to-noise ratio for a digital transmission system is 14 dB, and the bps/Hz figure is 0.8. Calculate the probability of bit error given that the modulation is OOK.
- 12.24. Explain what is meant by (a) *continuous frequency shift keying* (CFSK) and (b) *minimum shift keying* (MSK).
- 12.25. The received bit energy-to-noise density ratio for an MSK signal is 13 dB. Calculate the probability of bit error for (a) coherent detection and (b) noncoherent detection.
- 12.26. Explain what is meant by BPSK. The rms voltage of a received carrier that is modulated by a binary digital signal is $4 \mu\text{V}$. The digital bit rate is 32 Mbps. The noise figure of the receiver is 10 dB. Calculate the probability of bit error for (a) BPSK and (b) OOK modulation, given that coherent detection is used in both cases.
- 12.27. It is desired to operate a BPSK digital system with a probability of error no greater than 10^{-6} . Calculate the minimum E_b/N_o ratio required in decibels.
- 12.28. Explain what is meant by QPSK. Derive the trigonometric formulas showing the QPSK phase angles in the table in Fig. 12.9.11(c).
- 12.29. Explain what is meant by *carrier recovery*. Show how the carrier may be recovered from a QPSK signal by the use of a frequency quadrupler circuit.
- 12.30. Explain the operation of the Costa loop for carrier recovery. How does this method compare with the frequency multiplier method?
- 12.31. Explain what is meant by *phase ambiguity* and how this may be corrected.
- 12.32. Draw up a table similar to that in Fig. 12.11.1, showing how the sequence $d_k = \mathbf{1011001}$ would be encoded and decoded using DPSK.
- 12.33. Explain what is meant by *code rate* in connection with error control coding. The input bit rate to an error control encoder is 1.544 Mbps and the code rate is $\frac{7}{8}$. What is the output bit rate?
- 12.34. The probability of bit error for uncoded transmission is 10^{-4} , with polar transmission being used. Calculate the new probability of bit error if error control coding is employed with a code rate of 11/15.
- 12.35. Assuming that the probability of bit error with coded transmission is to be the same as for uncoded in Problem 12.34, calculate the factor by which the transmission time is increased.
- 12.36. The probability of bit error in a certain digital system is 0.001. Given that a codeword contains 8 bits, calculate (a) using the exact formula and (b) the approximate formula the probability of a codeword containing two errors.
- 12.37. For the values given in Problem 12.36, calculate the probability of getting less than three errors.
- 12.38. For the values given in Problem 12.36, calculate the probability of getting three or more errors.
- 12.39. The sequence **111 101 001 000** is received using triple redundancy encoding. What is the most likely message sequence?
- 12.40. For a binary polar transmission, the uncoded E_b/N_o ratio is 8.5 dB. Calculate the bit-error rate. Calculate the new E_b/N_o ratio when single parity encoding is used with $n = 10$, all other factors remaining the same. Calculate also the bit-error rate achieved with the single parity encoding.
- 12.41. The probability of bit error in transmission for an ARQ signal is 0.0001. What is the bit-error rate?
- 12.42. Explain what is meant by *forward error correction*. An FEC code can correct any number of errors up to three, the codeword length being 23 bits. Calculate the probability of a word error, given that the probability of bit error in transmission is 1 in 1000.

- 12.43.** Calculate the bit-error rate achieved after decoding for the system in Problem 12.42.
- 12.44.** Explain what is meant by *coding gain*. Using the data given in Example 12.13.4, plot the bit-error rates for coded and uncoded transmissions for $4 \text{ dB} \leq E_b/N_o \geq 12 \text{ dB}$ and determine the coding gain for a bit-error rate of 10^{-6} .
- 12.45.** Plot the bit error probability as a function of E_b/N_o for unipolar and polar waveforms using MATLAB.
- 12.46.** Generate the *eye diagram* on transmitting a random binary waveform using MATLAB.
- 12.47.** Generate the *on-off-keying (OOK)* waveform for the binary stream: **010111010001111**.
- 12.48.** Generate *amplitude shift keying (ASK)* waveform for the binary stream: **01011110001010**.
- 12.49.** Generate *frequency shift keying (FSK)* waveform for the binary stream: **0101101010000111**.
- 12.50.** Generate *phase shift keying (PSK)* waveform for the binary stream: **010110100001111**.
- 12.51.** Find the $(S/N)_{ratio}$ such that a unipolar binary system with AWGN has $P_e = 0.001$. What would be the error probability of a polar system with the same $(S/N)_{ratio}$?
- 12.52.** A binary system has AWGN with $N_o = 10^{-8}/r_b$. Find S_R for polar and unipolar signaling so that $p_e \leq 10^{-6}$.
- 12.53.** A polar binary system has 10 repeaters, each with $(S/N)_1 = 20 \text{ dB}$. Find P_e when the repeaters are *regenerative* and *nonregenerative*?
- 12.54.** A polar binary system with 20 repeaters has $P_e = 10^{-4}$. Find $(S/N)_1$ in dB when the repeaters are *regenerative* and *nonregenerative*.



Transmission Lines and Cables

13.1 Introduction

Transmission of information as an electromagnetic signal always occurs as a *transverse electromagnetic* (TEM) wave or as the combination of such waves as in a waveguide (see Chapter 14). The basic properties of the TEM wave are outlined in Appendix B, and propagation of these as radio waves is described in Chapter 15.

With transmission lines, the metallic conductors confine the TEM wave to the vicinity of the dielectric surrounding the conductors; as a result, some aspects of the transmission are best treated in terms of the *distributed circuit* parameters of the line, while others require the wave properties of the line to be taken into account. It must be clearly understood, however, that these are complementary views of the transmission; the voltages and currents on the line always accompany the TEM wave, and the particular viewpoint adopted usually depends on which properties are most easily measured.

Transmission lines may be *balanced* or *unbalanced* with respect to ground. The two basic types of transmission lines are the two-wire line, which is usually operated in the balanced mode [Fig. 13.1.1(a)], and the coaxial line, which is always operated in the unbalanced mode [Fig. 13.1.1(b)]. The electromagnetic field configurations for each type are shown in Fig. 13.1.1(c) and (d). In each case the direction of propagation of the TEM wave is into the page, and inspection shows that the *E* (electric) field is at right angles to the *H* (magnetic) field, and that both are at right angles to the direction of propagation, as required of a TEM wave.

With the coaxial line, the outer conductor forms a shield that confines the wave to the space between the conductors so that radiation from the line is negligible. However, the line is basically unbalanced since the external capacitance is between the outer conductor only and ground.

The two-wire line is normally operated in the balanced mode, the conductors being arranged so that they present equal capacitances to ground. (Of course, this may be a difficult condition to maintain in practice.) Radiation can occur from a two-wire line since the TEM wave can radiate out from the line as well as along it. The two-wire line is less expensive than the coaxial line and is used for the majority of low-frequency telephone

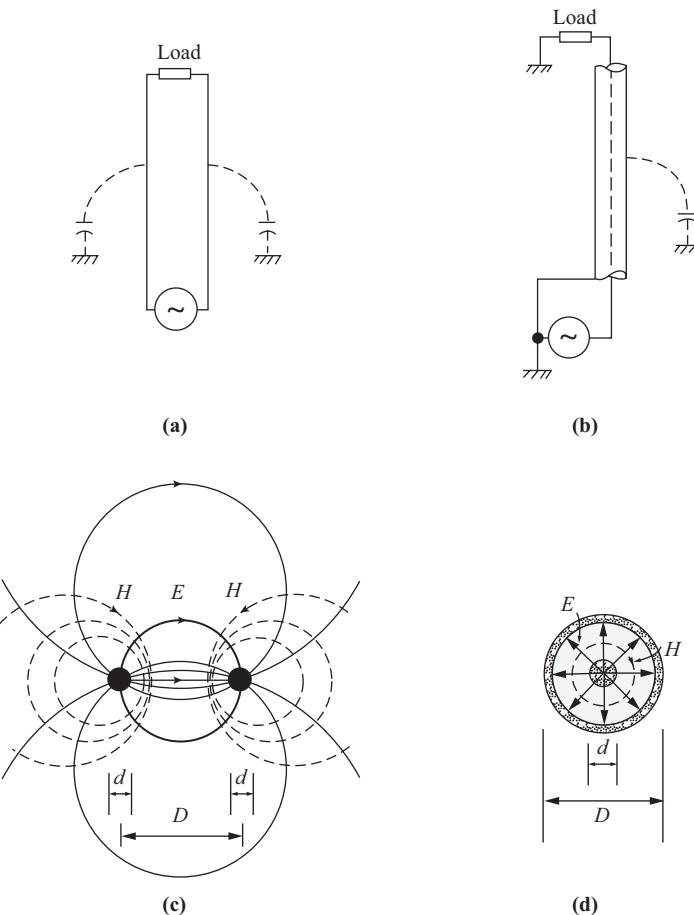


Figure 13.1.1 (a) Two-wire line, (b) coaxial line, (c) electromagnetic fields around a two-wire line, and (d) electromagnetic fields around a coaxial line.

circuits. Care must be taken to maintain balance conditions, as described in Section 13.14. For high-frequency circuits (including multichannel telephony and radio feeders), the coaxial line is used to minimize radiation; and where balanced radio antennas have to be connected to coaxial lines, special transmission line transformers known as *baluns* (*balanced* to *unbalanced*) are employed.

13.2 Primary Line Constants

From the circuit point of view, a transmission line will have series resistance and inductance, which together go to make up the series impedance of the conducting wires, and shunt conductance and capacitance of the dielectric between the conductors, which go to make up the shunt admittance of the line. A small length δx of the line may therefore be represented approximately by the filter section (Fig. 13.2.1), where, as part of the approximation, the values for both the *go* and *return* wires are included in the lumped components. The parameters R , L , G , and C shown in Fig. 13.2.1 are known as the *primary line constants*; these are the series resistance R , in ohms/meter; the series inductance L , in henries/meter; the shunt conductance G , in siemens/meter; and the shunt

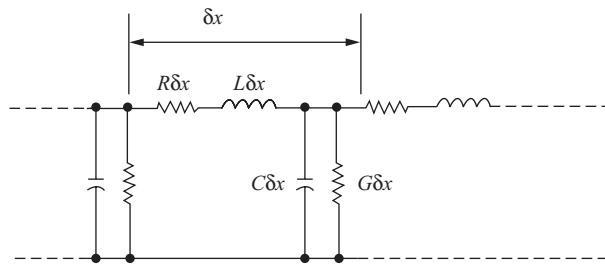


Figure 13.2.1 Circuit approximations for a short length δx of transmission line.

capacitance C , in farads/meter. The primary constants take into account both the *go* and *return* lines. They are constant in that they do not vary with voltage and current; however, they are frequency dependent to some extent. The series resistance R increases with frequency as a result of skin effect (see Section 1.6). The inductance L is almost independent of frequency for open lines, but tends to decrease with increasing frequency for screened cables. The capacitance C is almost independent of frequency, whereas the conductance G tends to increase with the frequency (that is, the shunt resistance decreases) because of increasing dielectric loss with increase in frequency.

No simple formulas can be given for the primary constants that cover all frequency ranges, but simplifications are possible for particular, well-defined operating ranges, for example, low or audio frequencies and high or radio frequencies. As will be shown in Section 13.4, the most practically useful characteristic of a line is the *characteristic impedance*, which at high frequencies is determined by the series inductance and shunt capacitance. These, as shown below, depend on line geometry, which therefore sets a limit on the range of characteristic impedances possible in practice.

For the two-wire line of Fig. 13.1.1(c), with the conductors embedded in a medium of permittivity ϵ (F/m) and permeability μ (H/m), and with the line dimensions in meters, the primary inductance and capacitance per unit length are given approximately by:

$$\text{Two-wire line: } L \cong \frac{\mu}{\pi} \ln\left(\frac{2D}{d}\right) \quad \text{H/m} \quad (13.2.1)$$

$$C \cong \frac{\pi\epsilon}{\ln(2D/d)} \quad \text{F/m} \quad (13.2.2)$$

For the coaxial line of Fig. 13.1.1(d), with dielectric of permittivity ϵ (F/m) and permeability μ (H/m), and again with the line dimensions in meters, the approximate forms of the equations are:

$$\text{Coaxial line: } L \cong \frac{\mu}{2\pi} \ln\left(\frac{D}{d}\right) \quad \text{H/m} \quad (13.2.3)$$

$$C \cong \frac{2\pi\epsilon}{\ln(D/d)} \quad \text{F/m} \quad (13.2.4)$$

13.3 Phase Velocity and Line Wavelength

In Appendix B the phase velocity of a TEM wave is given by Eq. (B.11) as

$$v_p = \frac{1}{\sqrt{\mu\epsilon}} \quad (13.3.1)$$

For free space the values are $\mu = \mu_0 = 4\pi \times 10^{-7}$ H/m and $\epsilon = \epsilon_0 = 8.854 \times 10^{-12}$ F/m, giving $v_p = 3 \times 10^8$ m/s. This is the velocity of light, normally denoted by the letter c . For a transmission line the permeability may be assumed equal to the free-space value, but the permittivity may differ from the free-space value, depending on the dielectric used. The expression for permittivity is $\epsilon = \epsilon_r \epsilon_0$ where ϵ_r is the relative permittivity (or dielectric constant). Substituting this in Eq. (13.3.1) along with the free-space values gives

$$v_p = \frac{c}{\sqrt{\epsilon_r}} \quad (13.3.2)$$

Typically, ϵ_r may range from 1 to 5, and thus the phase velocity of the TEM wave on the line may be less than the free-space value.

The wavelength of the wave is given by Eq. (B.4) as

$$\lambda = \frac{v_p}{f} \quad (13.3.3)$$

Substituting for v_p from Eq. (13.3.2) gives

$$\lambda = \frac{c}{f\sqrt{\epsilon_r}} \quad (13.3.4)$$

$$= \frac{\lambda_0}{\sqrt{\epsilon_r}} \quad (13.3.4)$$

where λ_0 is the free-space wavelength. The wavelength, as given by Eq. (13.3.4), is the value that must be used in transmission-line calculations; for example, the phase shift coefficient, which is the phase shift per unit length, is given by

$$\beta = \frac{2\pi}{\lambda} \quad (13.3.5)$$

where λ is as given by Eq. (13.3.4).

Another useful expression for the phase velocity can be obtained in terms of the inductance per unit length L and the capacitance per unit length C . From Eqs. (13.2.1) and (13.2.2) for the two-wire line, and Eqs. (13.2.3) and (13.2.4) for the coaxial line, it can be seen that

$$LC = \mu\epsilon \quad (13.3.6)$$

Hence, on substituting Eq. (13.3.6) in Eq. (13.3.1), the result is

$$v_p = \frac{1}{\sqrt{LC}} \quad (13.3.7)$$

13.4 Characteristic Impedance

Energy travels along a transmission line in the form of an electromagnetic wave, the wave set up by the signal source being known as the incident (or forward) wave. Only when the load impedance at the receiving end is a reflectionless match for the line, as discussed in Section 1.14, will all the energy be transferred to the load. If reflectionless matching is not achieved, energy will be reflected back along the line in the form of a

reflected wave (hence the name reflectionless matching). Because of the distributed nature of a transmission line, the question may be asked: Exactly to what impedance must the load be matched? This can be answered by considering a hypothetical line, infinite in length and for which no reflection can occur, since the incident wave never reaches the end. The ratio of maximum voltage to maximum current at any point on such a line is found to be constant, that is, independent of position. This ratio is known as the characteristic impedance Z_0 . Now, if a finite length of line is terminated in a load impedance $Z_L = Z_0$, this will appear as an infinite line to the incident wave since at all points, including the load termination, the ratio of voltage to current will equal Z_0 . Thus the characteristic impedance of a transmission line is the ratio of voltage to current at any point along the line on which no reflected wave exists.

With a sinusoidal signal of angular frequency ω rad/sec, the characteristic impedance in terms of the primary constants is found to be

$$Z_0 = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \quad \Omega \quad (13.4.1)$$

At low frequencies such that $R \gg \omega L$ and $G \gg \omega C$, the expression for Z_0 reduces to

$$Z_0 \approx \sqrt{\frac{R}{G}} \quad \Omega \quad (13.4.2)$$

and at high frequencies such that $R \ll \omega L$ and $G \ll \omega C$, it becomes

$$Z_0 \approx \sqrt{\frac{L}{C}} \quad \Omega \quad (13.4.3)$$

It will be observed that each limiting value is purely resistive (there is no j coefficient) and independent of frequency. Between these limits Z_0 is complex and frequency dependent, and it is found that for most practical lines it is capacitive. However, above a few tens of kilohertz for two-wire lines and a few hundred kilohertz for coaxial lines, the high-frequency approximation for Z_0 is sufficiently accurate for most practical purposes, and this will be the expression used throughout this chapter.

Substituting Eqs. (13.2.1) and (13.2.2) into Eq. (13.4.3) gives Z_0 in terms of line dimensions, permittivity, and permeability for the two-wire line:

$$Z_0 = \frac{1}{\pi} \sqrt{\frac{\mu}{\epsilon}} \ln\left(\frac{2D}{d}\right) \quad \Omega \quad (13.4.4)$$

And for the coaxial line, from Eqs. (13.2.3) and (13.2.4), Eq. (13.4.3) gives

$$Z_0 = \frac{1}{2\pi} \sqrt{\frac{\mu}{\epsilon}} \ln\left(\frac{D}{d}\right) \quad \Omega \quad (13.4.5)$$

For the dielectrics encountered in practice, the permeability will be equal to that of free space, $\mu = \mu_0 = 4\pi \times 10^{-7}$ H/m; and the permittivity will be given by $\epsilon = \epsilon_r \epsilon_0$, where $\epsilon_0 = 8.854 \times 10^{-12}$ F/m is the permittivity of free space and ϵ_r is the relative permittivity or dielectric constant. Substituting these into the impedance equations, Eqs. (13.4.4) and (13.4.5), gives

$$\text{Two-wire line: } Z_0 = \frac{120}{\sqrt{\epsilon_r}} \ln\left(\frac{2D}{d}\right) \quad \Omega \quad (13.4.6)$$

$$\text{Coaxial line: } Z_0 = \frac{60}{\sqrt{\epsilon_r}} \ln\left(\frac{D}{d}\right) \quad \Omega \quad (13.4.7)$$

In each case, it will be seen that for a given dielectric constant, the characteristic impedance is determined by the ratio D/d [Fig. 13.1.1(c) and (d)]. For dielectrics in common use, the dielectric constant will be within the range from 1 to 5, and practical limitations on the ratio D/d for each type of line limit Z_0 to the range of about 40 to 150 Ω for the coaxial line and 150 to 600 Ω for the two-wire line.

13.5 Propagation Coefficient

The propagation coefficient γ determines the variation of current or voltage with distance x along a transmission line. The current (and voltage) distribution along a matched line is found to vary exponentially with distance, the equations being

$$I = I_s e^{-\gamma x} \quad (13.5.1)$$

$$V = V_s e^{-\gamma x} \quad (13.5.2)$$

where I_s is the magnitude of the current and V_s is the magnitude of the voltage of the input or sending end of the line.

Like the characteristic impedance, the propagation coefficient also depends on the primary constants and the angular velocity of the signal. It is given by

$$\gamma = \sqrt{(R + j\omega L)(G + j\omega C)} \quad (13.5.3)$$

This is also a complex quantity and can be written as

$$\gamma = \alpha + j\beta \quad (13.5.4)$$

where α is known as the *attenuation coefficient* and determines how the voltage or current decreases with distance along the line, and β is known as the *phase-shift coefficient* and determines the phase angle of the voltage (or current) variation with distance. The phase-shift coefficient is the phase shift per unit length, and since a phase shift of 2π radians occurs over a distance of one wavelength λ , then

$$\beta = \frac{2\pi}{\lambda} \quad (13.5.5)$$

To see how the propagation coefficient affects the current, consider the current equation

$$I = I_s e^{-(\alpha + j\beta)x} \quad (13.5.6)$$

$$= I_s e^{-\alpha x} (e^{-j\beta x}) \quad (13.5.7)$$

The last equation can be represented graphically as shown in Fig. 13.5.1(a). The length of the phasor line represents $I_s e^{-\alpha x}$, and the angle of rotation from the reference line represents βx .

The attenuation is often expressed in terms of a unit known as the *neper*. The magnitude of the current (on the matched line) is

$$|I| = I_s e^{-\alpha x} \quad (13.5.8)$$

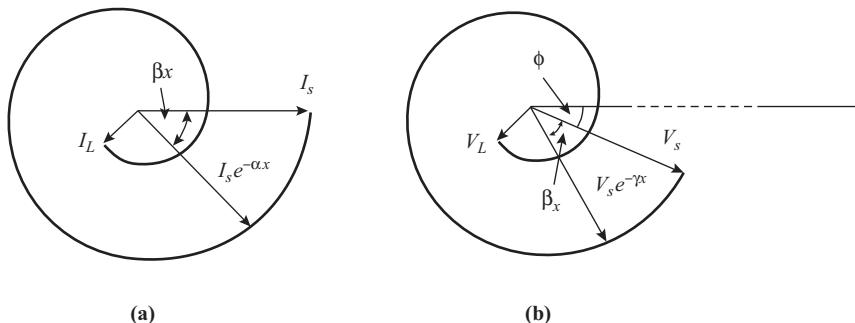


Figure 13.5.1 (a) Current and (b) voltage phasors diagrams.

and the attenuation of the current, in nepers, is defined as

$$N = -\ln\left(\frac{|I|}{I_s}\right) = -\ln(e^{-\alpha x}) \quad (13.5.9)$$

$$= \alpha x \text{ nepers} \quad (13.5.10)$$

The attenuation coefficient α can therefore be alternatively expressed as the attenuation in nepers per unit length of line:

$$\alpha = \frac{N}{x} \quad (13.5.11)$$

EXAMPLE 13.5.1

The input current in a matched line is 50 mA and the local current is 1 mA. The line is 1 km long. Calculate (a) the total attenuation in nepers and (b) the attenuation coefficient.

SOLUTION (a)

$$N = -\ln(1/50)$$

$$= 3.9 \text{ nepers}$$

(Note that by including a minus sign in the defining equation for the neper the attenuation comes out a positive number.)

(b)

$$\alpha = \frac{3.9}{1000}$$

$$= 3.9 \text{ mN/m}$$

The neper is a useful unit in theoretical work, but in practical work the decibel is more commonly used. In Appendix A, the relationship between the decibel value D and the neper value N of a given current ratio (Eq. A.8) is shown to be

$$D = 8.686N \quad (13.5.12)$$

Hence, for the transmission line,

$$D = 8.686\alpha x \text{ decibels} \quad (13.5.13)$$

where $N = \alpha x$. Therefore, if the attenuation coefficient is *defined* in decibels per unit length, say $[\alpha]$, then

$$[\alpha] = 8.686\alpha \text{ dB/m} \quad (13.5.14)$$

EXAMPLE 13.5.2

The attenuation coefficient of a line is 0.0006 N/m. Determine the attenuation coefficient in (a) dB/m and (b) dB/mile.

SOLUTION (a) $[\alpha] = 8.686\alpha$, and since α numerically equals the attenuation in nepers/meter,

$$\begin{aligned} [\alpha] &= 8.686 \times 0.0006 \\ &= \mathbf{0.00521 \text{ dB/m}} \end{aligned}$$

(b) Since there are 1609 m in 1 mile, the attenuation coefficient in dB/mi is

$$0.00521 \times 1609 = \mathbf{8.4 \text{ dB/mi}}$$

For a matched line, $V = IZ_0$ at any point along the line, and for $Z_0 = Z_0 \angle -\phi$ (assuming a leading phase angle), the voltage–distance phasor diagram will be similar to that for current, as shown in Fig. 13.5.1(b). The length of the phasor is modified by Z_0 , and the reference line for the voltage is displaced by $-\phi$.

Both α and β are determined by the primary constants and the frequency, since

$$\gamma = \alpha + j\beta = \sqrt{(R + j\omega L)(G + j\omega C)} \quad (13.5.15)$$

The square-root expression can be expanded by means of the binomial expansion, and an approximation sufficiently accurate for most practical purposes is to take the expansion to the third term. This results in the expression

$$\alpha \approx \frac{R}{2Z_0} + \frac{GZ_0}{2} \quad (13.5.16)$$

where Z_0 in this particular case is $\sqrt{L/C}$ (that is, the high-frequency limit to the characteristic impedance). In most practical lines G is very small (almost zero), so the further approximation

$$\alpha \approx \frac{R}{2Z_0} \quad (13.5.17)$$

is usually valid.

As an example, the primary constants for a coaxial cable at a frequency of 10 MHz were determined approximately as

$$L = 234 \text{ nH/m}$$

$$C = 93.5 \text{ pF/m}$$

$$R = 0.568 \Omega/\text{m}$$

$$G = 0$$

Therefore, at a frequency of 10 MHz,

$$Z_0 \cong \sqrt{\frac{234 \times 10^{-9}}{93.5 \times 10^{-12}}} = 50 \Omega \quad (13.5.18)$$

and

$$\alpha = \frac{0.568}{2 \times 50} = 0.00568 \text{ N/m} \quad (13.5.19)$$

The attenuation, in dB/m [α], is $0.00568 \times 8.686 = 0.0493$ dB/m. The major variation in R with frequency is due to skin effect, for which R is proportional to the square root of frequency, and this will cause the attenuation to vary likewise.

The binomial expansion for γ results in the following expression for β :

$$\beta \cong \omega \sqrt{LC} \left[1 + \frac{1}{8} \left(\frac{R}{\omega L} - \frac{G}{\omega C} \right)^2 \right] \quad (13.5.20)$$

An interesting situation arises when

$$\frac{R}{\omega L} = \frac{G}{\omega C} \quad (13.5.21)$$

The phase-shift coefficient is then seen to be equal to

$$\beta = \omega \sqrt{LC} \quad (13.5.22)$$

This meets the condition, discussed in the following section, required for distortionless transmission (that is, β is proportional to ω). The condition for distortionless transmission is seen to be

$$\frac{R}{\omega L} = \frac{G}{\omega C}$$

and this can be rearranged as

$$\frac{R}{G} = \frac{L}{C} \quad (13.5.23)$$

In any practical line, R and G both will be as small as possible in order to keep losses at a minimum. The ratio L/C will not, in general, be equal to R/G , and therefore, if distortionless transmission is desired, the ratio L/C will have to be altered to equal R/G . In all practical cases this is achieved by increasing L , a technique known as *loading*. One common form of loading is to add inductance coils in series with the line at regularly spaced intervals. However, it is not practical to load to achieve completely distortionless conditions, as it is found that this would require an inordinately large value of inductance. Furthermore, large inductance decreases phase velocity [see Eq. (13.3.7)], which may introduce unacceptable delays on long-distance telephone circuits. It is interesting to note that the introduction of pulse code modulation on normal telephone lines requires the removal of the loading coils in order to increase the bandwidth of the lines.

13.6 Phase and Group Velocities

The phase velocity of an electromagnetic wave has already been discussed briefly in Section 13.3. For wave motion generally, the following simple relationship exists for frequency f , wavelength λ , and phase velocity v_p (see Appendix B):

$$\lambda f = v_p \quad (13.6.1)$$

Since $\beta = 2\pi/\lambda$ and $\omega = 2\pi f$,

$$\frac{\omega}{\beta} = v_p \quad (13.6.2)$$

It can be seen therefore that while β is proportional to ω , the phase velocity will be constant (that is, independent of frequency), and therefore all component waves making up a signal will be transmitted at the same velocity v_p . This is the distortionless transmission condition referred to in the previous section.

The situation can occur where β is not proportional to ω , as, for example, in the general case given by Eq. (13.5.20). Component sine waves of a signal will be transmitted with different velocities, and the question then arises, at what velocity does the signal wave travel? An answer to this question can be obtained by considering two sinusoidal waves differing in frequency by a small amount $\delta\omega$. The combined wave is shown in Fig. 13.6.1, where, for clarity, the individual sine waves are assumed to have equal amplitudes, and line attenuation is ignored. The composite signal is seen to consist of high-frequency waves (many zero crossings on the time axis) modulated by a low-frequency envelope. Detailed analysis shows that the envelope travels along the line with a velocity given by

$$v_g = \frac{\delta\omega}{\delta\beta} \quad (13.6.3)$$

where the subscript g signifies *group velocity* (the velocity at which the group of two sine waves travels). It is also the velocity at which the energy is propagated along the line.

In the limit the group velocity is given by the differential coefficient of ω with respect to β , and if β varies rapidly with ω , serious distortion will result.

The condition stated previously for distortionless transmission was that v_p be constant. A more general condition will now be stated: for distortionless transmission, the β/ω graph should be a straight line, which may be expressed as

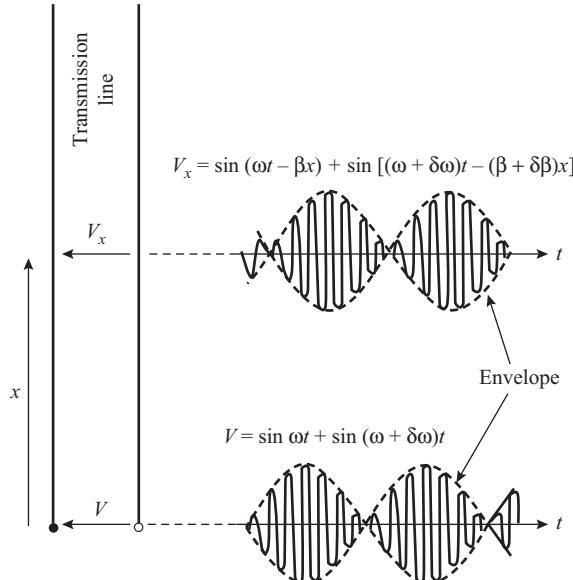


Figure 13.6.1 Wave group used in determination of group velocity.

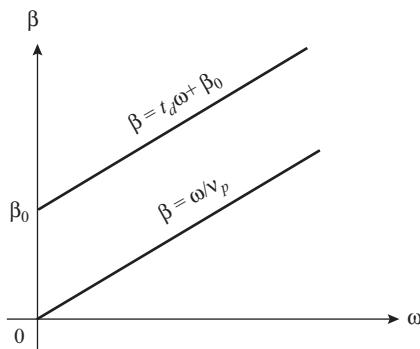


Figure 13.6.2 β/ω graph for distortionless transmission.

$$\beta = t_d \omega + \beta_0 \quad (13.6.4)$$

where t_d , known as the *group delay time*, is the slope of the line, and β_0 is the intercept, which must be equal to $\pm n\pi$ for distortionless transmission. This graph is sketched in Fig. 13.6.2, along with the line representing the simpler condition, $\beta = \omega/v_p$.

It will be seen that t_d is the reciprocal of the group velocity, and therefore, in addition to the constraint that $\beta_0 = \pm n\pi$, v_g must be constant (for constant t_d) for distortionless transmission. This more general set of conditions covers situations where distortionless transmission is achieved even though the simpler condition ($\beta/\omega = \text{constant}$) is violated.

13.7 Standing Waves

When the load impedance does not match the line impedance, part of the energy in the incident wave is reflected at the load. This gives rise to a reflected wave traveling back along the line toward the source. If the source impedance does not match the line, a further reflection will take place, and, in this way, multiple reflections can be set up at both load and source. The overall effect can be treated as the resultant of a single incident and a single reflected wave. These can be represented by rotating phasors in the manner of Fig. 13.5.1, but with the reflected wave phasor rotating in the opposite direction to the incident wave phasor [Fig. 13.7.1(a)]. This is because from a given point an increase in x away from the source is a decrease in distance x toward the point of reflection.

Clearly, at certain values of x , the phasors will be in opposition, giving rise to voltage minima [Fig. 13.7.1(b)], while at other values of x they will coincide [Fig. 13.7.1(c)], giving rise to voltage maxima. The voltage as a function of distance x is sketched in Fig. 13.7.1(d), where it can be seen to go through a series of maxima and minima. This voltage pattern is stationary as regards position and is therefore referred to as the *voltage standing wave (VSW)*. It is essential to grasp, however, that at any given point, the voltage is time varying at the frequency of the input signal. Thus, at point x_1 , corresponding to the first minimum, the voltage alternates sinusoidally between the peak values $\pm V_{1\text{ min}}$, as shown in Fig. 13.7.1(e).

Another important feature of the voltage standing-wave pattern is that the distance between successive minima (and between successive maxima) is $\lambda/2$. This is easily shown from the fact that when they are at a minimum the phasors form a straight line [Fig. 13.7.1(b)]. On moving from position x_1 to x_2 , the phasors must

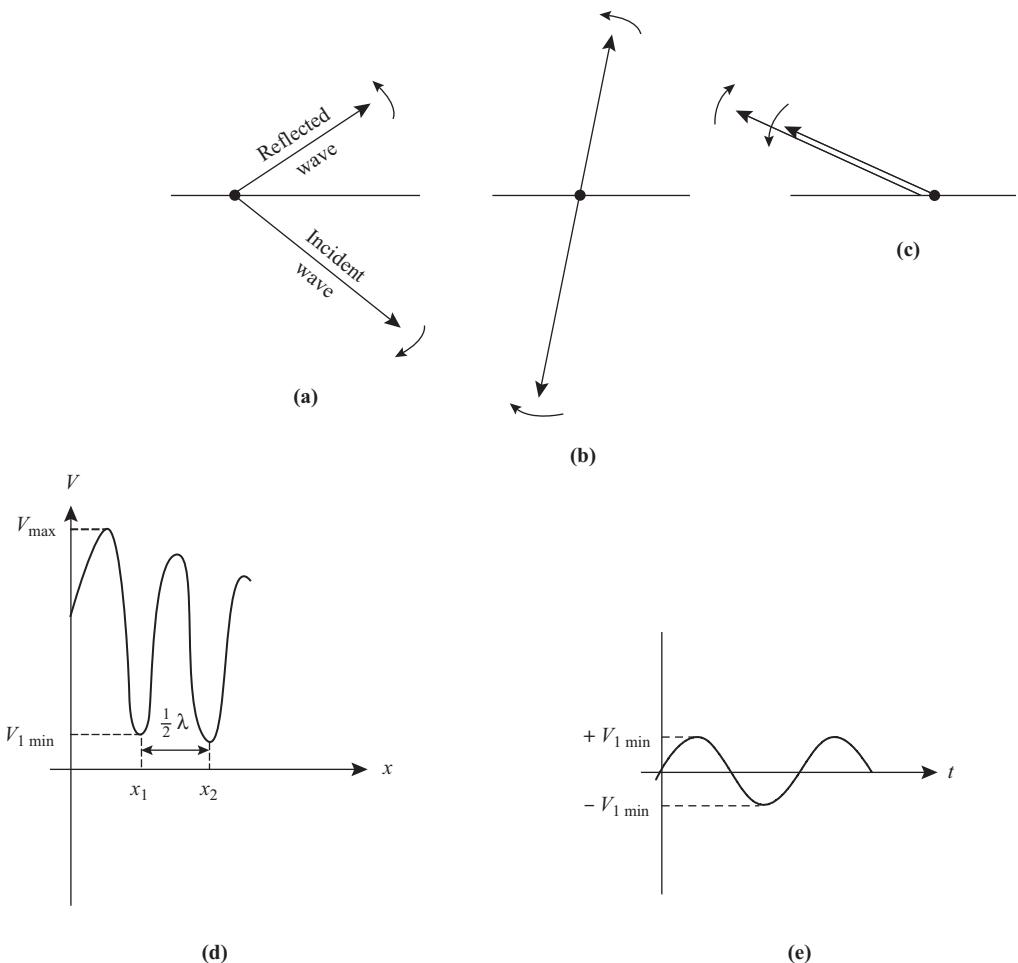


Figure 13.7.1 (a) Phasor representation of forward and reflected waves, (b) phasors at a voltage minimum, (c) phasors at a voltage maximum, (d) voltage standing wave, and (e) the voltage–time variation.

rotate to interchange positions (that is, each phasor rotates by a phase angle of π radians). But each phasor also rotates by an angle given by $\beta(x_2 - x_1)$, where $\beta = 2\pi/\lambda$, the phase-shift coefficient. Therefore,

$$\frac{2\pi}{\lambda}(x_2 - x_1) = \pi \quad (13.7.1)$$

$$\therefore (x_2 - x_1) = \frac{1}{2}\lambda \quad (13.7.2)$$

In a similar way, it is easy to show that the distance between a minimum and the following maximum is $\frac{1}{4}\lambda$.

A current standing wave will also occur. For a purely resistive Z_0 , the incident current wave will always be in phase with the incident voltage wave, while the reflected current wave will always be antiphase to the reflected voltage wave. This happens because either the electric or the magnetic field of the reflected wave

must reverse direction, as shown in Fig. B.1(b). A consequence of the antiphase relationship between reflected voltage and current is that the current maxima occur along with the voltage minima, and the current minima along with the voltage maxima, for the standing wave patterns (bearing in mind that these are the phasor sums of incident and reflected waves). As already mentioned, this condition only holds when Z_0 is purely resistive. It enters into the theory of slotted-line measurements discussed in Section 13.10.

13.8 Lossless Lines at Radio Frequencies

For many applications at radio frequencies, the losses in a transmission line are small enough to be ignored (for example, in short lengths of good-quality cable), and therefore the attenuation coefficient α can be set equal to zero. The propagation coefficient is then given by

$$\gamma = j\beta \quad (13.8.1)$$

It is often more convenient to measure distances from the load end rather than from the sending end. Denoting this by ℓ , an equation for the incident voltage, similar in form to Eq. (13.5.2), may be written:

$$V_i = V_I e^{j\beta\ell} \quad (13.8.2)$$

Here, V_i is the incident wave voltage at any distance ℓ from the load; V_I , which may be complex, is the value at the load ($\ell = 0$); and $j\beta$ replaces γ of Eq. (13.5.2). A positive exponential is used in Eq. (13.8.2) compared to the negative exponential in Eq. (13.5.2) since ℓ is measured in the opposite direction to x .

A similar equation can be written for the reflected voltage wave:

$$V_r = V_R e^{-j\beta\ell} \quad (13.8.3)$$

Here, V_R (which may be complex) is the value of the reflected voltage at the load, and, of course, a negative exponential is used since the reflected wave's phase changes in the opposite sense to that of the incident wave. At any point on the line,

$$V = V_i + V_r \quad (13.8.4)$$

In particular, at the load ($\ell = 0$),

$$V_L = V_I + V_R \quad (13.8.5)$$

The equations for current are

$$\text{Incident wave: } I_i = \frac{V_i}{Z_0} \quad (13.8.6)$$

$$\text{Reflected wave: } I_r = -\frac{V_r}{Z_0} \quad (13.8.7)$$

Note the minus sign for the reflected current wave, signifying the 180° phase change discussed in the previous section.

The resultant current at any point ℓ from the load is

$$I = I_i + I_r \quad (13.8.8)$$

$$= \frac{V_i - V_r}{Z_0} \quad (13.8.9)$$

In particular, the load current is

$$I_L = \frac{V_I - V_R}{Z_0} \quad (13.8.10)$$

Normally it is not necessary to know V_I and V_R separately, the ratio V_R/V_I entering more into calculations. This ratio is termed the *voltage reflection coefficient* Γ_L :

$$\Gamma_L = \frac{V_R}{V_I} \quad (13.8.11)$$

Note that Γ_L is defined in terms of the incident and reflected waves *at the load*.

From Eqs. (13.8.5) and (13.8.6), the load impedance may be expressed as

$$\begin{aligned} Z_L &= \frac{V_L}{I_L} \\ &= \frac{V_I + V_R}{V_I - V_R} Z_0 \end{aligned} \quad (13.8.12)$$

from which

$$\Gamma_L = \frac{Z_L - Z_0}{Z_L + Z_0} \quad (13.8.13)$$

The fact that the voltage reflection coefficient is determined solely by the load impedance and the characteristic impedance makes it an important practical parameter.

Three specific load conditions that occur frequently are (i) matched load, (ii) short-circuit load, and (iii) open-circuit load. For each of these, the following obtain:

(i) Matched load, $Z_L = Z_0$: $\Gamma_L = 0$

(ii) Short circuit, $Z_L = 0$: $\Gamma_L = -1$

(iii) Open circuit, $Z_L = \infty$: Here it is first necessary to rearrange the equation as

$$\begin{aligned} \Gamma_L &= \frac{1 - Z_0/Z_L}{1 + Z_0/Z_L} \\ &= 1 \quad (\text{as } Z_L \text{ goes to infinity}) \end{aligned} \quad (13.8.14)$$

It is left as an exercise for the student to show that the corresponding load conditions for the three cases are

- (i) $V_L = V_1$ and $I_L = V_I/Z_0$
- (ii) $V_L = 0$ and $I_L = 2V_I/Z_0$
- (iii) $V_L = 2V_1$ and $I_L = 0$

13.9 Voltage Standing-wave Ratio

The *voltage standing-wave ratio* (VSWR) is defined as

$$\text{VSWR} = \frac{V_{\max}}{V_{\min}} \quad (13.9.1)$$

where V_{\max} and V_{\min} are as shown in Fig. 13.7.1. The line is assumed lossless so that the maxima all have the same value V_{\max} , and the minima all have the value V_{\min} .

As previously shown [Fig. 13.7.1(c)], a maximum occurs when the phasors are in phase:

$$V_{\max} = |V_I| + |V_R| \quad (13.9.2)$$

$$= |V_I|(1 + |\Gamma_L|) \quad (13.9.3)$$

(Note: $|V_I| + |V_R|$ is not the same as $|V_I + V_R|$, which is the modulus of the load voltage).

A minimum occurs when the phasors are antiphase:

$$V_{\min} = |V_I|(1 - |\Gamma_L|) \quad (13.9.4)$$

Therefore,

$$\text{VSWR} = \frac{|V_I|(1 + |\Gamma_L|)}{|V_I|(1 - |\Gamma_L|)}$$

or

$$\text{VSWR} = \frac{(1 + |\Gamma_L|)}{(1 - |\Gamma_L|)} \quad (13.9.5)$$

The VSWR can range in value from unity to infinity; that is,

$$1 \leq \text{VSWR} \leq \infty \quad (13.9.6)$$

Ideally, the VSWR should equal 1, as this represents a matched condition, and practical adjustments on RF transmission lines are often aimed at minimizing the VSWR. (Note that the VSWR is always a real number that has no imaginary part.)

The equation for VSWR can be rearranged to give

$$|\Gamma_L| = \frac{\text{VSWR} - 1}{\text{VSWR} + 1} \quad (13.9.7)$$

13.10 Slotted-line Measurements at Radio Frequencies

Use is made of the voltage standing wave at high frequencies to determine unknown impedances that would otherwise be difficult to measure. In principle, the method requires only a determination of the VSWR and the distance of the first minimum from the load (the minimum rather than the maximum is chosen in practice because it is more sharply defined). The apparatus consists of a slotted section of coaxial line through which a probe can sample the electric field and hence the voltage standing wave. The output signal from the probe is often converted to a signal of lower frequency and then amplified and rectified to produce dc output proportional to the voltage standing-wave amplitude; alternatively, for simpler measurements, the probe output may be directly rectified and the dc read by a microammeter. The distance of the probe from the load can be read directly off a calibrated distance scale. There will always be specific corrections to make for a given apparatus to allow for end effects, details of which should be included in the manufacturer's handbook. The theory of the measurement technique is as follows.

At distance ℓ from the load, the voltage reflection coefficient is found, using Eqs. (13.8.2) and (13.8.3), to be

$$\begin{aligned}\Gamma &= \frac{V_r}{V_i} \\ &= \Gamma_L e^{-j2\beta\ell}\end{aligned}\quad (13.10.1)$$

Since Γ_L is complex, it may be written in terms of its modulus and phase angle as

$$\Gamma_L = |\Gamma_L| e^{j\phi_L} \quad (13.10.2)$$

Combining Eqs. (13.10.1) and (13.10.2) gives

$$\begin{aligned}\Gamma &= |\Gamma_L| e^{j(\phi_L - 2\beta\ell)} \\ &= |\Gamma_L| \angle \phi_L - 2\beta\ell\end{aligned}\quad (13.10.3)$$

In particular, let ℓ_{\min} represent the distance from the load to the first voltage minimum; then the voltage reflection coefficient at the minimum is

$$\Gamma_{\min} = |\Gamma_L| \angle \phi_L - 2\beta\ell_{\min} \quad (13.10.4)$$

The angle $(\phi_L - 2\beta\ell_{\min})$ is the phase angle of the reflected voltage with respect to the incident voltage at the first voltage minimum. At a voltage minimum, the two voltages are in antiphase; therefore, the reflected voltage lags the incident voltage by π radians. This is because the incident voltage advances in phase as ℓ increases (shown by the $+2\beta\ell$ term), while the reflected voltage lags (as shown by the $-2\beta\ell$ term). Therefore,

$$\phi_L - 2\beta\ell_{\min} = -\pi$$

from which it follows that

$$\phi_L = \left(\frac{4\ell_{\min}}{\lambda} - 1 \right) \pi \quad (13.10.5)$$

This shows that the angle ϕ_L can also be determined by slotted line measurements, and since the modulus $|\Gamma_L|$ is determined from the VSWR measurement, as shown by Eq. (13.9.7), the reflection coefficient at the load is known. This in turn allows the load impedance to be found, using Eq. (13.8.13), to be

$$Z_L = Z_0 \frac{1 + \Gamma_L}{1 - \Gamma_L} \quad (13.10.6)$$

Although ℓ_{\min} was specified as the distance to the first voltage minimum, in fact any voltage minimum can be used (assuming losses can be ignored) since minima are separated by $n\lambda/2$, and this will simply add $n2\pi$ radians to ϕ_L , which has no effect on the principal value.

EXAMPLE 13.10.1

Measurements on a $50\text{-}\Omega$ slotted line gave a VSWR of 2.0 and a distance from load to first minimum of 0.2λ . Determine both the equivalent series and equivalent parallel circuits for Z_L .

SOLUTION From Eq. (13.9.7),

$$|\Gamma_L| = \frac{\text{VSWR} - 1}{\text{VSWR} + 1} = \frac{2 - 1}{2 + 1} = \frac{1}{3}$$

Also,

$$\begin{aligned}\phi_L &= \pi(4 \times 0.2 - 1) \\ &= -36^\circ\end{aligned}$$

But

$$\begin{aligned}e^{j\phi_L} &= \cos \phi_L + j \sin \phi_L \quad (\text{Euler's identity}) \\ &= 0.809 - j0.588\end{aligned}$$

and

$$\begin{aligned}|\Gamma_L| e^{j\phi_L} &= \frac{0.809 - j0.588}{3} \\ &= 0.27 - j0.196\end{aligned}$$

From Eq. (13.10.6),

$$\begin{aligned}Z_L &= 50 \frac{1 + (0.27 - j0.196)}{1 - (0.27 - j0.196)} \\ &= 50 \frac{1.27 - j0.196}{0.73 + j0.196}\end{aligned}$$

The series equivalent is $Z_L = R_s + jX_s$ and, therefore,

$$\begin{aligned}R_s + jX_s &= 50 \frac{(1.27 - j0.196)(0.73 - j0.196)}{0.73^2 + 0.196^2} \\ &= 77.8 - j34.3\end{aligned}$$

from which

$$R_s = 77.8 \Omega$$

$$|X_s| = 34.3 \Omega \quad (\text{capacitive})$$

The parallel equivalent is $Y_L = G + jB$, where $Y_L = 1/Z_L$, $G = 1/R_p$, and $B = 1/X_p$. Therefore, from the above expression for Z_L ,

$$\begin{aligned}
 G + jB &= \frac{0.73 + j0.196}{50(1.27 - j0.196)} \\
 &= \frac{(0.73 + j0.196)(1.27 + j0.196)}{50(1.27^2 - 0.196^2)} \\
 &= 10.76 + j4.75 \text{ mS}
 \end{aligned}$$

from which

$$\begin{aligned}
 R_p &= \frac{1}{G} \\
 &= \frac{10^3}{10.76} \\
 &= 92.9 \Omega
 \end{aligned}$$

and

$$\begin{aligned}
 |X_p| &= \left| \frac{1}{B} \right| \\
 &= \frac{10^3}{4.75} \\
 &= 211 \Omega \quad (\text{capacitive})
 \end{aligned}$$

13.11 Transmission Lines as Circuit Elements

The voltage reflection coefficient at any point on the line is defined as

$$\Gamma = \frac{V_r}{V_i} \quad (13.11.1)$$

It will be seen that Eq. (13.8.12) for the voltage reflection coefficient at the load is a particular case of this.

Following the same line of reasoning used to derive Eq. (13.10.6) for the load impedance, the impedance at any point on the line any distance from the load can be written in terms of the voltage reflection coefficient at that point as

$$Z = Z_0 \frac{1 + \Gamma}{1 - \Gamma} \quad (13.11.2)$$

Certain cases are of particular importance:

Case (i): $Z_L = 0$ (*a short circuit*), Fig. 13.11.1(a). From Eq. (13.8.13), $\Gamma_L = -1$, and hence, from Eq. (13.10.1), $\Gamma = -e^{-j2\beta\ell}$. Substituting this in Eq. (13.11.2) gives

$$\begin{aligned}
 Z &= Z_0 \frac{1 - e^{-j2\beta\ell}}{1 + e^{-j2\beta\ell}} \\
 &= jZ_0 \tan\beta\ell
 \end{aligned} \quad (13.11.3)$$

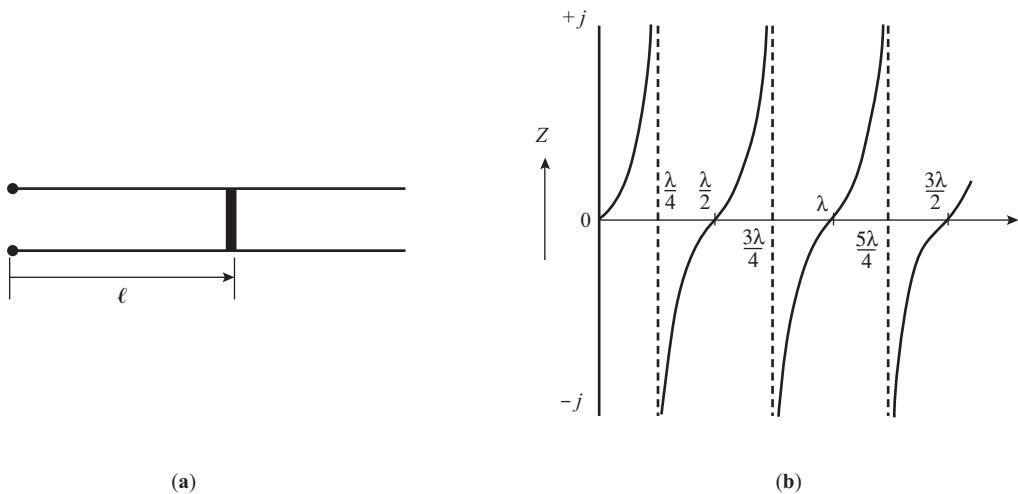


Figure 13.11.1 (a) Short-circuited line. (b) Reactance variation with line length.

The latter step makes use of the trigonometric identity $j \tan \theta = (1 - e^{-j2\theta})/(1 + (e^{-j2\theta}))$. The graph of Z/ℓ is shown in Fig. 13.11.1(b), where it is seen that for $0 \leq \ell \leq 4$ and $\lambda/2 \leq \ell \leq 3\lambda/4$, and so on, Z is inductive ($+j$). In practice, to obtain a variable inductive reactance, the short circuit on the line would be variable in such manner that ℓ could be varied between 0 and $\lambda/4$. (The other possible ranges are usually avoided since the longer line lengths are undesirable practically.)

The equivalent inductance L_{eq} is obtained by equating

$$j\omega L_{eq} = jZ_0 \tan \beta \ell \quad \left(0 \leq \ell \leq \frac{\lambda}{4}\right) \quad (13.11.4)$$

or

$$L_{eq} = \frac{Z_0}{\omega} \tan \beta \ell \quad (13.11.5)$$

EXAMPLE 13.11.1

A 50Ω short-circuited line is 0.1λ in length, at a frequency of 500 MHz. Calculate (a) the equivalent inductive reactance and (b) the equivalent inductance.

SOLUTION

$$\beta \ell = 2\pi \times 0.1$$

$$= 36^\circ$$

(a)

$$\begin{aligned} Z &= j50 \times \tan 36^\circ \\ &= j50 \times 0.7265 \\ &= \mathbf{j36.33 \Omega} \end{aligned}$$

(b)

$$\begin{aligned} L_{eq} &= \frac{36.33}{2\pi \times 500 \times 10^6} \times 10^9 \text{ nH} \\ &= \mathbf{11.6 \text{ nH}} \end{aligned}$$

Because the inductive reactance is not directly proportional to frequency, but to $\tan \beta\ell$, L_{eq} shows a frequency dependence. In the previous example, let the frequency be doubled; then

$$\ell = 0.2\lambda$$

$$= 72^\circ$$

$$Z = j50 \times \tan 72^\circ$$

$$= j153.9 \Omega \quad (\text{well over four times the previous value})$$

$$L_{\text{eq}} = \frac{153.9}{2\pi \times 1000 \times 10^6} \times 10^9 \text{nH}$$

$$= 24.5 \text{nH}$$

From Fig. 13.11.1(b) it is also seen that $\frac{1}{4}\lambda \leq \ell \leq \frac{1}{2}\lambda$ (or $\frac{3}{4}\lambda \leq \ell \leq \lambda, \dots$); that is, the impedance appears capacitive ($-j$). The equivalent capacitive impedance is given by

$$-j \frac{1}{\omega C_{\text{eq}}} = jZ_0 \tan \beta\ell \quad \left(\frac{1}{4}\lambda \leq \ell \leq \frac{1}{2}\lambda \right) \quad (13.11.6)$$

or

$$C_{\text{eq}} = \frac{1}{\omega Z_0 \tan \beta\ell} \quad (13.11.7)$$

For example, let $\ell = 3\lambda/8$ at a frequency of 500 MHz; then

$$\beta\ell = 135^\circ$$

$$Z = j50 \times \tan 135^\circ$$

$$= -j50 \times \tan 45^\circ$$

$$= -j50 \Omega$$

$$C_{\text{eq}} = \frac{10^{12}}{2\pi \times 500 \times 10^6 \times 50} \text{pF}$$

$$= 6.4 \text{ pF}$$

Case (ii): $Z_L = \infty$ (an open circuit), Fig. 13.11.2(a). In this case, $\Gamma_L = +1$, so Z becomes

$$Z = Z_0 \frac{1 + e^{-j2\beta\ell}}{1 - e^{-j2\beta\ell}}$$

$$= -jZ_0 \cot \beta\ell \quad (13.11.8)$$

The graph of Z/ℓ is shown in Fig. 13.11.2(b), where it will be seen that when $0 \leq \ell \leq \lambda/4$ the open-circuited line appears capacitive, and when $\lambda/4 \leq \ell \leq \lambda/2$, it appears inductive (with the pattern repeating at $1/2 \lambda$ intervals). Although the open-circuited line finds some use, it is not as practically convenient as the short-circuited line. The short circuit provides a mechanical support for the end of the line, it is easily adjustable, and the length ℓ is well defined. However, open-circuited lines are readily fabricated in microstrip and stripline construction.

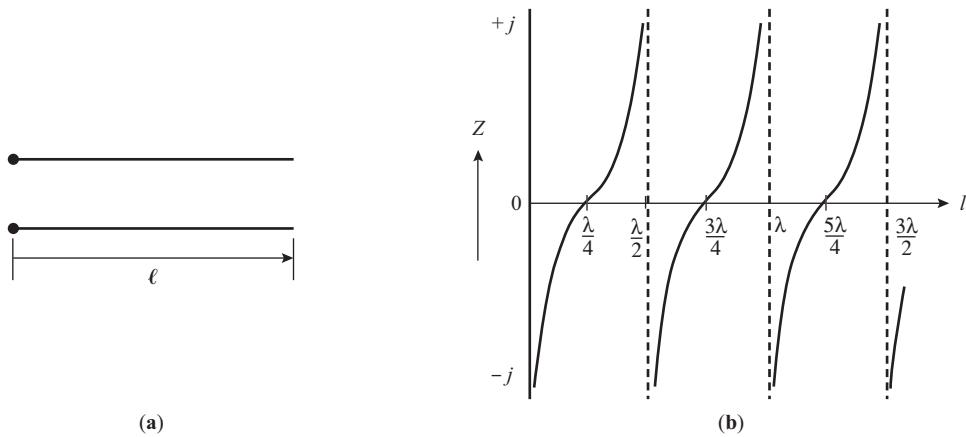


Figure 13.11.2 (a) Open-circuited line, (b) Reactance variation with line length.

Case (iii): $\ell = \lambda/4$

$$\begin{aligned}
 \beta\Omega &= \frac{2\pi}{\lambda} \frac{\lambda}{4} = \frac{\pi}{2} \\
 Z &= Z_0 \frac{1 + \Gamma_L e^{-j\pi}}{1 - \Gamma_L e^{-j\pi}} \\
 &= Z_0 \frac{1 + \Gamma_L}{1 - \Gamma_L} \\
 &= Z_0 \frac{Z_0}{Z_L} \\
 &= \frac{Z_0^2}{Z_L} \tag{13.11.9}
 \end{aligned}$$

This important relationship shows that the impedance as seen at the input to the line is Z_0^2/Z_L , and therefore a $\lambda/4$ section of line may be used to transform an impedance value from Z_L to Z_0^2/Z_L . For this reason, the $\lambda/4$ section is often known as a *quarter-wave transformer*.

Suppose, for example, it is required to match a $73\text{-}\Omega$ antenna to a $600\text{-}\Omega$ feeder line at a frequency of 150 MHz, shown in Fig. 13.11.3. To match the main feeder, the impedance Z'_L must equal $600\ \Omega$. Therefore,

$$Z'_L = \frac{(Z'_L)^2}{Z_L}$$

and

$$\begin{aligned}
 Z'_0 &= \sqrt{Z_L Z'_L} \\
 &= \sqrt{73 \times 600} \\
 &= 209\ \Omega
 \end{aligned}$$

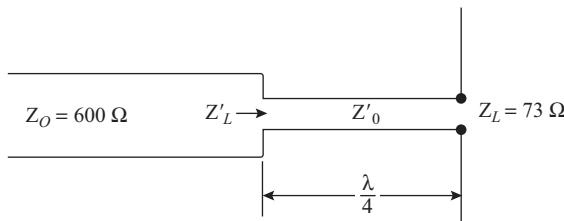


Figure 13.11.3 $\lambda/4$ transformer matching.

With this arrangement, there will be no standing waves on the main feeder line (600Ω); that is, all the power transmitted along the feeder will pass smoothly into the $\lambda/4$ section. There will be a mismatch between the $\lambda/4$ section and the 73Ω load, so a standing wave will occur on the $\lambda/4$ section; but, in practice, this can be tolerated because the matching section is considerably shorter than the main feeder. As can be easily verified, the reflection coefficient is -0.52 with, and -0.78 without, the $\lambda/4$ matching section.

Since the $\lambda/4$ section is a resonant element (that is, it is only fully effective at the frequency that makes the length exactly $\lambda/4$), it is only useful over a narrow range of frequencies (for example, voice-modulated radio waves). However, special broadbanding techniques are available.

Two cases of special interest arise when $Z_L = 0$ (short circuit) and $Z_L = \infty$ (open circuit). For the short-circuit load,

$$Z = \frac{Z_0^2}{0} = \infty \quad (13.11.10)$$

That is, the $\frac{1}{4}\lambda$ short-circuited section appears as a parallel resonant circuit (high impedance). In practice, line resistance will restrict Z to a finite, but high, value; the equivalent Q factor of the circuit will be high, on the order of 3000.

For the open-circuit line,

$$Z = \frac{Z_0^2}{\infty} = 0 \quad (13.11.11)$$

That is, the open-circuit $\frac{1}{4}\lambda$ section appears as a series resonant circuit (low impedance). Again, in practice, line resistance will result in Z being a small positive resistance, not zero.

13.12 Smith Chart

The work involved in transmission-line calculations may be considerably reduced by using a special chart known as a *Smith chart* (devised by P. H. Smith in 1944). The theory behind the chart is too involved to be given here, but its use will be explained.

The chart is based on the relationship given by Eq. (13.11.2). Normalized values of impedance are frequently used. Normalized impedance is simply the ratio of Z/Z_0 and is denoted by the lower-case letter z :

$$z = \frac{Z}{Z_0} \quad (13.12.1)$$

It is important to note that normalized impedance is dimensionless. In terms of normalized impedance, Eq.(13.11.2) becomes

$$z = \frac{1 + \Gamma}{1 - \Gamma} \quad (13.12.2)$$

In terms of normalized resistance r and normalized reactance x , the normalized impedance is $z = r + jx$. Any point on the Smith chart shows the four quantities r , x , $|\Gamma|$, and ϕ . Point z_1 on Fig. 13.12.1(a) shows $r_1 = 0.55$, $x_1 = 1.6$, $|\Gamma_1| = 0.746$, and $\phi_1 = 60^\circ$; that is,

$$\begin{aligned} z_1 &= 0.55 + j1.6 \\ \Gamma_1 &= 0.746 /60^\circ \end{aligned}$$

Similarly, point z_2 represents $z_2 = 0.55 - j0.3$, and $\Gamma_2 = 0.34 \angle -135^\circ$. To find the actual impedance values corresponding to z_1 and z_2 , the characteristic impedance of the line must be known. Suppose this is 50Ω ; then $Z_1 = 50(0.55 + j1.6) = 27.5 + j80 \Omega$, and $Z_2 = 50(0.55 - j0.3) = 27.5 - j15 \Omega$.

Admittance values may also be shown in the Smith chart. Defining normalized admittance by

$$y = \frac{1}{z} \quad (13.12.3)$$

and then, using Eq. (13.12.2), we have

$$y = \frac{1 - \Gamma}{1 + \Gamma} \quad (13.12.4)$$

Noting that $-\Gamma = |\Gamma|/180^\circ + \phi$, Eq. (13.12.4) can be written as

$$y = \frac{1 + |\Gamma|/180^\circ + \phi}{1 - |\Gamma|/180^\circ + \phi} \quad (13.12.5)$$

This is similar to the relationship (Fig. 13.12.2) on which the chart is based, except that the angle coordinate must be read as $(180^\circ + \phi)$, where ϕ is the phase angle for the voltage coefficient. Normalized admittance may be written in terms of normalized conductance g and normalized susceptance b as $y = g + jb$, so the same scales as used for r , x , $|\Gamma|$, and ϕ may also be used for g , b , $|\Gamma|$, and $(180^\circ + \phi)$.

In Fig. 13.12.1(b) the point $y_3 = 0.55 + j1.6$ is shown. The corresponding voltage reflection coefficient values are $|\Gamma| = 0.746$ and $\phi_3 = -120^\circ$. The angle relationship shown on the chart is $180^\circ + \phi_3 = 60^\circ$. Another example shown is $y_4 = 1 - j2$. The corresponding voltage reflection coefficient is $0.707 \angle -225^\circ$. This can also be expressed as $0.707 \angle 135^\circ$. These are normalized values where $y = Y/Y_0$ and $Y_0 = 1/Z_0$. Again, assuming a value of $Z_0 = 50 \Omega$, the actual admittance for y_3 is $Y_3 = Y_0 y_3 = (0.55 + j1.6)/50 = 1.1 + j32 \text{ mS}$. The equivalent parallel components of resistance and reactance are

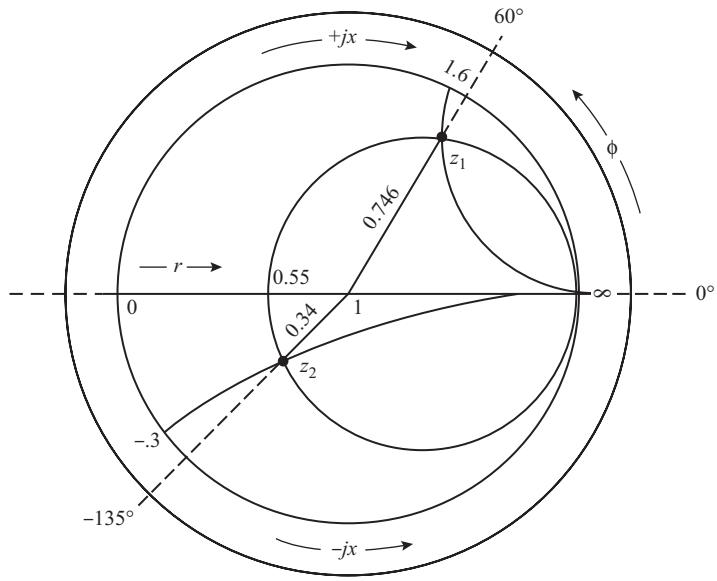
$$R_p = \frac{1}{1.1 \times 10^{-3}} = 909 \Omega$$

$$X_p = -\frac{1}{32 \times 10^{-3}} = -31.25 \Omega$$

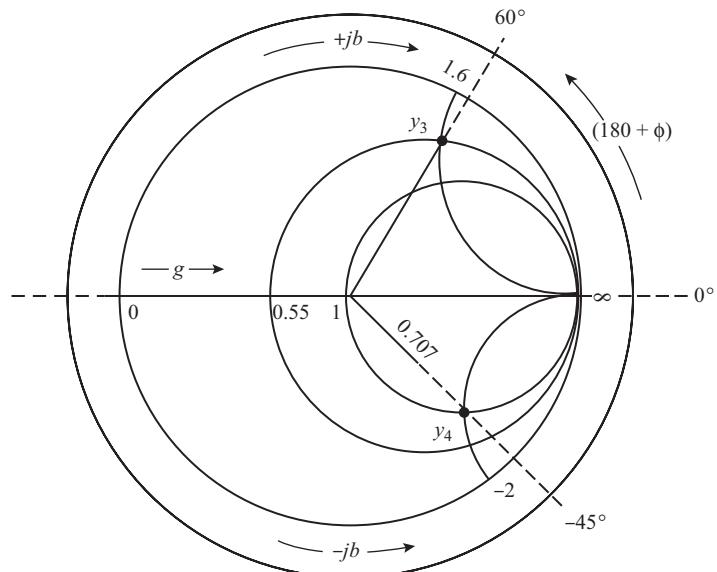
The negative sign shows that the reactance is capacitive.

Two important points on the chart are the impedance at a voltage maximum and the impedance at a voltage minimum. At a voltage maximum the phase angle of the voltage reflection coefficient is zero; hence

$$\Gamma_{\max} = |\Gamma_L|/0^\circ = |\Gamma_L|$$



(a)



(b)

Figure 13.12.1 (a) Impedance points z_1 and z_2 on the Smith chart. (b) Admittance points y_3 and y_4 on the Smith chart.

Therefore,

$$\begin{aligned} z_{\max} &= \frac{1 + |\Gamma_L|}{1 - |\Gamma_L|} \\ &= \text{VSWR} \end{aligned} \quad (13.12.6)$$

This last relationship is obtained from Eq. (13.9.5). Thus, the normalized impedance at a voltage maximum is seen to be purely resistive and *equal* in magnitude to VSWR and so must lie on the r -axis as shown in Fig. 13.12.2(a).

At a voltage minimum, the phase angle of the voltage reflection coefficient is 180° ; and therefore,

$$\begin{aligned} z_{\min} &= \frac{1 - |\Gamma_L|}{1 + |\Gamma_L|} \\ &= \frac{1}{\text{VSWR}} \end{aligned} \quad (13.12.7)$$

The normalized impedance at a voltage minimum, therefore, is also purely resistive, but is equal in magnitude to $1/\text{VSWR}$. This point must also lie on the r -axis, as shown in Fig. 13.12.2(a). The Smith chart axes are arranged so that a circle centered on $r = 1$ passes through both of these points. The VSWR value for such a circle *applies for every point on the circle*. For example, the normalized impedance of $0.4 + j0.75$ results in a VSWR of 4, as shown in Fig. 13.12.2(b). Care must always be taken not to confuse the “ r -circles” with the “VSWR-circles.” The VSWR circle in Fig. 13.12.2(b) is shown dashed.

Another use to which the VSWR circle is put is in finding the admittance corresponding to a given impedance value, and vice versa. The normalized admittance y is found diametrically opposite z on the VSWR circle, as shown in Fig. 13.12.2(b). For example, for $z = 1.4 + j1.7$ in the figure, y would be found as $y = 0.29 - j0.35$. The impedance–admittance transformation provided on the Smith chart need not be associated with transmission lines, and in fact the chart provides a graphical means of solving the series–parallel equivalence of circuits. Some specific examples will be shown shortly.

The Smith chart is used with slotted line measurements to determine impedance and admittance. A distance scale is added to the chart, the zero reference for this scale being the 180° phase angle point. As shown by Eq. (13.10.3), the voltage reflection coefficient phase angle is given by $\phi_L - 4\pi\ell/\lambda$. The normalized distance is ℓ/λ , and as ℓ increases, the phase angle decreases, so that from the 180° reference point, the decreasing phase angle scale can also be calibrated in *wavelengths toward generator*. Likewise, for ℓ decreasing, that is, going toward load from a voltage minimum, the corresponding increasing scale for phase angle can be calibrated in *wavelengths toward load*. These scales are shown in Fig. 13.12.3(b).

The total circumferential length of the chart is limited to 0.5 since the standing-wave pattern repeats itself at this interval.

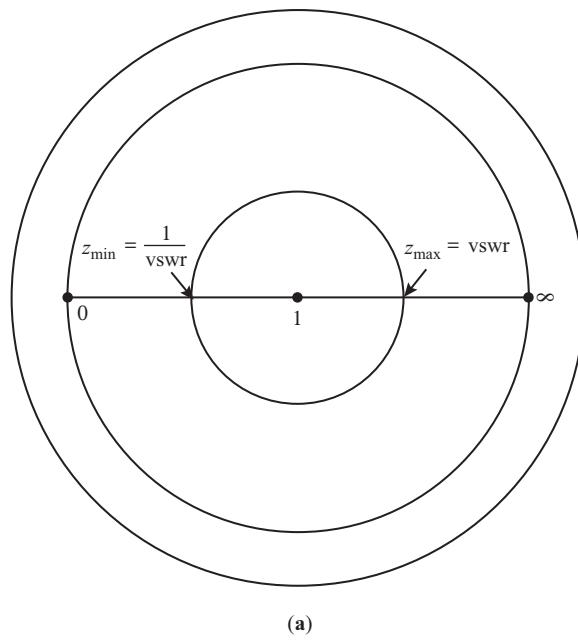
EXAMPLE 13.12.1

Rework Example 13.10 using the Smith chart. Data from Example 13.10.1 are $\text{VSWR} = 2:1$, $\ell_{\min} = 0.2$, and $Z_0 = 50\Omega$.

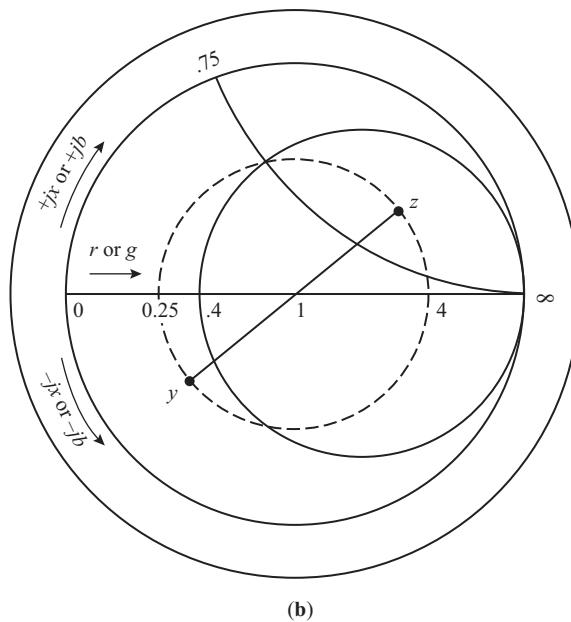
SOLUTION Draw the $\text{VSWR} = 2$ circle to cut the r -axis at 2 and 0.5 as shown in Fig. 13.12.4. Move along the *wavelengths-toward-load* scale a distance of 0.2, and draw a straight line from this point to the chart center. Where this line cuts the VSWR circle, point z_L in Fig. 13.12.4 gives the load impedance.

From the chart,

$$z_L = 1.55 - j0.7$$



(a)



(b)

Figure 13.12.2 (a) Impedance points corresponding to voltage maxima and minima, (b) A VSWR = 4 circle shown dotted, and the impedance-admittance z - y transformation on this.

Thus

$$\begin{aligned} Z_L &= 50(1.55 - j0.7) \\ &= 77.5 - j35\Omega \end{aligned}$$

This compares with $77.8 - j34.3 \Omega$ obtained by calculation in Example 13.10.1.

To find the equivalent admittance, move diametrically opposite from z_L to y_L on the VSWR circle, as shown in Fig. 13.10.1. From the chart,

$$y_L = 0.54 + j0.24$$

Therefore,

$$\begin{aligned} Y_L &= \frac{0.54 + j0.24}{50} \\ &= 10.8 + j4.8 \text{ mS} \end{aligned}$$

This compares with $10.76 + j4.75 \text{ mS}$ obtained by calculation in Example 13.10.1.

This example shows that to find the load impedance point, given the position of a voltage minimum, it is necessary only to move the distance ℓ_{\min}/λ toward load on the chart, and the z_L point is located where the radius line cuts the VSWR circle. It also shows how much less work is involved in using the Smith chart to solve the problem compared to calculation.

EXAMPLE 13.12.2

Rework Example 13.11.1 using the Smith chart. Data from Example 13.11.1 are $Z_0 = 50 \Omega$, $\ell/\lambda = 0.1$, and the load is a short circuit.

SOLUTION See Fig. 13.12.4. With a short-circuited load, the $\text{VSWR} = \infty$, and therefore the VSWR circle coincides with the normalized reactance (or susceptance) scale. Part of the scale is shown in heavy outline in Fig. 13.12.4. The problem is to find the input reactance at 0.1λ toward the generator. Therefore, moving along the VSWR circle a distance 0.1 toward the generator gives a normalized input impedance of

$$z_{in} = j0.725$$

Thus

$$\begin{aligned} Z_{in} &= 50 \times j0.725 \\ &= j36.25\Omega \end{aligned}$$

This compares with $j36.33 \Omega$ obtained by calculation in Example 13.11.1.

It is important to observe that *clockwise movement from any position on the chart gives a shift in position toward generator, while counterclockwise movement gives a shift in position toward load*. It is not necessary that such shifts take place from the zero origin on the wavelengths scale.

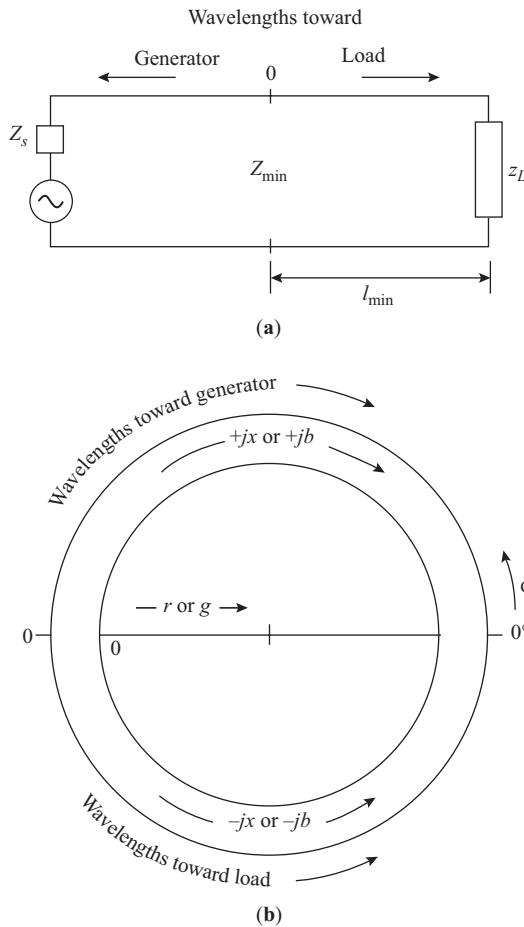


Figure 13.12.3 (a) Line lengths from a voltage minimum. (b) Normalized distance scale on the Smith chart.

Reflectionless matching of transmission lines can be achieved by using a reactive stub to tune out the reactance or susceptance at the correct position on the line. The position of the stub, its length, its characteristic impedance, and whether it should be open or short-circuited are all parameters that must be taken into consideration. In the method to be described here, the characteristic impedance of the stub is taken to be the same as that of the main feeder to be matched. Where the stub is connected in series with the main line, it is best to work with impedances and reactances. Where the stub is connected in parallel, as shown in Fig. 13.12.5(a), which is the most convenient arrangement for coaxial lines, it is best to work with admittances and susceptances. The problem to be solved is that of finding the position ℓ_1 , the length ℓ_2 , and whether the stub should be open or short-circuited. This problem is easily solved using the Smith chart. In terms of normalized admittances, the length ℓ_1 must transform the load admittance y_L into an admittance $y = 1 + jb$. The stub must add a susceptance $-jb$ so that the effective load admittance as seen by the main feeder is $(1 + jb) - jb = 1$. The actual admittance at this point is therefore $Y_0 \times 1 = Y_0$; that is, the line is matched.

Referring to Fig. 13.12.5(b), the first step on the Smith chart is to enter the point z_L , draw the VSWR circle, and locate the y_L point diametrically opposite. Of course, if y_L is known, it may be entered directly on the chart, and the VSWR circle may be drawn. The second step is to move from y_L in the direction

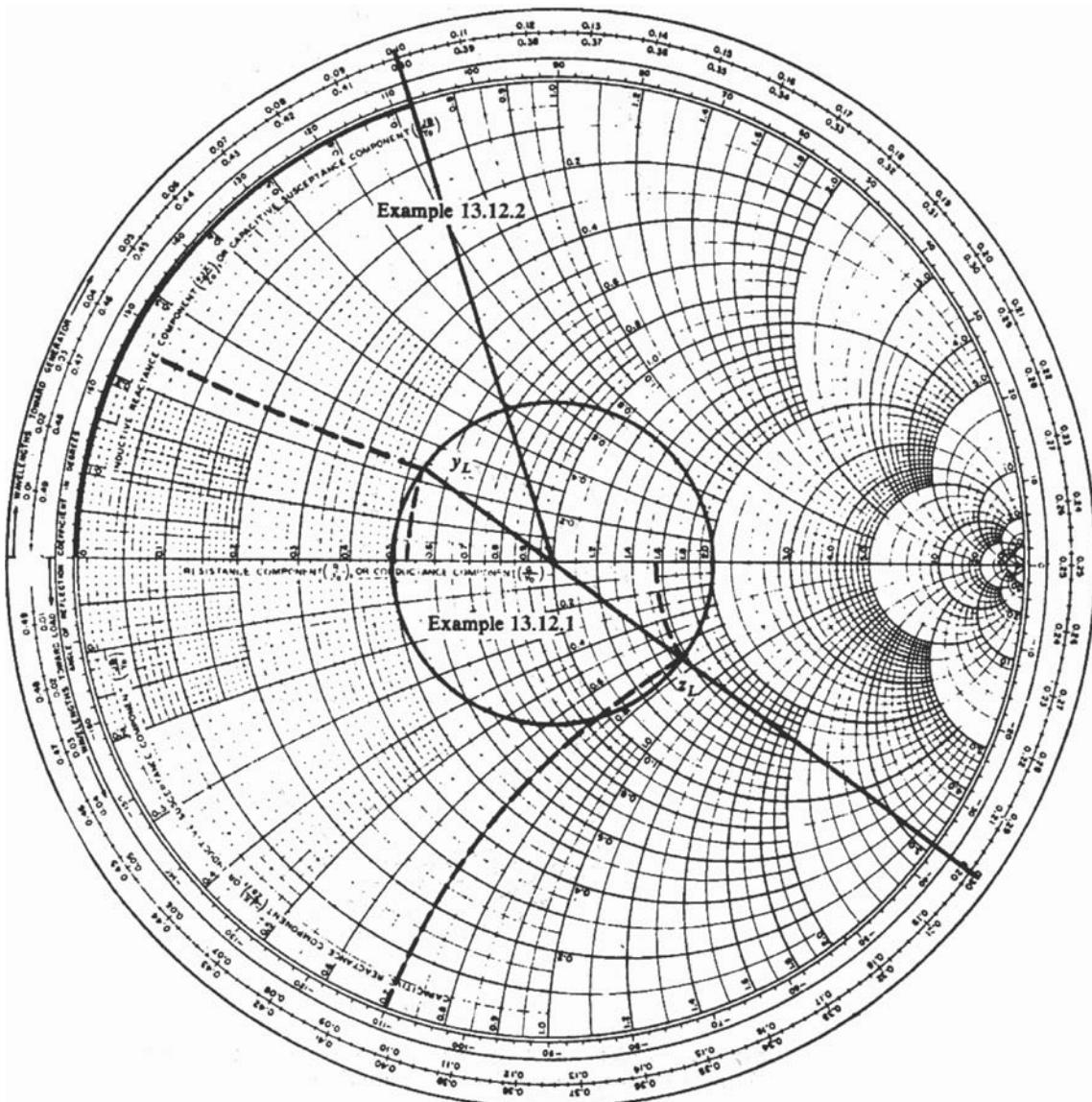


Figure 13.12.4 Smith chart solutions for Examples 13.12.1 and 13.12.2.

wavelengths toward generator, to the point where the VSWR circle cuts the $g=1$ circle. This gives the length ℓ_1 , as shown on Fig. 13.12.5(b), and the stub susceptance, jb . The third step is to go to the $-jb$ point on the chart and move in the direction wavelengths toward load, keeping in mind that it is the stub load that is referred to. The load point is reached when either zero susceptance or infinite susceptance is reached. Zero susceptance means that the stub load must be an open circuit, and infinite susceptance means that it must be

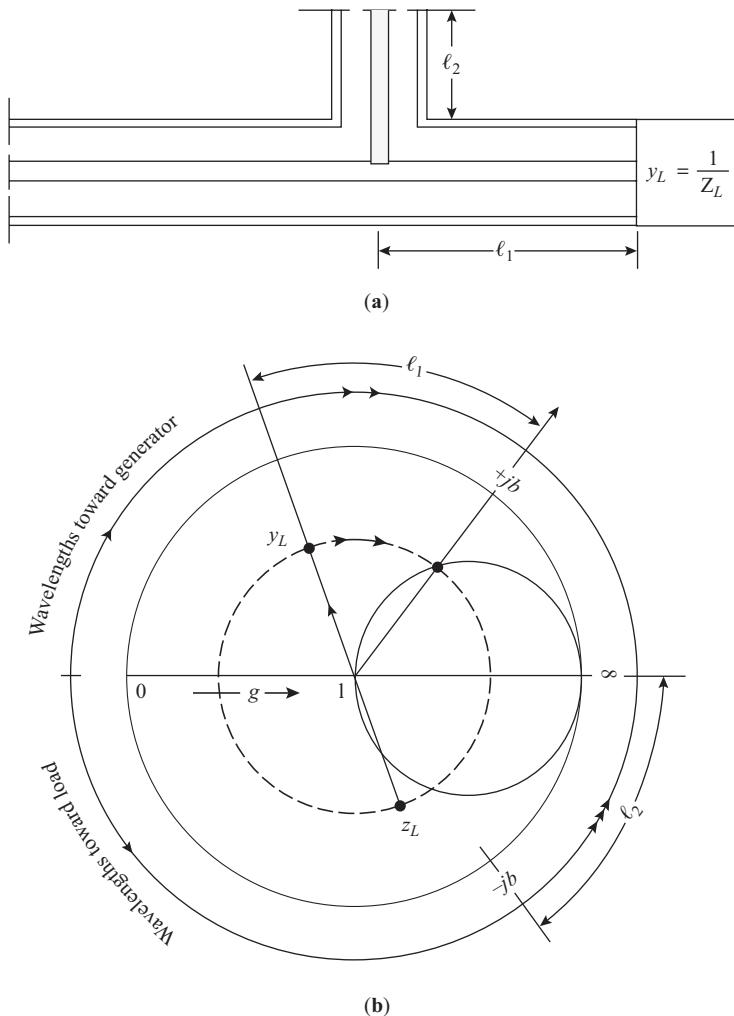


Figure 13.12.5 (a) Single stub matching. (b) The steps required on a Smith chart to determine the lengths of ℓ_1 and ℓ_2 .

a short circuit. For the conditions shown in Fig. 13.12.5(b), the infinite susceptance is reached first. The length ℓ_2 is determined as shown, and the stub load required in this case is a short circuit.

Single stub matching suffers from the disadvantage that the position ℓ_1 cannot easily be changed, and therefore the method is really suitable only for a fixed load. Where various loads may be encountered, double stub matching is often used, one such arrangement being shown in Fig. 13.12.6(a). As with single stub matching, the objective is to transfer the load point onto the $g = 1$ circle and then tune out the susceptance using stub ℓ_2 . In the arrangement shown, an arbitrary length ℓ of line transforms the load admittance from y_L to y'_L , this being a fixed transformation as shown in Fig. 13.12.6(b). Stub length ℓ_1 is now adjusted to bring y'_L onto the $\frac{1}{8}\lambda$ -displaced $g = 1$ circle at point y_1 . To achieve this, the length ℓ_1 is adjusted so that the

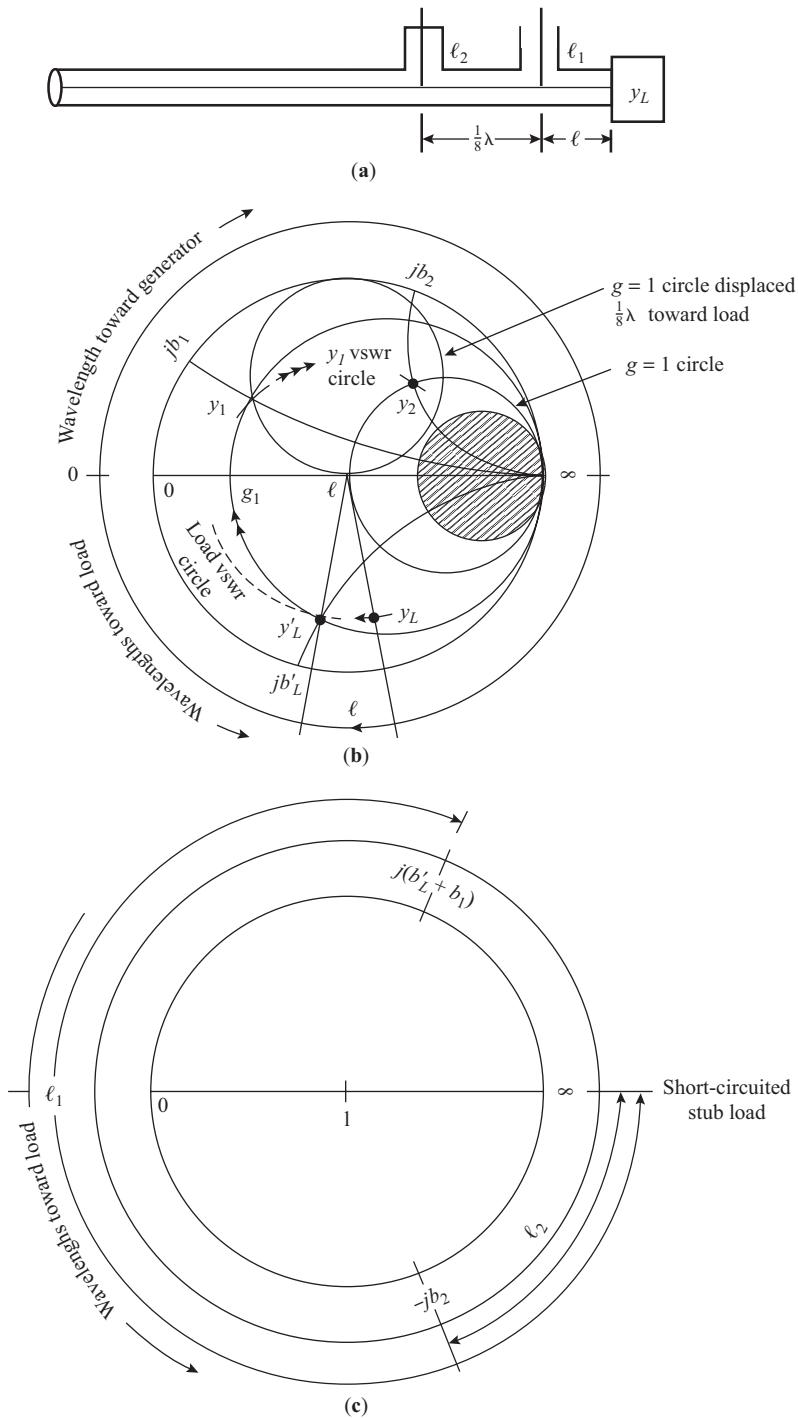


Figure 13.12.6 (a) Double-stub matching; (b) admittance transformations; (c) short-circuit stub lengths.

stub adds susceptance $j(b'L + b_1)$ in parallel with the existing $-jb'L$. The conductance remains constant at g_1 . At this stage, there is no movement along the main line, the stub ℓ_1 being adjusted to move the total susceptance from $-jb'L$ to $+jb'_1$. Any point on the $\frac{1}{8}\lambda$ -displaced $g=1$ circle is automatically transferred to the $g=1$ circle on moving the $\frac{1}{8}\lambda$ distance from stub 1 to stub 2 in the *toward generator* direction. Thus point y_1 is transformed into point y_2 . Length ℓ_2 is now adjusted to tune out the susceptance $+jb_2$. The required stub lengths are shown on Fig. 13.12.6(c). Since both stubs are shown short circuited, the stub load admittance in each case is infinite. Stub 1 has to provide a positive susceptance $+j(b'L + b_1)$ for the example shown, and the displacement must be *toward (stub) load*, which is infinite admittance. In this example, ℓ_1 is seen to be greater than 0.25λ , since half the circumference of the chart is 0.25λ . Stub 2 has to provide a susceptance $-jb_2$ from the short circuit, and so ℓ_2 is found as shown on Fig. 13.12.6(c).

Since stub lengths ℓ_1 and ℓ_2 are both adjustable, the system can be readjusted for different loads. One disadvantage is that if the load point y'_L should fall within the circle shown shaded, it could not be brought onto the $\frac{1}{8}\lambda$ -displaced $g=1$ circle, and so points within the shaded circle cannot be matched. For the $\frac{1}{8}\lambda$ spacing, examination of the Smith chart shows that the forbidden region is enclosed by the $g=2$ circle.

13.13 Time-domain Reflectometry

The basic arrangement for a *time-domain reflectometry* (TDR) system is shown in Fig. 12.12.1(a). The method utilizes direct measurement of the incident wave and the reflected wave as functions of time, hence the name “time-domain reflectometry.” The incident wave is supplied by a step generator, the step voltage having a very fast rise time. The oscilloscope records the outgoing step V_i and, at a somewhat later time, the reflected voltage V_r . The time between V_i and V_r is a measure of the length of the cable.

The nature of the load shows up in the shape of the reflected wave. As shown in Section 13.8, for an open circuit $\Gamma_L = 1$, the incident wave is totally reflected at the load. Therefore, assuming a lossless line, the oscilloscope trace will be as shown in Fig. 13.12.1(b). A short circuit results in a $\Gamma_L = -1$, so the incident step is totally reflected with a phase reversal, and therefore the oscilloscope trace is as shown in Fig. 13.12.1(c).

A purely resistive load of value $2Z_0$ results in a reflection coefficient of

$$\begin{aligned}\Gamma_L &= \frac{2Z_0 - Z_0}{2Z_0 + Z_0} \\ &= \frac{1}{3}\end{aligned}\tag{13.13.1}$$

That is,

$$V_R = \frac{V_I}{3}\tag{13.13.2}$$

The trace will therefore be as shown in Fig. 13.13.1(d).

An example of a complex load is the parallel *RC* circuit shown in Fig. 13.13.1(e). Initially, the capacitor behaves as a short circuit (assuming it is initially uncharged). However, C will immediately start to charge up, the voltage following an exponential law. When C is fully charged, it can be considered an open circuit, so the line is effectively terminated in R. The reflection coefficient for the final condition is therefore

$$\Gamma_L = \frac{R - Z_0}{R + Z_0}\tag{13.13.3}$$

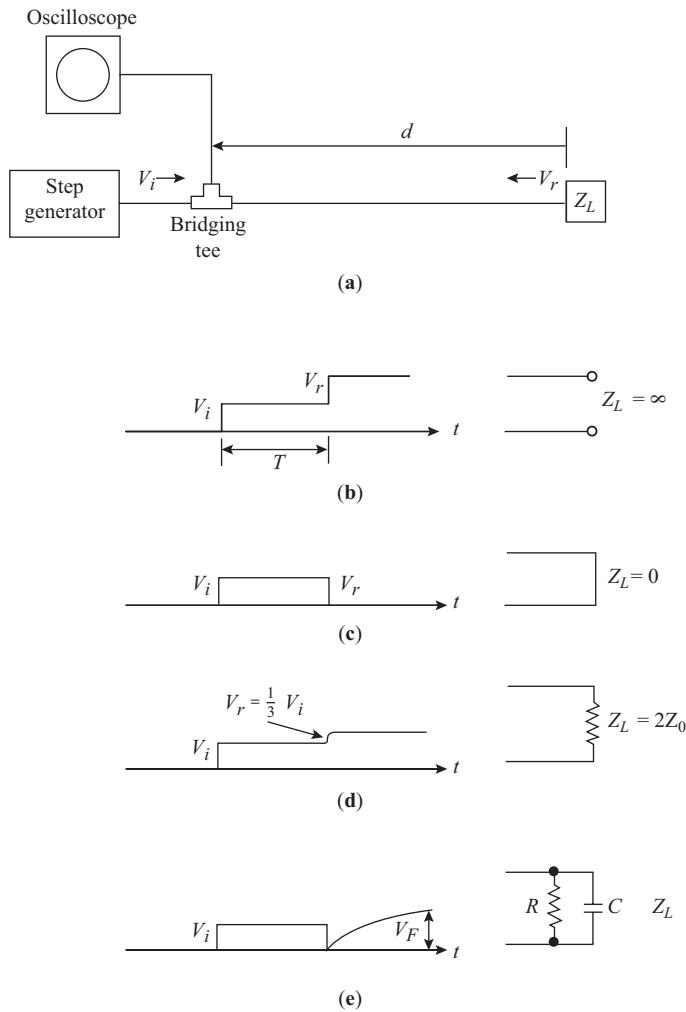


Figure 13.13.1 (a) Time-domain reflectometry system. Displays for (b) open-circuited termination, (c) short circuit, (d) resistive load $= 2Z_0$, and (e) parallel RC load.

and the final voltage V_F is

$$\begin{aligned}
 V_F &= V_I + V_R \\
 &= V_I(1 + \Gamma_L) \\
 &= V_I \left(1 + \frac{R - Z_0}{R + Z_0} \right)
 \end{aligned} \tag{13.13.4}$$

The oscilloscope trace is shown in Fig. 13.13.1(e).

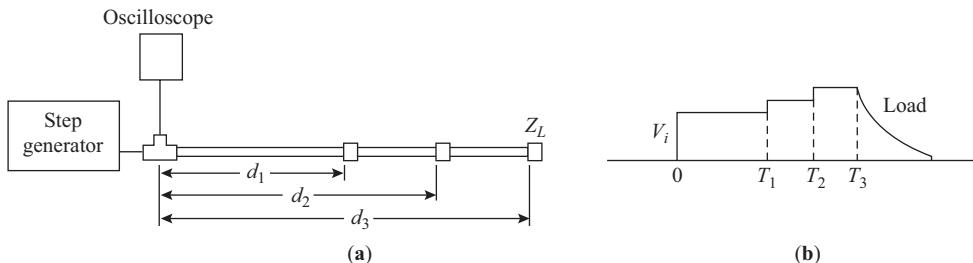


Figure 13.13.2 (a) TDR system used to determine discontinuities; (b) TDR display.

Knowing the elapsed time T between the incident and reflected voltages and the phase velocity on the cable, the cable length between the bridging tee and the point of reflection can be found. Thus d [Fig. 13.13.1(a)] is given by

$$d = v_p \frac{T}{2} \quad (13.13.5)$$

T is the round-trip time; therefore, it is necessary to divide by 2 as shown.

This feature is particularly important in locating faults and sources of mismatch on a cable, which show up as reflections. The cable shown in Fig. 13.13.2(a), for example, may result in a trace as shown in Fig. 13.13.2(b). The phase velocity can be found by measuring T for a known length of cable. The various distances [Fig. 13.13.2(a)] are then

$$d_1 = v_p \frac{T_1}{2} \quad (13.13.6)$$

$$d_2 = v_p \frac{T_2}{2} \quad (13.13.7)$$

$$d_3 = v_p \frac{T_3}{2} \quad (13.13.8)$$

TDR is a very versatile system of measurement for determining transmission-line properties, and fuller details will be found in the *Application Notes* 62 and 67 issued by the Hewlett-Packard Company.

13.14 Telephone Lines and Cables

The simplest type of line (apart from the single overhead wire with ground return) is the two-wire overhead line. The wire is usually made of cadmium copper, which provides better mechanical strength than hard-drawn copper. Bare wires are used where possible, but where insulation or protection is required, a polyvinyl chloride (PVC) covering is used.

Power lines (60 Hz) can induce interference in open-wire lines, but this can be eliminated by *transposing* the telephone wires. Figure 13.14.1(a) shows how interference may be picked up through both inductive and capacitive coupling. The induced interference voltages E_1 and E_2 act in opposition, but because they are not in general equal in magnitude, the resultant interference voltage $E_1 - E_2$ will be different from zero.

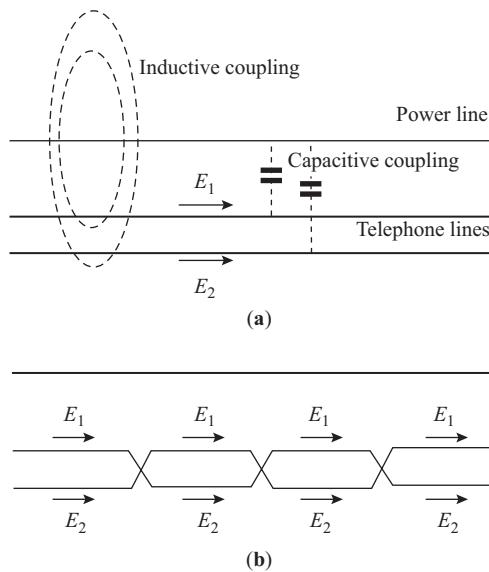


Figure 13.14.1 (a) Induced interference; (b) line transposition to eliminate interference.

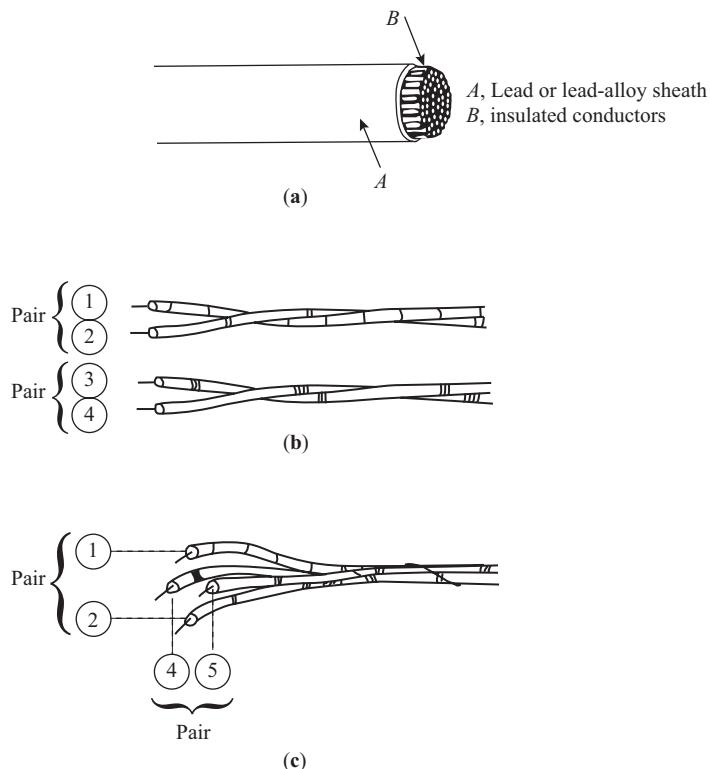


Figure 13.14.2 (a) Lead-sheathed multicore cable; (b) twin-cable assembly; (c) star-quad cable assembly. (Courtesy BICC Telephone Cables Div.)

By transposing the wires as shown in Fig. 13.14.1(b), the induced voltages in each wire can be equalized so that they cancel.

Where open-wire lines are not feasible, the wires are formed into a multicore cable assembly, which is usually carried underground in ducting. For this type of cable the wires are annealed copper, which is flexible, and each wire is insulated with high-grade paper tape. A lead-alloy sheath is extruded over the wires to form the cable as shown in Fig. 13.14.2(a).

For larger cable assemblies, for example those required between exchanges, the wires of a given pair are either twisted together in pairs to form a *twin cable* [Fig. 13.14.2(b)], or they may be twisted to form a *star quad* [Fig. 13.14.2(c)]. These are then assembled in larger units either as a concentric assembly or in unit construction. The latter is favored for use in local distribution networks, as branching into smaller units is then conveniently and tidily achieved.

For higher-frequency signals (for example, multicarrier telephony or television), coaxial cables are normally employed, the constructional features of which are shown in Fig. 13.14.3(a). These, in turn, may form part of a composite cable assembly as shown in Fig. 13.14.3(b).

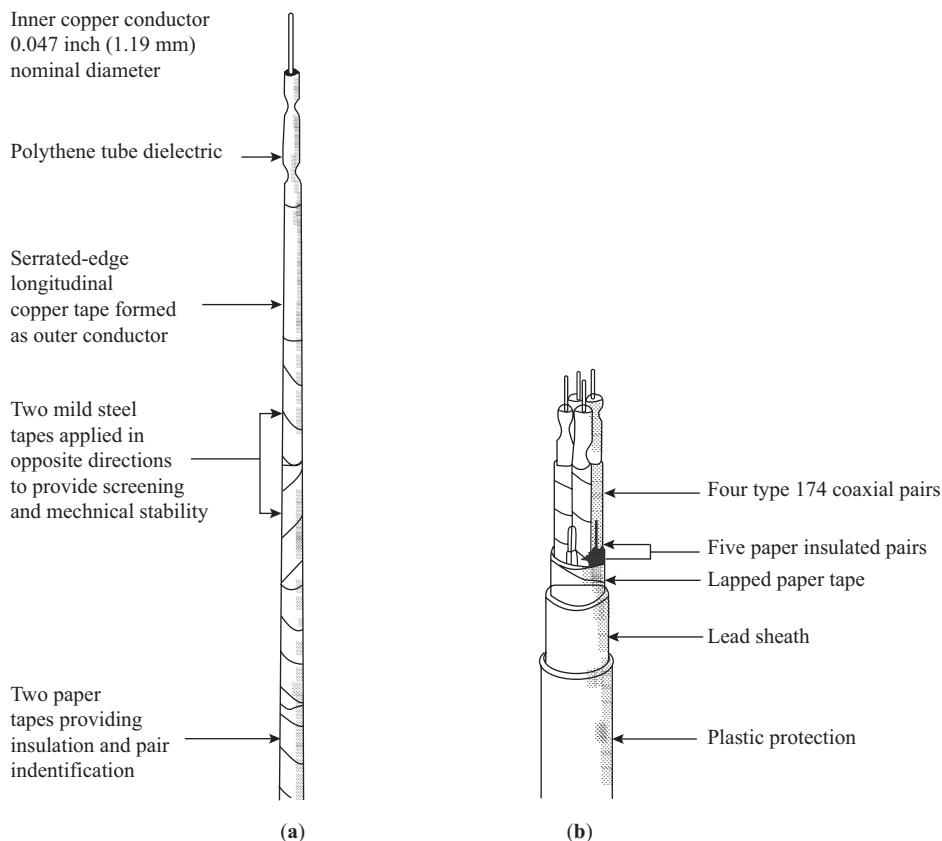


Figure 13.14.3 (a) Coaxial cable; (b) Composite cable assembly. (Courtesy BICC Telephone Cables Div.)

13.15 Radio-frequency Lines

For low-power applications, lines with solid dielectrics, such as the twin feeder or the coaxial line, are normally used. With high-power lines (such as would be used to feed radio-frequency power to antennas in the kilowatt range), the solid dielectric is omitted to keep losses at a minimum, and an open-type construction is used with well-spaced support insulators, so that the main dielectric is air.

13.16 Microstrip Transmission Lines

Microwave integrated circuit assemblies utilize special forms of transmission lines, the two most common types being the microstrip line and the stripline. The basic structures for both types, along with their electromagnetic field configurations, are shown in Fig. 13.16.1. Because the microstrip assembly is open on the top, the field distribution tends to be complex, and this affects the characteristic impedance of the lines. However, the open structure is easier to fabricate, and discrete components are readily added to the circuit. With the stripline, the fields are confined to the dielectric region, and this is more like a distorted version of the coaxial line field distribution shown in Fig. 13.1.1.

Although the conductor patterns for a given circuit configuration may be similar for both systems, the constructional details are very different. A typical substrate material for microstrip is Alumina, which is physically hard and can withstand the high temperatures and high vacuum conditions encountered during the deposition of the conductors. Both thick-film and thin-film deposition methods are used in practice. The nominal

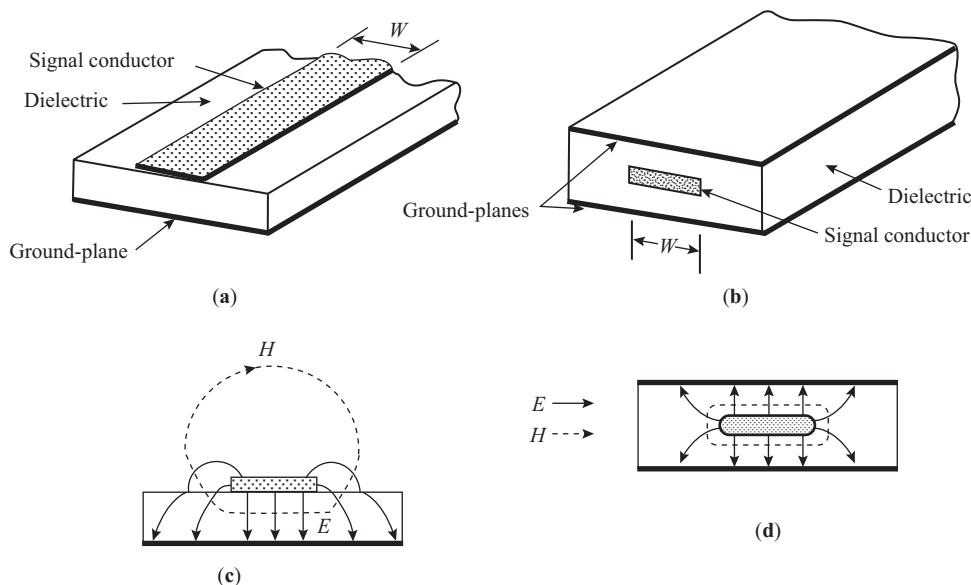


Figure 13.16.1 (a) Microstrip; (b) stripline; (c) electric field E and magnetic field H for microstrip; (d) E and H for stripline.

relative permittivity of Alumina is 9. An open view of a microstrip assembly is shown in Fig. 13.16.2, along with the equivalent circuit.

Stripline circuits are made using printed-circuit-board methods. However, special board material having low loss at microwave frequencies and very good mechanical stability must be used. Various board materials are available such as woven fiber glass and polyolefin, the relative permittivity for the latter being 2.32. Boards may come with copper cladding on one or both sides. The cladding is specified in ounces, which refers to the number of ounces of copper used for each square yard of surface. Common weights are 1 oz and 2 oz, corresponding to copper thicknesses of 0.0014 in. and 0.0028 in., respectively. Figure 13.16.3 shows an exploded view of a stripline assembly. Two printed circuit boards are used, each having a ground plane. The lower surface on the upper board is left blank, while the conductor pattern is etched on the upper surface of the lower board. When ready for assembly, the two substrates are tightly and uniformly clamped together between metal casings. The residual air gap around the conductor pattern may be filled using a silicone grease or a polyethylene laminate, or in some cases the clamping pressure may be relied on to produce sufficient flow of the dielectric around the conductors.

The characteristic impedance of both types of line depends on the conductor line width, thickness, and separation from substrate, and also on the relative permittivity of the substrate. In the case of microstrip, an effective relative permittivity must be used that takes into account the combined substrate—air dielectric space. Design formulas and charts for determining characteristic impedance are available in specialist

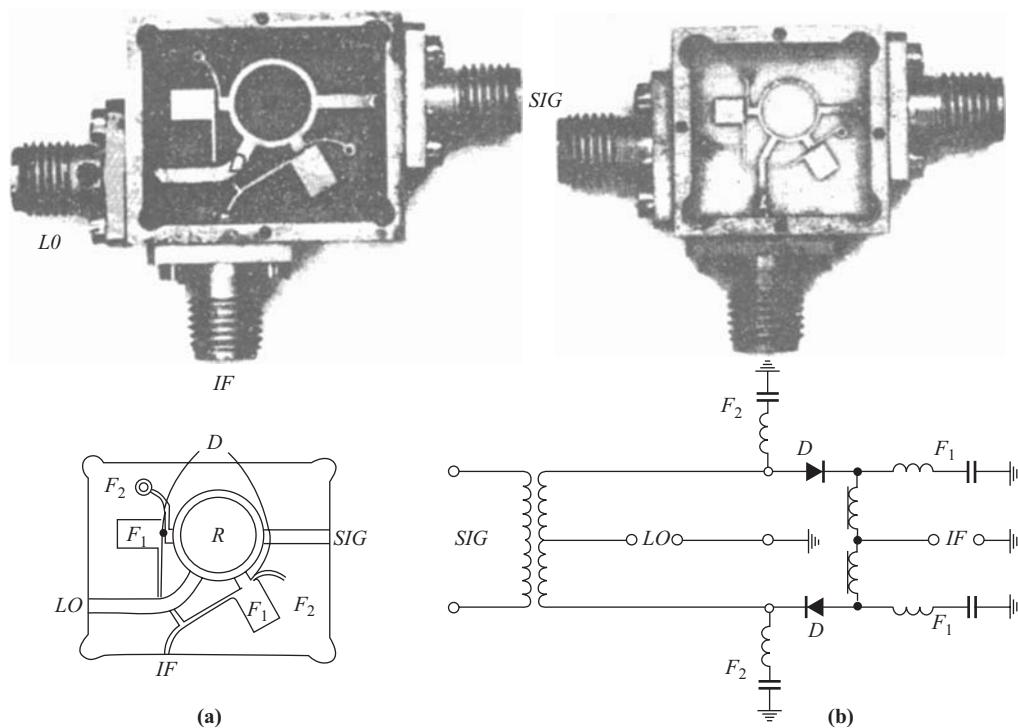


Figure 13.16.2 (a) Balanced mixer for 12 GHz: left, on quartz glass; right, on aluminum oxide. *R*, hybrid ring. *D*, Schottky-barrier diodes. *F₁*, band-stop filters for 12 GHz. *F₂*, band-stop filters for the second harmonic of the oscillator frequency. *LO*, oscillator input. *SIG*, aerial input. *IF*, intermediate-frequency output. (b) Low-frequency equivalent of the mixer circuit (a). (Courtesy J. H. C. van Heuven and A. G. van Nie, *Philips Tech. Rev* 32, 1971.)

publications dealing with the design of these circuits. Practical values of characteristic impedance range from about 5 to 110 Ω . The line impedance increases as the line width w decreases, because a narrower line results in an increased series inductance and decreased shunt capacitance, and, as shown by Eq. (13.4.3), both effects tend to increase Z_0 .

Matching networks similar to those shown in Figs. 13.12.5 and 13.12.6 can be designed using microstrip and stripline, and the method of using the Smith chart to solve the matching problem is equally applicable, but the values must be normalized to the characteristic impedance of the line being used. A main advantage of stripline and microstrip is that various line sections having different characteristic impedances can be readily constructed together in the same circuit, resulting in versatile matching networks and other types of circuits. Because of the mixture of characteristic impedances encountered in such situations, the Smith chart does not provide a convenient method of solution, but in fact the matching elements are easily calculated by direct means. Such networks make use of the circuit elements described in Section 13.11. Specifically, open- and short-circuited line lengths of $\frac{1}{8}\lambda$ and $\frac{3}{8}\lambda$ are commonly used, as these provide pure reactances. From Eq. (13.11.3), the impedance Z of a $\lambda/8$ short-circuited length of line is given by

$$Z = jZ_0 \tan\left(\frac{2\pi}{\lambda} \frac{\lambda}{8}\right) = jZ_0 \tan\left(\frac{\pi}{4}\right) = jZ_0 \quad (13.16.1)$$

This shows that the input impedance of such a stub is purely reactive and numerically equal to the characteristic impedance of the stub. In a similar manner, a length of $3\lambda/8$ results in

$$Z = -jZ_0 \quad (13.16.2)$$

Usually, it is more convenient to work with admittance values, which are simply the reciprocals of the impedances.

$$\frac{\lambda}{8} \text{ short-circuited: } Y = -jY_0 \quad (13.16.3)$$

$$\frac{3\lambda}{8} \text{ short-circuited: } Y = jY_0 \quad (13.16.4)$$

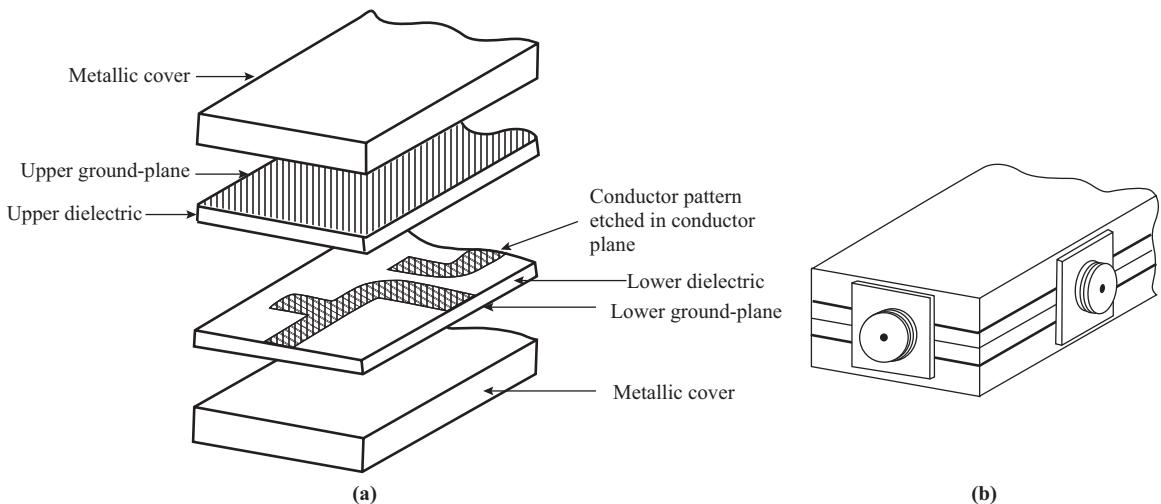


Figure 13.16.3 (a) Exploded view of a stripline assembly; (b) assembled stripline with edge connectors.

Open-circuited stubs may also be used, and the admittances for the two cases are readily derived from Eq. (13.11.8) as

$$\frac{\lambda}{8} \text{ open-circuited: } Y = jY_0 \quad (13.16.5)$$

$$\frac{3\lambda}{8} \text{ open-circuited: } Y = -jY_0 \quad (13.16.6)$$

Along with stub sections, the $\lambda/4$ transformer section is also used, Eq. (13.11.9) being applicable in this case. The use of stripline (and microstrip) for construction of a matching network is illustrated in the following example.

EXAMPLE 13.16.1

To optimize the low-noise performance of a microwave amplifier, the transistor must be fed from a source having an admittance $Y = 0.05 = j0.03$ S. The actual source impedance is 50Ω . Design a suitable stripline matching network.

SOLUTION The schematic for the matching network is shown in Fig. 13.16.4(a). Writing the required source admittance as $Y = G + jB$, the $\lambda/4$ section has to transform the 50Ω source into $1/G = 1/0.05 = 20\Omega$. Using Eq. (13.11.9),

$$20 = \frac{Z_0^2}{50}$$

Therefore,

$$Z_0 = \sqrt{20 \times 50} = 31.62 \Omega$$

The stub must add a susceptance of $jB = -j0.03$ siemens. Equation (13.16.3) shows that this can be accomplished through the use of a short-circuited stub of characteristic admittance Y_0 , where

$$-j0.03 = -jY_0$$

Thus

$$Z_0 = \frac{1}{Y_0} = 33.33 \Omega$$

In practice, to minimize the effect of the shunt to series connection, the shunt stub is usually balanced about the series line as shown in Fig. 13.16.4(b). This is equivalent to having two stubs in parallel, each of susceptance $-j0.15$ S and, therefore, each having a characteristic impedance of 66.66Ω . Capacitive short circuits are used at the ends of these stubs.

13.17 Use of Mathcad in Transmission Line Calculations

Computer program packages such as Mathcad offer a very powerful means of solving equations. One advantage of using such programs is that they eliminate much of the algebraic manipulation needed in order to reduce equations to explicit forms for solving by manual methods. Computer solutions often made graphical

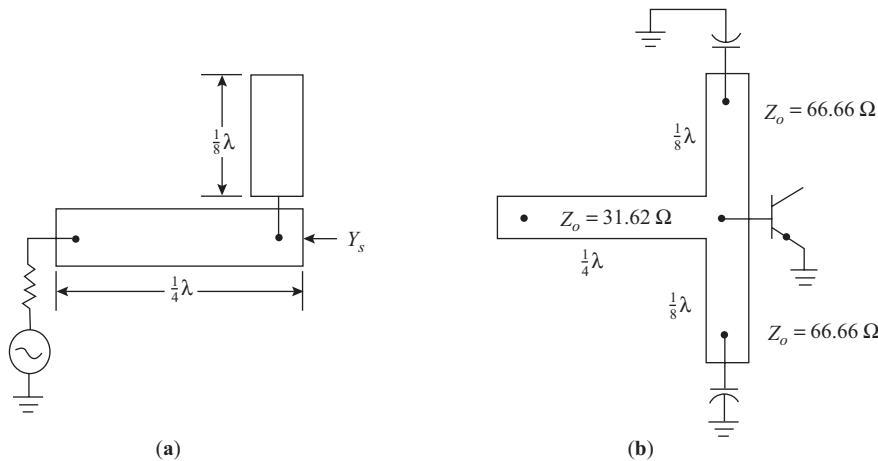


Figure 13.16.4 (a) Matching network schematic for Example 13.16.1; (b) stripline layout.

solutions such as the use of the Smith chart redundant. Graphical methods do, however, have certain advantages of their own. Significant practical trends can often be identified very quickly from a graphical presentation, which might otherwise be difficult to spot, and which makes graphical methods the preferred choice by the experienced user. Of course, we have to be careful not to obscure the significance of the physical processes occurring by relying too much on computer solutions, and until we are thoroughly familiar with the underlying theory of a topic, the computer solutions should embody enough detail so that the practical aspects of the solutions can be followed. In other words, although it might be possible to produce computer solutions in a very compact form, at the start it is better to take the time and trouble to write the steps out in sufficient detail so that the process is understood also. Mathcad is highly suitable for this approach, and in this section three examples are presented to illustrate some of the strengths of the computer solutions.

As a first example, Example 13.10.1 will be reworked using Mathcad.

EXAMPLE 13.17.1

This is a repeat of Example 13.10.1 using Mathcad to solve the problem.

Given Data: VSWR := 2, $\ell_{\min} := 0.2$ $Z_0 := 50 \Omega$

Computations:

$$\phi_L := (4\Omega_{\min} - 1) \cdot \pi \quad \text{Eq. (13.10.5)}$$

$$\rho_L := \frac{\text{VSWR} - 1}{\text{VSWR} + 1} \quad \text{Eq. (13.9.7)}$$

$$\Gamma_L := \rho_L e^{j \cdot \phi_L} \quad \text{Eq. (13.10.2)}$$

$$Z_L := Z_0 \frac{1 + \Gamma_L}{1 - \Gamma_L} \quad \text{Eq. (13.10.6)}$$

$$Z_L = 77.7 - 34.3j \cdot \Omega$$

Thus, for the equivalent series components, the resistance is 77.7Ω , and the reactance is 34.3Ω capacitive, as indicated by the minus sign. The equivalent parallel component values are found as follows:

$$Y_L := \frac{1}{Z_L}$$

$$R_p := \frac{1}{Re(Y_L)}$$

$$X_p := \frac{-1}{Im(Y_L)}$$

$$R_p = 92.8 \Omega$$

$$X_p = 210.6 \Omega$$

This example was also solved using the Smith chart, as Example 13.12.1. Comparing the three methods, it is seen that the Mathcad solution consists of a sequence of fairly simple equations derived from basic definitions. It must be borne in mind that the derivations themselves involve some algebraic manipulation, but this is kept to a minimum.

As a second example, consider the situation where the load impedance is known, along with the characteristic impedance and the propagation coefficient of the line, and it is required to find the input of a given length of line.

EXAMPLE 13.17.2

A transmission line is terminated in a load impedance of $30 - j23 \Omega$. Determine the input impedance of a 50-cm length of line, given that the characteristic impedance is 50Ω and the wavelength on the line is 45 cm. The line may be assumed lossless.

SOLUTION Given data: $Z_L := (30 - j23) \Omega$ $Z_0 := 50 \Omega$
 $\lambda := 45 \text{ cm}$ $\ell := 50 \text{ cm}$

Computations:

$$\beta := \frac{2\pi}{\lambda} \quad \text{Eq. (13.5.5)}$$

$$\Gamma_L := \frac{Z_L - Z_0}{Z_L + Z_0} \quad \text{Eq. (13.8.13)}$$

To set up the basic voltage and current equations, a reference incident voltage of 1 V will be used.

$$V_I := 1V$$

$$V_R := V_I \Gamma_L \quad \text{Eq. (13.8.1)}$$

$$V_i := V_I e^{j\beta l} \quad \text{Eq. (13.8.2)}$$

$$V_r := V_I e^{-j\beta l} \quad \text{Eq. (13.8.3)}$$

$$V := V_i + V_r \quad \text{Eq. (13.8.4)}$$

$$I := \frac{V_i - V_r}{Z_0} \quad \text{Eq. (13.8.9)}$$

$$Z := \frac{V}{I}$$

$$Z = 23.5 + 5.1j\Omega$$

As a final example, Mathcad will be used to evaluate the voltage standing wave ratio (VSWR) as a function of frequency for a quarter-wave matching transformer. Solving this problem by conventional means would be very difficult, yet with Mathcad it still only requires the basic equations to be set up as functions of frequency.

EXAMPLE 13.17.3

A 600-Ω main line is matched to a 73-Ω load through a quarter-wave transformer. Plot the voltage standing, wave ratio on the main line, and determine the VSWR at a frequency of $0.9f_0$, where f_0 is the quarter-wave resonant frequency.

SOLUTION Given data: $Z_0 := 600 \Omega$, $Z_L := 73 \Omega$

The quarter-wave section is resonant at some frequency f_0 corresponding to a wavelength λ_0 , and to see how the VSWR varies with frequency a normalized variable $F = f/f_0$ will be used. It will also be noted that $F = \lambda_0/\lambda$. The product of propagation coefficient and quarter-wavelength is therefore $\beta * \lambda_0/4 = \pi F/2$. This will be denoted as $\theta(F)$.

$$F := 0.8, 0.81, \dots, 1.2, \quad \theta(F) := \frac{2\pi F}{4}$$

Computations: For matching, the effective load impedance on the main line must equal the characteristic impedance of the main line; thus

$$Z'_L := Z_0$$

$$Z'_0 := \sqrt{Z'_L Z_L}, \quad \text{from Eq. (13.11.9)}$$

$$\Gamma_L := \frac{Z_L - Z'_0}{Z_L + Z'_0}$$

As before, the incident voltage will be normalized to unity: $V_I := 1 V$. At the input to the quarter-wave section:

$$V_i(F) := V_I e^{j\theta(F)} \quad V_r(F) := \Gamma_L V_I e^{-j\theta(F)}$$

$$V_{in}(F) := V_i(F) + V_r(F) \quad I_{in}(F) := \frac{V_i(F) - V_r(F)}{Z'_0}$$

$$Z_{in}(F) := \frac{V_{in}(F)}{I_{in}(F)}$$

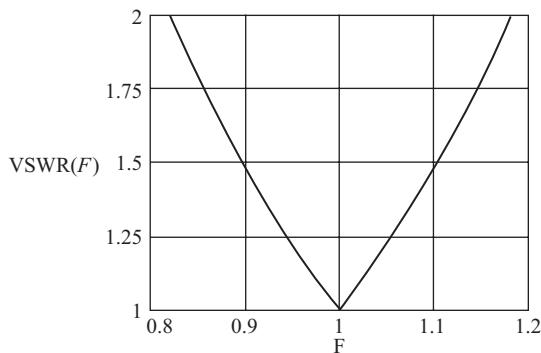


Figure 13.17.1 VSWR versus frequency plot for Example 13.17.3.

This input impedance forms the load on the $600\text{-}\Omega$ main line, and hence the voltage reflection coefficient for the main line is

$$\Gamma_L(F) := \frac{Z_{in}(F) - Z_0}{Z_{in}(F) + Z_0}$$

and the VSWR is

$$\text{VSWR}(F) := \frac{1 + |\Gamma_L(F)|}{1 - |\Gamma_L(F)|}$$

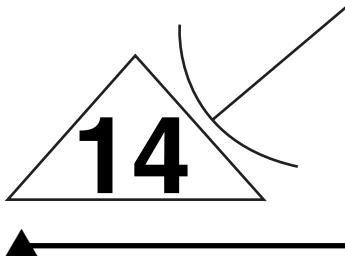
This can be plotted as shown in Fig. 13.17.1. The VSWR at $F = 0.9$ is obtained as
VSWR(0.9) = 1.479

PROBLEMS

- 13.1. Explain what is meant by the characteristic impedance of a transmission line. Explain why you would expect an infinitely long uniform line to have an input impedance equal to the characteristic impedance.
- 13.2. The primary constants (which may be assumed independent of frequency) for a transmission line are $R = 0.5 \text{ }\Omega/\text{m}$, $L = 250 \text{ nH/m}$, $C = 100 \text{ pF/m}$, and $G = 10^{-6} \text{ S/m}$. Plot the variation of $|Z_0|$ against frequency for this line.
- 13.3. For the transmission line in Problem 13.2, calculate the attenuation coefficient and the phase-shift coefficient at a frequency of 20 MHz.
- 13.4. Explain what is meant by the propagation coefficient of a transmission line. The propagation coefficient for a transmission line is given as $(0.0005 + j\pi/10)\text{m}^{-1}$. Determine, for a matched-line length of 50 m, (a) the total attenuation in nepers, (b) the total attenuation in decibels, and (c) the total phase shift.

- 13.5.** The propagation coefficient γ for a transmission line is given by $\gamma = 0.01 + j(\pi/10)$. The line is 32 m long and is correctly terminated to avoid reflections. Plot the phasor diagram showing the variation of current over the length of line.
- 13.6.** For the transmission line in Problem 13.4, determine the attenuation, in nepers, and in decibels, of 15 km of line.
- 13.7.** Explain what is meant by the *group delay time* and why this should be constant for a transmission line. Given that $\beta = \omega\sqrt{LC}$, $L = 1\mu\text{H/m}$, and $C = 11.11\text{ pF/m}$, calculate (a) the group delay time and (b) the group velocity.
- 13.8.** Explain what is meant by the *voltage reflection coefficient* for a transmission line. The voltage reflection coefficient on a $50\text{-}\Omega$ line is $0.7\angle 30^\circ$. Determine the load impedance terminating the line.
- 13.9.** A $300\text{-}\Omega$ line is terminated in a load impedance of $100 + j200\text{ }\Omega$. Determine the voltage reflection coefficient.
- 13.10.** Explain how standing waves may be set up on a transmission line, and state the relationship between voltage and current standing waves for a line whose characteristic impedance is purely resistive.
- 13.11.** Calculate the voltage standing-wave ratio for a lossless line of characteristic impedance $50\text{ }\Omega$ when it is terminated in (a) $100\text{ }\Omega$ resistive, and (b) $(30-j50)\text{ }\Omega$.
- 13.12.** Sketch the variation of input impedance with length for a lossless transmission line terminated in a short circuit. A section of $50\text{-}\Omega$ line, 0.15λ in length, is terminated in a short circuit. Determine the effective input reactance of the section, showing clearly whether it is inductive or capacitive.
- 13.13.** Calculate the shortest shorted stub length required to produce a reactance equal to (a) Z_0 and (b) $-Z_0$. Express the length as a fraction of a wavelength.
- 13.14.** A $600\text{-}\Omega$ lossless line is terminated in a 5-pF capacitor. The line is operated at a frequency that makes it $\frac{1}{4}\lambda$ long. Determine the nature of the input impedance of the line, giving the equivalent inductance or capacitance value.
- 13.15.** Determine, using a Smith chart, the admittance value corresponding to an impedance of $150 + j80\text{ }\Omega$. Give the admittance value in Siemens, and also give the equivalent parallel component values in ohms.
- 13.16.** A load impedance of $80 - j70\text{ }\Omega$ is connected to a $50\text{-}\Omega$ transmission line. Find (a) the VSWR, (b) the distance, in terms of wavelength, to the first voltage minimum, and (c) the impedance at this minimum.
- 13.17.** Measurements on a $50\text{-}\Omega$ slotted line gave the following results: $\text{VSWR} = 3.0$; distance from load to first voltage minimum = 0.4λ . Determine, using a Smith chart, the value of the load impedance.
- 13.18.** An impedance $100 - j50\text{ }\Omega$ is used to terminate a $50\text{-}\Omega$ lossless line. The line is 0.6λ long. Determine, using a Smith chart, (a) the VSWR, (b) the input impedance, and (c) the impedance at a voltage maximum.
- 13.19.** A load admittance of $4 - j10\text{ mS}$ is to be matched to a $50\text{-}\Omega$ line system using single stub matching. Determine the position for the stub to be connected in parallel with the main line, the length of the stub, and whether the stub is to be open or short circuited.
- 13.20.** Repeat Problem 13.19 for a series-connected stub.
- 13.21.** A double stub matching section is arranged as shown in Fig. 13.12.6(a) with the distance $\ell = 0.1\lambda$. Determine which of the following normalized admittances can be matched: (a) $0.6 + j1.2$, (b) $0.6 - j1.2$, (c) $1.1 + j1.3$, (d) $3 - j2$
- 13.22.** A load impedance of $75 - j80\text{ }\Omega$ is to be matched to a $50\text{-}\Omega$ line using the double stub section of Problem 13.21. Determine the required stub lengths in wavelengths.

- 13.23.** In a time-domain reflectometer measurement on a terminated transmission line, the reflected voltage was found to be a vertical downward step, of magnitude one-half the input step. If the line can be assumed lossless and has a characteristic impedance of 100Ω , determine the value of load impedance.
- 13.24.** Sketch the form of trace expected on a TDR measuring set when applied to a line terminated in a series RL circuit. Give equations for the asymptotic values on the trace.
- 13.25.** A transmission line is made of two parallel copper No. 18 AWG wires (1.024 mm in diameter) spaced 10 mm apart and embedded in a plastic spacer made of polyvinyl chloride (PVC). The wire has a resistance of $1.984 \Omega/\text{km}$, and the insulation has a leakage conductivity of $2.5 \times 10^{-10} \text{ S/km}$. Find the primary line constants. For PVC, $\epsilon_r = 3.3$. Find the characteristic impedance for the line.
- 13.26.** Repeat Problem 13.25 for a coaxial line made of the same wire and insulating material, with a copper sheath of inside diameter of 2 cm. Assume the leakage to have the same value.
- 13.27.** A microwave amplifier requires to be fed from a source impedance of $12 + j8 \Omega$. The actual source impedance is 50Ω . Determine the impedance values for a suitable stripline matching section consisting of a $\lambda/4$ transformer section and a parallel-connected $3\lambda/8$ stub, and state whether the stub should be open or short circuited.
- 13.28** Plot the *characteristic impedance* Z_0 of a two-wire transmission line as a function of inter-wire spacing D . Assume that 32AWG wire is used.
- 13.29** Plot the *characteristic impedance* Z_0 of a two-wire transmission line as a function of diameter of wires d . Assume that inter-wire spacing is 4mm.
- 13.30** Repeat the above exercises for coaxial cables.
- 13.31** Plot, using MATLAB, β versus ω for coaxial cables.
- 13.32** Plot the reactance Z against line length, l , using MATLAB, when $0 \leq l \leq \lambda$, of a short-circuited transmission line.
- 13.33** Repeat the above exercise for an open-circuited transmission line.



Waveguides

14.1 Introduction

At frequencies higher than about 3000 MHz, transmission of electromagnetic waves along lines and cables becomes difficult, mainly because of the losses that occur both in the solid dielectric needed to support the conductors, and in the conductors themselves. It is possible to transmit an electromagnetic wave down a metallic tube called a *waveguide*. The most common form of waveguide is rectangular in cross section, as shown in Fig. 14.1.1(a). Induced currents in the walls of the waveguide give rise to power losses, and to minimize these losses the waveguide wall resistance is made as low as possible. Because of skin effect, the currents tend to concentrate near the inner surface of the guide walls, and these are sometimes specially plated to reduce resistance.

Apart from determination of losses, the walls of a waveguide may be considered to be perfect conductors. Two important boundary conditions result, which determine the mode of propagation of an electromagnetic wave along a guide: (1) electric fields must terminate normally on the conductor, that is the tangential component of the electric field must be zero [Fig. 14.1.1(b)]; (2) magnetic fields must lie entirely tangentially along the wall surface, that is, the normal component of the magnetic field must be zero [Fig. 14.1.1(c)]. Knowing these boundary conditions provides a simple way of visualizing how the various modes of waveguide transmission occur, which will be discussed in the following sections.

Although the microwave frequency spectrum extends from about 300 MHz to 300 GHz, transmission lines can be utilized for the lower part of the range. Above about 3000 MHz, waveguides become necessary where large amounts of power have to be transmitted.

14.2 Rectangular Waveguides

The boundary conditions already referred to preclude the possibility of a waveguide supporting *transverse electromagnetic* (TEM) wave (see Appendix B) propagation, since the magnetic field is at right angles to the direction of propagation (along the axis of the waveguide) and therefore would have to terminate normally

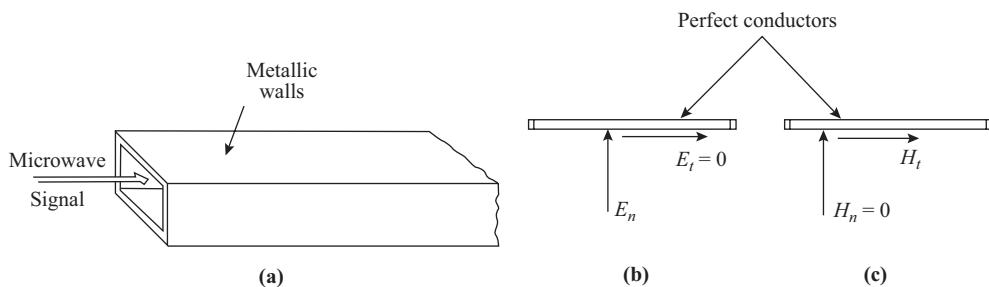


Figure 14.1.1 (a) Rectangular waveguide, (b) Electric field boundary conditions, (c) Magnetic field boundary conditions.

to the sidewalls, which cannot happen. One possible solution is for the magnetic field to form loops along the direction of propagation lying parallel to the top and bottom walls and tangentially to the sidewalls [see Fig. 14.2.1(a)].

The variation of electric field as a function of distance along the direction of propagation is shown in Fig. 14.2.1(b) and along the cross section in Fig. 14.2.1(c). The propagation mode sketched in Fig. 14.2.1 is known as a *transverse electric (TE) mode* because the electric field is entirely transverse to the direction of propagation. (It is also known as an *H mode*, signifying that part of the magnetic field lies along the direction of propagation.)

Subscripts are used to denote the number of half-cycles of variation that occur along the a and b sides. As shown in Fig. 14.2.1, one half-cycle (one maximum) occurs along the a side and none along the b side; this mode is therefore referred to as the TE_{10} mode. The TE_{10} mode is the dominant mode in waveguide transmission, for, as will be shown, it supports the lowest-frequency waveguide mode.

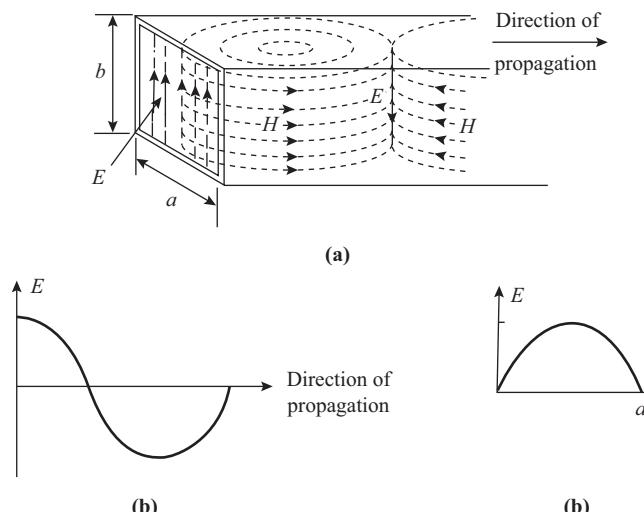


Figure 14.2.1 (a) One possible field configuration for a waveguide. (b) Electric field amplitude along guide axis. (c) Electric field amplitude along width.

Properties of the TE₁₀ Mode

Figure 14.2.2(a) shows how a TE₁₀ wave may be formed as the resultant of two TEM waves crossing each other. At the points of intersection shown, the individual waves add vectorially. At those points where the electric fields reinforce each other, shown by double crosses and double dots, the magnetic field is directly up and down; at the points where the electric fields cancel, shown by dots and crosses together, the magnetic field is directed left and right. Metallic walls may be placed along the L-R direction as shown in Fig. 14.2.2(b) without violating the boundary conditions, as it will be seen that the electric field is zero and the magnetic field tangential at these walls. Metallic top and bottom walls can be put in position, as the electric field will terminate normally on these and the magnetic field lies parallel to them.

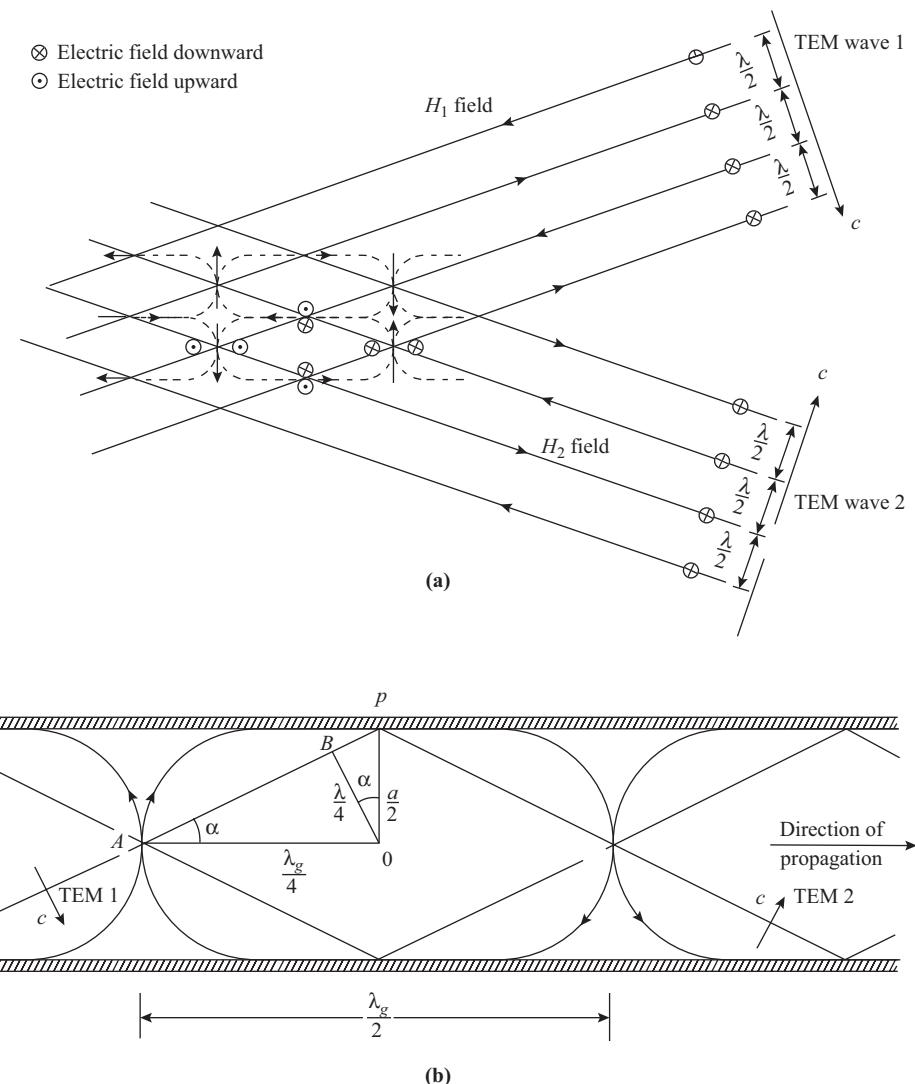


Figure 14.2.2 (a) Generating the waveguide mode from two TEM waves. (b) Geometry used in determining TE₁₀ properties.

The direction of the magnetic field loops alternates as shown in Fig. 14.2.2(b), one loop occupying a distance of a half wavelength in the guide. This will be different from a half wavelength of the individual TEM waves and is shown as $\frac{1}{2}\lambda_g$, where λ_g is termed the guide wavelength.

The frequency of the TE wave will be the same as that of the TEM waves (that is, it will have the same number of cycles in a second). Let v_p be the phase velocity of the TE wave; then, using the general relationship $\lambda_g = v_p$, derived in Appendix B, we have

$$\text{TE wave: } \lambda_g f = v_p \quad (14.2.1)$$

$$\begin{aligned} \text{TEM wave: } \lambda_f &= c \\ &\text{(in air)} \end{aligned} \quad (14.2.2)$$

from which it follows that

$$v_p = c \frac{\lambda_g}{\lambda} \quad (14.2.3)$$

From the geometry of Fig. 14.2.2(b), it can be seen that

$$\cos \alpha = \frac{\lambda}{2a} \quad (\text{from the right-angled triangle } OBP) \quad (14.2.4)$$

$$\sin \alpha = \frac{\lambda}{\lambda_g} \quad (\text{from the right-angled triangle } ABO) \quad (14.2.5)$$

and since

$$\sin^2 \alpha + \cos^2 \alpha = 1$$

it follows that

$$\left(\frac{\lambda}{\lambda_g}\right)^2 + \left(\frac{\lambda}{2a}\right)^2 = 1$$

or

$$\frac{1}{\lambda_g^2} = \frac{1}{\lambda^2} - \frac{1}{(2a)^2} \quad (14.2.6)$$

where λ_g is the guide wavelength of the TE_{10} mode, λ is the free-space wavelength of either TEM wave, and a is the broad dimension of the guide.

From Eq. (14.2.6), it can be seen that when $\lambda = 2a$ the guide wavelength becomes infinite; this corresponds to the individual TEM waves bouncing from side to side with no velocity component directed along the guide. Clearly, then, $\lambda = 2a$ represents the longest-wavelength TEM wave that can be induced into a guide, as a TE mode will not be generated for longer wavelengths. For shorter TEM wavelengths, Eq. (14.2.6) shows that the guide wavelength is real and positive, so that TE-mode propagation takes place.

The term $2a$ is referred to as the cutoff wavelength λ_c of the TE_{10} mode, and, rewriting Eq. (14.2.6), we have

$$\frac{1}{\lambda_g^2} = \frac{1}{\lambda^2} - \frac{1}{\lambda_c^2} \quad (14.2.7)$$

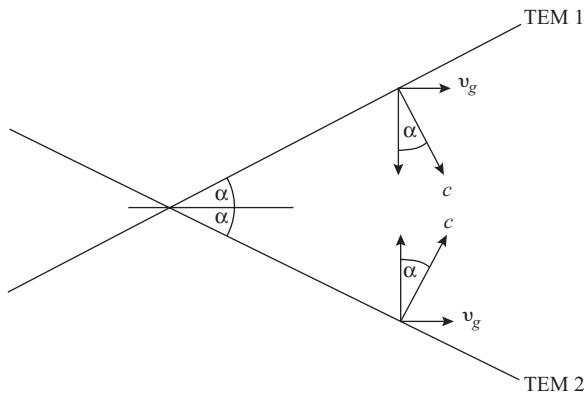


Figure 14.2.3 Group velocity as the horizontal components of light velocity c .

From Eqs. (14.2.3) and (14.2.5), it follows that

$$v_p = \frac{c}{\sin \alpha} \quad (14.2.8)$$

Since the sine of an angle is never greater than unity, v_p can never be less than c ; it can, of course, be greater.

The individual waves have to travel a zigzag path, and therefore the velocity component along the guide, which is the velocity at which the energy in the waves is conveyed, is less than c . From Fig. 14.2.3 this component, known as the group velocity v_g , is

$$v_g = c \sin \alpha \quad (14.2.9)$$

Combining Eqs. (14.2.8) and (14.2.9),

$$v_g v_p = c^2 \quad (14.2.10)$$

Equation (14.2.9) can also be rewritten in terms of wavelengths, since these are the quantities that are normally known:

$$v_g = c \frac{\lambda}{\lambda_g} \quad (14.2.11)$$

Unless otherwise stated, it is assumed that the waveguide dielectric is dry air, for which free-space permittivity and permeability apply.

EXAMPLE 14.2.1

A rectangular waveguide has a broad wall dimension of 0.900 in. and is fed by a 10-GHz carrier from a coaxial cable as shown in Fig. 14.2.4. Determine whether a TE_{10} wave will be propagated, and, if so, find its guide wavelength, phase, and group velocities.

SOLUTION

$$\begin{aligned}a &= 0.900 \text{ in.} \\&= 2.286 \text{ cm}\end{aligned}$$

Therefore,

$$\begin{aligned}\lambda_c &= 2 \times 2.286 \\&= 4.572 \text{ cm} \\&\lambda = \frac{3 \times 10^8}{10^{10}} \\&= 3 \text{ cm}\end{aligned}$$

Note that it is the free-space wavelength, and not the wavelength along the coaxial cable, that is used. Therefore, $\lambda_c > \lambda$, and a TE₁₀ wave will propagate. Hence

$$\frac{1}{\lambda_g^2} = \frac{1}{\lambda^2} - \frac{1}{\lambda_c^2}$$

Therefore,

$$\begin{aligned}\lambda_g &= \frac{\lambda}{\sqrt{1 - (\lambda/\lambda_c)^2}} \\&= \frac{3}{\sqrt{1 - 0.431}} \\&= 3.975 \text{ cm}\end{aligned}$$

$$\begin{aligned}v_p &= c \frac{\lambda_g}{\lambda} \\&= 3 \times 10^8 \times 1.325 \\&= 3.975 \times 10^8 \text{ m/s}\end{aligned}$$

$$\begin{aligned}v_g &= c \frac{\lambda}{\lambda_g} \\&= \frac{3 \times 10^8}{1.325} \\&= 2.264 \times 10^8 \text{ m/s}\end{aligned}$$

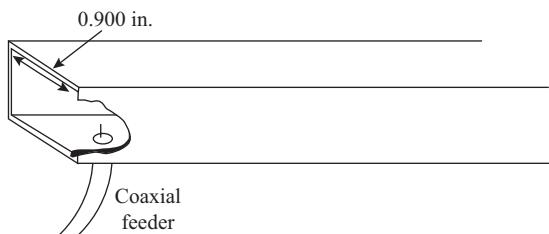


Figure 14.2.4 Example 14.2.1.

The *wave impedance* is defined as the ratio of the transverse components of electric to magnetic fields. In synthesizing the TE₁₀ mode, the two TEM waves are of equal amplitude and time phase; thus $E_1 = E_2 = E$ and $H_1 = H_2 = H$. The entire electric field E is transverse, along dimension b , while from Fig. 14.2.2 it is easily ascertained that the transverse component of H lies along dimension a and is

$$H_a = H \sin \alpha \quad (14.2.12)$$

Hence the wave impedance for the TE₁₀ mode is

$$\begin{aligned} Z_\omega &= \frac{E_b}{H_a} \\ &= \frac{E}{H \sin \alpha} \end{aligned} \quad (14.2.13)$$

In Appendix B the wave impedance of a TEM wave in free space is shown to be [Eq. (B.10)]

$$\begin{aligned} Z_0 &= \frac{E}{H} \\ &= \sqrt{\frac{\mu_0}{\epsilon_0}} \end{aligned} \quad (14.2.14)$$

Thus, substituting Eqs. (14.2.14) and (14.2.5) in Eq. (14.2.13) gives

$$Z_\omega = Z_0 \frac{\lambda_g}{\lambda} \quad (14.2.15)$$

$$= 120\pi \frac{\lambda_g}{\lambda} \Omega \quad (14.2.16)$$

when the numerical values for μ_0 and ϵ_0 are substituted.

The importance of the wave impedance concept is that it can be used in an analogous manner to the characteristic impedance of a transmission line, so the theory of reflections, standing waves, and Smith charts developed in Chapter 13 can also be applied to waveguides.

Standing Waves

Consider a section of waveguide closed by a perfectly conducting sheet at the end. The boundary conditions require that the fields adjust so that the electric field is zero and the magnetic field is entirely tangential at the closure (that is, the field patterns are similar to those at the walls of the guide). The resultant wave pattern can be accounted for in terms of the incident TE wave and a reflected TE wave, the combination of which sets up a standing-wave pattern along the guide, similar to that described in Section 13.7 for transmission lines. It is important to realize that the resultant wave pattern is stationary in space and varies in time, whereas the traveling wave shown in Fig. 14.2.2(b) is time invariant, but moves along the guide. Figure 14.2.5 emphasizes this point. On the left are shown the conditions in a short-circuited guide at time intervals of one-fourth the periodic time of the wave, the reference time being chosen at a maximum field condition. The sequence on the right shows how a single traveling wave would vary over the same time intervals. (For clarity, only the magnetic field loops are shown.)

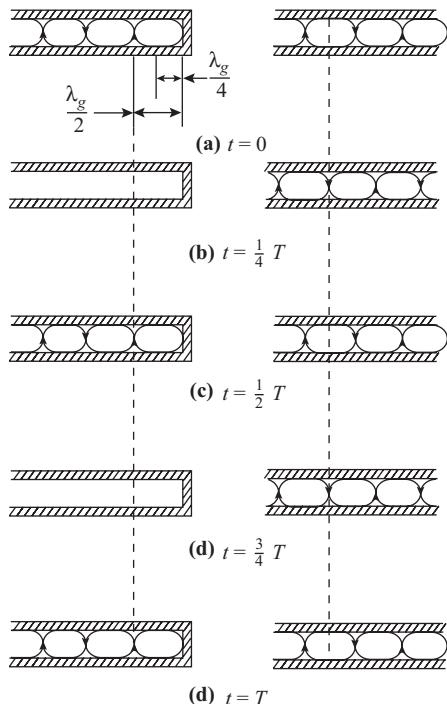


Figure 14.2.5 Comparing standing waves (left) to traveling waves (right).

It will be observed that at $\frac{1}{4}\lambda_g$ from the closed end the transverse magnetic field is zero, whereas the electric field is a maximum. Therefore, the wave impedance is infinite at this section across the guide. At the $\frac{1}{2}\lambda_g$ section, the electric field is zero and the transverse magnetic field component is a maximum, so that the wave impedance is zero. Thus the quarter-wavelength and half-wavelength sections of guide have the same transformation properties as those of the transmission-line sections described in Section 13.11.

Waveguide Terminations

To terminate a waveguide without causing reflections, the terminating impedance must equal the wave impedance of the incident wave. Consider first a thin sheet of resistive material closing the end of the waveguide [Fig. 14.2.6(a)]. Between the top and bottom of the guide, the sheet presents a resistance of path length b and cross-sectional area ta , where t is the thickness of the sheet; a and b are the guide dimensions defined previously. Only the sheet resistance between top and bottom need be considered, since the current direction is at right angles to the H -field at the sheet. Let ρ be the resistivity of the sheet material; then the terminating resistance is

$$R_T = \rho \frac{b}{ta} \quad (14.2.17)$$

The *sheet resistivity* R_s is defined as

$$R_s = \frac{\rho}{t} \quad \Omega \text{ per square} \quad (14.2.18)$$

The units of the square do not matter. R_s will have the same value for a square inch, a square centimeter, and so on. What does matter is that the units of ρ and t should be consistent. If ρ is in $\Omega\text{-m}$, then t must be in meters and R_s will be in ohms per square. If ρ is in $\mu\Omega\text{-in.}$, t must be in inches and R_s will be in $\mu\Omega$ per square.

Now, the current I in the sheet supports the tangential component of the magnetic field H_T , given by

$$H_T = \frac{I}{a} \quad (14.2.19)$$

Also, the voltage across the sheet gives rise to an electric-field component E_T :

$$E_T = \frac{V}{b} \quad (14.2.20)$$

Therefore,

$$\frac{E_T}{H_T} = \frac{Va}{Ib} \quad (14.2.21)$$

But $V/I = R_T = R_s(b/a)$, and substituting this into Eq. (14.2.21) gives

$$\begin{aligned} \frac{E_T}{H_T} &= R_s \frac{b}{a} \frac{a}{b} \\ &= R_s \end{aligned} \quad (14.2.22)$$

E_T and H_T will consist of three components: incident, reflected, and transmitted waves, shown in Fig. 14.2.6(a) as W_i , W_r , and W_t , respectively. The transmitted wave can be eliminated by extending the guide $\frac{1}{4}\lambda_g$ beyond the termination sheet, as shown in Fig. 14.2.6(b) and short-circuiting the end. The $\frac{1}{4}\lambda_g$ shorted section presents an infinite impedance; therefore, no wave can be transmitted into it. Next, the sheet resistivity R_s is made equal to the incident-wave impedance Z_ω :

$$Z_\omega = R_s \quad (14.2.23)$$

This ensures that $E_T/H_T = E_i/H_i = E/H$; that is, the incident wave is absorbed by the sheet just as though it were continuing down an infinitely long waveguide in which the wave impedance was Z_ω (or R_s). Thus the reflected wave is eliminated.

The arrangement discussed illustrates two important principles, that of sheet resistivity and the idea of using a $\frac{1}{4}\lambda_g$ shorted section to isolate the load from external impedances. The arrangement is difficult to implement in practice, however, largely because it requires an adjustable short circuit on the $\frac{1}{4}\lambda_g$ section, which is rather critical to adjust. More practical arrangements for terminating a guide are shown in Fig. 14.2.7. The resistive strip is set in the plane of the maximum electric field, and the taper ensures a gradual change in electric field, which reduces reflections to a negligible level. Either form of taper shown in Fig. 14.2.7 is satisfactory, and the dimensions shown are typical of those used in practice. A carbon-coated strip may be used,

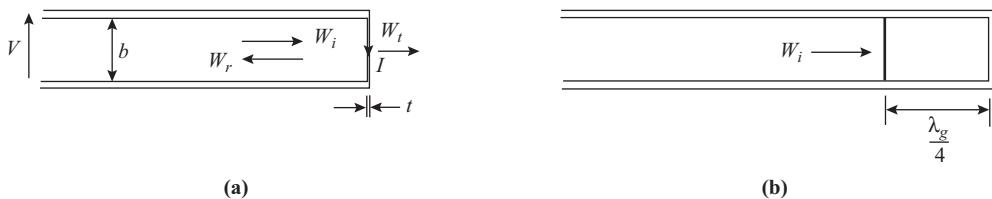


Figure 14.2.6 (a) Waveguide terminated with sheet resistance. $\lambda_g/4$ section added to termination.

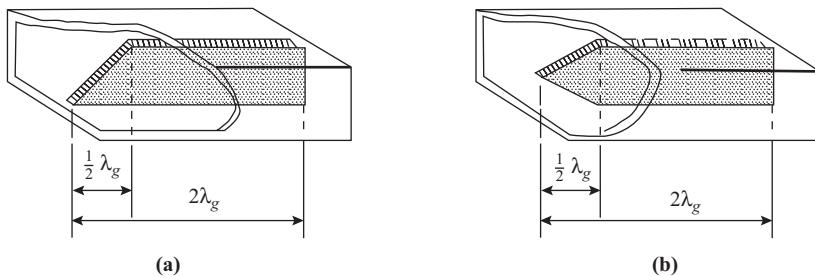


Figure 14.2.7 Practical waveguide terminations.

or for more stable operation, a glass strip covered with a thin film of metal, which, in turn, has a thin dielectric coating for protection. Surface resistivities of the order of 500Ω per square are typical.

Attenuators

An arrangement similar to that shown in Fig. 14.2.7 can be used to provide attenuation in a waveguide. Two common methods are shown in Fig. 14.2.8. In Fig. 14.2.8(a), the thin resistive sheet can be moved from the sidewall, where it produces minimum attenuation, to the center of the guide, where it produces maximum attenuation. The mechanical drive for the sheet is often fitted with a micrometer control so that fine adjustment of attenuation can be made and accurately calibrated. The flap attenuator of Fig. 14.2.8(b) is simple to construct, and, as shown in the next section, the slot position is such that radiation is minimized. However, some radiation does occur, and this type is not used for accurate work.

Currents in Walls

As already stated, the traveling-wave pattern is time invariant; that is, the pattern appears to keep its shape as it moves down the guide. At any given cross section of the guide, however, the magnetic field (and the electric field) appears to vary in time as the loops of alternate polarity sweep by. This gives rise to induced currents in the walls of the guide, at right angles to the magnetic field. The currents for the TE_{10} mode are as shown in Fig. 14.2.9(a). This pattern moves down the guide along with the field pattern at the phase velocity. It is

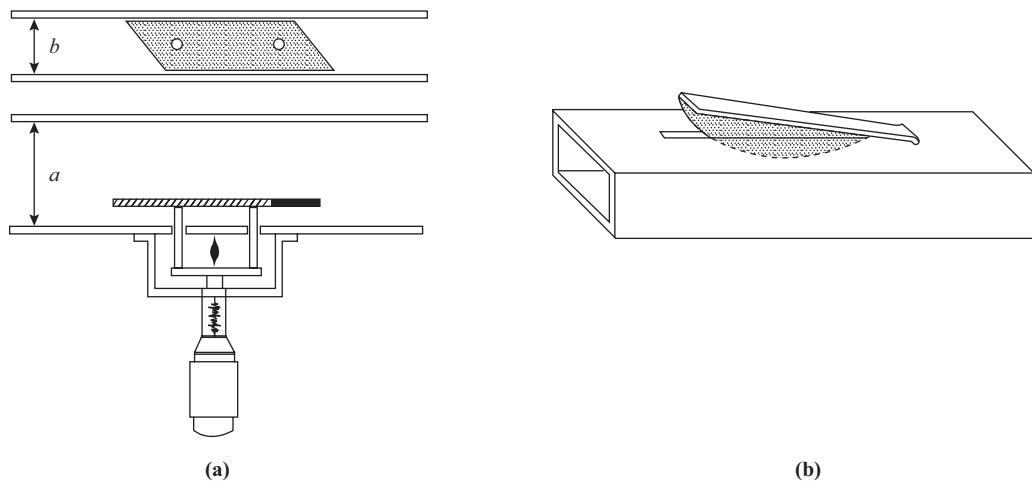


Figure 14.2.8 Waveguide attenuators.

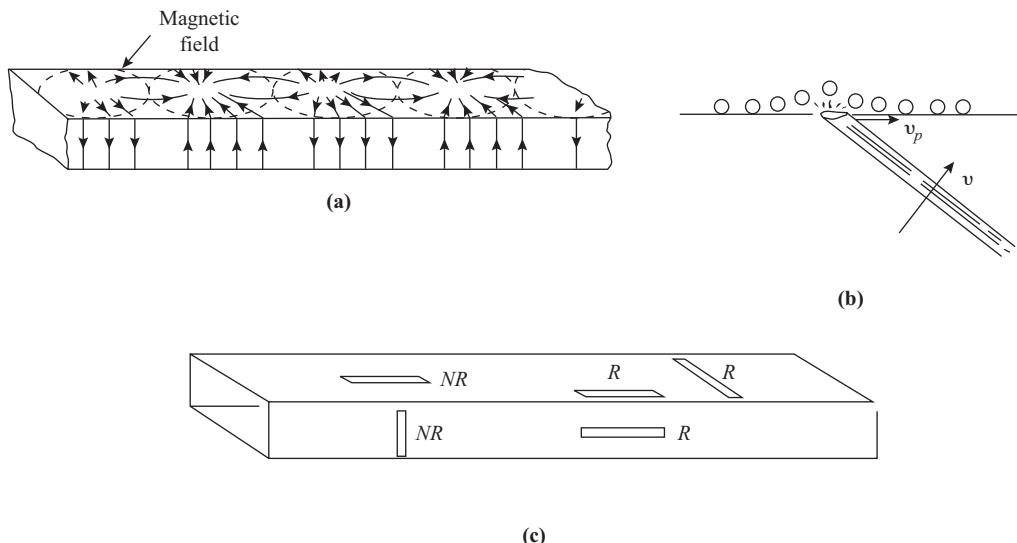


Figure 14.2.9 (a) Currents in walls for the TE_{10} mode. (b) Sea-wave analogy illustrating phase velocity. (c) Nonradiating (NR) and radiating (R) slots.

emphasized that the pattern moves at the phase velocity, not the current (since that would imply electrons moving faster than the speed of light!). In effect, the current pattern builds up and decays as the TE wave sweeps by. An analogy can be drawn with a sea wave moving along a sea wall, creating a splash [Fig. 14.2.9(b)]. The splash travels at the phase velocity (which is faster than the velocity at which the wave approaches the wall), and people standing at the edge of the sea wall will move back as the splash passes them, so the ripple in the line moves at the phase velocity.

Knowing the current pattern is important, as it enables slots to be correctly positioned in the walls, to serve various purposes. Slots that do not noticeably interrupt the current flow are termed *nonradiating*, since they result in minimum disturbance of the internal fields, and therefore little electromagnetic energy leaks through them. The positions of two nonradiating slots, labeled NR, are shown in Fig. 14.2.9(c). It will be seen that a nonradiating slot can be placed along the center of the broad wall of the guide, and use is made of this in the attenuator of Fig. 14.2.8(b). Another application is the slotted waveguide, which enables standing waves to be measured, the technique being similar to that described in Section 13.10 for transmission lines.

Slots that do interrupt the current flow, such as those labeled R in Fig. 14.2.9(c), are termed *radiating* slots. These produce maximum disturbance of the internal fields, which results in energy being radiated. Radiating slots form the basis of slot antennas.

Contacts and Joints

The properties of a short-circuited section can be used to provide an electrical short circuit without the necessity of providing a solid mechanical contact at the point of short circuit. This principle is incorporated into the design of some types of flanges used for coupling guide sections together and in the design of movable short-circuiting contacts. Two $\frac{1}{4}\lambda_g$ transformations take place, as shown in Fig. 14.2.10(a). The top $\frac{1}{4}\lambda_g$ section transforms the solid short circuit at the top to an open circuit at the junction of the two $\frac{1}{4}\lambda_g$ sections. Here a mechanical joint occurs, but since this is a high-impedance point the currents are small and the joint resistance is not important. The second $\frac{1}{4}\lambda_g$ section transforms the open circuit back to a short circuit at the entry point. Application of the principle to a coupling flange is illustrated in Fig. 14.2.10(b) and to a movable short

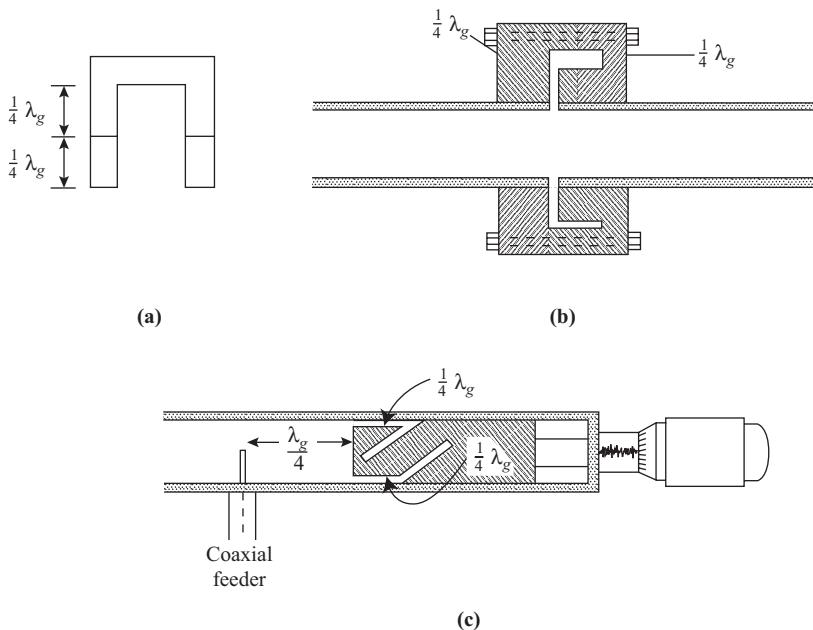


Figure 14.2.10 Two $\lambda_g/4$ transforming sections in series. (b) Coupling flange utilizing (a). (c) Movable short-circuiting contact utilizing (a). The contact illustrated provides a $\lambda_g/4$ shorted section for the coaxial feeder.

circuit in Fig. 14.2.10(c). In the latter figure, a coaxial feeder is shown by way of example, the $\frac{1}{4}\lambda_g$ short-circuited section performing the same function as that shown in Fig. 14.2.6(b).

Reactive Stubs

Equivalent reactive elements can be introduced into a waveguide in a variety of ways, a common method being to use an adjustable screw stub as shown in Fig. 14.2.11. When the stub is only a short way in, as shown in Fig. 14.2.11(a), it acts as a capacitor, as it produces an increase in the electric flux density of the wave in the vicinity of the stub.

With the screw all the way in, such that it forms a post between the top and bottom of the guide [Fig. 14.2.11(b)], a path is provided for induced currents that set up a magnetic field; such a post therefore acts as an inductor. An equivalent series LC circuit is formed when the screw is sufficiently far in for both components to be significant but neither dominant, such as is shown in Fig. 14.2.11(c). Screw stubs are used singly and in groups of two and three to provide matching devices between a waveguide and load.

14.3 Other Modes

Higher modes can occur in waveguides: an example of the TE_{20} mode is sketched in Fig. 14.3.1(a). *Transverse magnetic (TM) modes* can also occur, as the TM_{11} mode sketched in Fig. 14.3.1(b) illustrates. It can be shown that the cutoff wavelength for TE_{mn} and TM_{mn} modes in general is given by

$$\left(\frac{1}{\lambda_c}\right)^2 = \left(\frac{m}{2a}\right)^2 = \left(\frac{n}{2b}\right)^2 \quad (14.3.1)$$

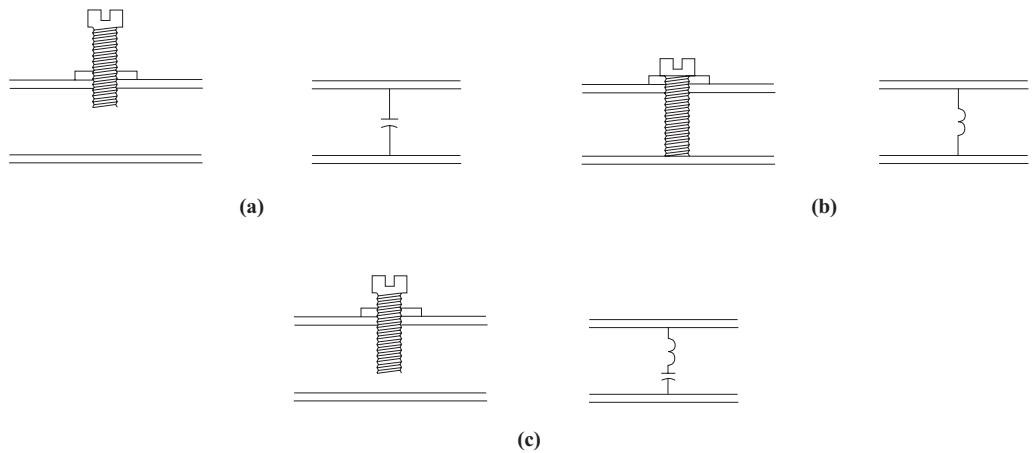


Figure 14.2.11 Reactive stubs: (a) Capacitive stub, (b) Series LC stub.

where m and n are integers. Equation (14.2.24) reduces to $\lambda_c = 2a$ for $m = 1$ and $n = 0$, which is the condition already introduced in Eq. (14.2.7) for the TE_{10} mode.

The TM_{11} mode is the lowest TM mode that can occur, as the boundary conditions exclude the TM_{10} mode. For transmission purposes, only the TE_{10} mode is used, and the guide dimensions are chosen, in conjunction with the input frequency, to cut off all but the dominant mode (TE_{10}). For example, for a standard guide WR 90, the dimensions are

$$\begin{aligned} \text{Outside walls: } & 1.000 \times 0.500 \text{ in.} \\ \text{Wall thickness: } & 0.050 \text{ in.} \end{aligned}$$

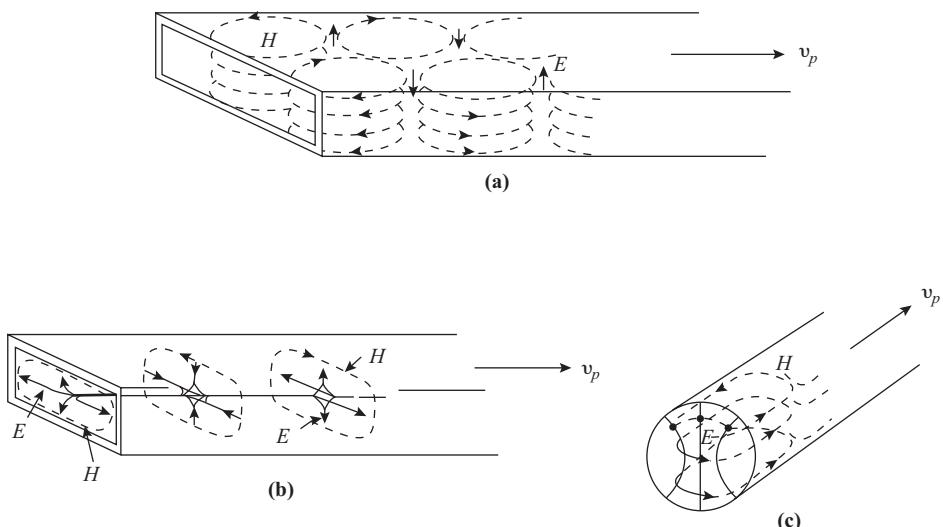


Figure 14.3.1 Other waveguide modes: (a) TE_{20} mode, (b) TM_{11} mode, (c) TE_{11} circular waveguide mode.

The inside wall dimensions are, therefore,

$$\begin{aligned}a &= 2.286 \text{ cm} \\b &= 1.016 \text{ cm}\end{aligned}$$

The cutoff wavelengths for some of the various modes are

$$\begin{aligned}\text{TE}_{10}: \quad \lambda_c &= 2a \\&= 4.572 \text{ cm}\end{aligned}$$

$$\begin{aligned}\text{TE}_{20}: \quad \lambda_c &= a \\&= 2.286 \text{ cm}\end{aligned}$$

$$\begin{aligned}\text{TE}_{01}: \quad \lambda_c &= 2b \\&= 2.032 \text{ cm}\end{aligned}$$

$$\left. \begin{aligned}\text{TE}_{11} \\ \text{TM}_{11}\end{aligned} \right\} \quad \left(\frac{1}{\lambda_c} \right)^2 = \left(\frac{1}{2a} \right)^2 + \left(\frac{1}{2b} \right)^2$$

Therefore,

$$\lambda_c = 1.857 \text{ cm}$$

These results are tabulated as follows:

Mode	$\lambda_c(\text{cm})$	$f_c(\text{GHz})$
TE_{10}	4.572	6.56
TE_{20}	2.286	14.1
TE_{01}	2.032	14.8
TE_{11}	1.857	16.2

The recommended operating frequency range for the WR90 guide is 8.20 to 12.40 GHz, and it will be seen that only the TE_{10} mode will be excited. Of course, if a higher frequency is fed into the guide, higher modes will be excited; for example, if a frequency of 15 GHz is used, the first three modes may exist simultaneously in the guide. For all these modes, including the circular mode described in the next paragraph, Eq. (14.3.1) applies.

Circular guides also support waveguide modes; the TE_{11} circular mode is sketched in Fig. 14.3.1(c). (The subscripts here denote circumferential and radial variations of the fields, which are much more complex than the modes in the rectangular guide; only the TE_{11} circular mode will be considered here.) The cut-off wavelength is related to the diameter of the guide, the value derived from theory being

$$\lambda_c = 1.71d \quad (14.3.2)$$

where d is the diameter of the circular guide.

Circular guides have special properties that enable them to be used for rotating joints, and the TE_{10} mode has the unusual characteristic of attenuation becoming less as the frequency is increased, which is attractive for transmissions at the higher microwave frequencies. However, the mechanical problems of making and

maintaining precise dimensions in a circular guide are much more formidable than in a rectangular guide, and the latter finds greater use for transmission purposes.

PROBLEMS

- 14.1. Determine λ_c , λ_g , λ , v_p , and v_g for standard waveguide WR62 for which the outside dimensions are 0.702 X 0.391 in. and the wall thickness is 0.040 in. when the exciting frequency is 10 GHz.
- 14.2. Determine the TE_{10} wave impedance for the WR62 guide of Problem 14.1
- 14.3. Explain how standing waves can be produced in a waveguide. What is the spacing, in terms of wavelength, between successive minima along the standing-wave pattern?
- 14.4. By analogy with the transmission-line theory presented in Section 13.8, determine the reflection coefficient for the electric field of a TE_{10} wave of wave impedance 680Ω terminated as shown in Fig. 14.2.6(b), the sheet resistivity of the load being $500 \Omega/\text{square}$.
- 14.5. Also by analogy to the transmission theory presented in Section 13.9, determine the electric field standing-wave ratio for the guide conditions of Problem 14.4.
- 14.6. What is meant by sheet resistivity, and how does this differ from the resistivity of a material? What are the units for sheet resistivity?
- 14.7. Illustrate on a Smith chart the transformations that take place in the $\frac{1}{2}\lambda_g$ short-circuited section shown in Fig. 14.2.10(a).
- 14.8. Explain the difference between radiating and nonradiating slots in a TE_{10} mode waveguide.
- 14.9. A standard waveguide, WR187, has the following dimensions: outside walls 2.000×1.000 in., wall thickness 0.064 in. Determine the highest frequency that can be transmitted if only the TE_{10} mode is permitted. What modes can exist in it if it is excited by a wave of frequency 6.00 GHz?
- 14.10. Determine (a) cutoff wavelength, (b) guide wavelength, (c) phase constant, (d) phase velocity, and (e) wave independence in the case of a hollow rectangular metallic waveguide of dimensions 4 cm and 2 cm, respectively, when the applied frequency is 4 GHz. Assume that TE_{22} mode gets transmitted through the waveguide.
- 14.11. Repeat the above problem with dimensions 8 cm and 4 cm, and with propagating frequency 12 GHz. Assume that TE_{10} mode is propagating.
- 14.12. Determine the cutoff frequency for the TE_{10} mode in a rectangular waveguide of dimensions 8 cm and 4 cm, respectively, given that the distance between a voltage maximum and a minimum is 4 cm.
- 14.13. Determine the cutoff frequency for the TE_{11} mode in a rectangular waveguide of dimensions 10 cm and 5 cm, respectively, given that the distance between a voltage maximum and a minimum is 5 cm.
- 14.14. Compute the value of characteristic impedance of a microstrip transmission line, given that $l = 5\mu\text{m}$, $\omega = 2\mu\text{m}$, and $\epsilon_r = 12$.
- 14.15. Repeat the above exercise for $l = 5\mu\text{m}$, $\omega = 10\mu\text{m}$, and $\epsilon_r = 12$.
- 14.16. Calculate the quality factor of a *cavity resonator*, given that $f = 10\text{GHz}$, $\mu_r = 1$, $a = 4\text{cm}$; $R_s = 2m\Omega$.
- 14.17. A cavity resonator has dimensions of 12 cm, 6 cm, and 3 cm, respectively. Compute its *resonant frequency*.
- 14.18. A *micro-strip* line has $\epsilon_r = 5$ and $Z_o = 50\Omega$. Compute the values of *line inductance* and *line capacitance*.
- 14.19. A micro-strip line has substrate thickness of 1 mm and $\epsilon_r = 10$. The operating frequency is 12 GHz. The conductor width is 0.8 mm and thickness is 0.2 mm. Calculate Z_o *surface skin resistivity* of Copper and *conductor attenuation*.



Radio-wave Propagation

15.1 Introduction

Radio communications use electromagnetic waves propagated through the earth's atmosphere or space to carry information over long distances without the use of wires. Radio waves with frequencies ranging from about 100 Hz in the ELF band to well above 300 GHz in the EHF band have been used for communications purposes, and more recently radiation in and near the visible range (near 1000 THz, or 10^{15} Hz) have also been used. Figure 15.8.1 shows the frequency-band designations in common use.

Some of the basic properties of a *transverse electromagnetic* (TEM) wave are described in Appendix B. Although the electric and magnetic fields exist simultaneously, in practice, antennas are designed to work through one or other of these fields. Antennas are described in Chapter 16. Basically, to launch an electromagnetic wave into space, an electric charge has to be accelerated, which in practice means that the current in the radiator must change with time (for example, be alternating). In this chapter, sinusoidal or cosinusoidal variations will be assumed unless stated otherwise.

15.2 Propagation in Free Space

Mode of Propagation

Consider first an average power P_T , assumed to be radiated equally in all directions (isotropically). This will spread out spherically as it travels away from the source, so that at distance d , the power density in the wave, which is the power per unit area of wavefront, will be

$$P_{Di} = \frac{P_T}{4\pi d^2} \text{ W/m}^2 \quad (15.2.1)$$

This is so because $4\pi d^2$ is the surface area of the sphere of radius d , centered on the source. P_{Di} stands for isotropic power density.

It is known that all practical antennas have directional characteristics; that is, they radiate more power in some directions at the expense of less in others. The directivity gain is the ratio of actual power density along the main axis of radiation of the antenna to that which would be produced by an isotropic antenna at the same distance fed with the same input power. Let G_T be the *maximum* directivity gain of the transmitting antenna; then the power density along the direction of maximum radiation will be

$$\begin{aligned} P_D &= P_{Di}G_T \\ &= \frac{P_T G_T}{4\pi d^2} \end{aligned} \quad (15.2.2)$$

A receiving antenna can be positioned so that it collects maximum power from the wave. When so positioned, let P_R be the power delivered by the antenna to the load (receiver) under matched conditions; then the antenna can be considered as having an effective area (or aperture) A_{eff} , where

$$\begin{aligned} P_R &= P_D A_{\text{eff}} \\ &= \frac{P_T G_T}{4\pi d^2} A_{\text{eff}} \end{aligned} \quad (15.2.3)$$

It can be shown that for any antenna, the ratio of maximum directivity gain to effective area is

$$\frac{A_{\text{eff}}}{G} = \frac{\lambda^2}{4\pi} \quad (15.2.4)$$

Here, λ is the wavelength of the wave being radiated. Letting G_R be the maximum directivity gain of the receiving antenna, we have

$$\frac{P_R}{P_T} = G_T G_R \left(\frac{\lambda}{4\pi d} \right)^2 \quad (15.2.5)$$

This is the fundamental equation for free-space transmission. Usually it is expressed in terms of frequency f , in megahertz, and distance d , in kilometers. As shown in Appendix B, $\lambda f = c$, and on substituting this in Eq. (15.2.5) and doing the arithmetic, which is left as an exercise for the reader, the result obtained is

$$\frac{P_R}{P_T} = G_T G_R \frac{0.57 \times 10^{-3}}{(df)^2} \quad (15.2.6)$$

By expressing the power ratios in decibels, Eq. (15.2.6) can be written as

$$\left(\frac{P_R}{P_T} \right)_{\text{dB}} = (G_T)_{\text{dB}} + (G_R)_{\text{dB}} - (32.5 + 20 \log_{10} d + 20 \log_{10} f) \quad (15.2.7)$$

The third term in parentheses on the right-hand side of Eq. (15.2.7) is the loss, in decibels, resulting from the spreading of the wave as it propagates outward from the source. It is known as the transmission path loss, L . Thus

$$L = (32.5 + 20 \log_{10} d + 20 \log_{10} f)_{\text{dB}} \quad (15.2.8)$$

where d is in kilometers and f in megahertz.

Equation (15.2.7) then becomes

$$\left(\frac{P_R}{P_T} \right)_{\text{dB}} = (G_T)_{\text{dB}} + (G_R)_{\text{dB}} - (L)_{\text{dB}} \quad (15.2.9)$$

EXAMPLE 15.2.1

In a satellite communications system, free-space conditions may be assumed. The satellite is at a height of 36,000 km above earth, the frequency used is 4000 MHz, the transmitting antenna gain is 15 dB, and the receiving antenna gain is 45 dB. Calculate (a) the free-space transmission loss and (b) the received power when the transmitted power is 200 W.

SOLUTION (a)
$$\begin{aligned} L &= 32.5 + 20 \log_{10} 36,000 + 20 \log_{10} 4000 \\ &= 196 \text{ dB} \end{aligned}$$

(b)
$$\left(\frac{P_R}{P_T} \right)_{\text{dB}} = 15 + 45 - 196 = -136 \text{ dB}$$

This is a power ratio of 0.25×10^{-13} , and since $P_T = 200 \text{ W}$,

$$\begin{aligned} P_R &= 200 \times 0.25 \times 10^{-13} \\ &= 5 \times 10^{-12} \text{ W} = 5 \text{ pW} \end{aligned}$$

Frequently, it is required to know the electric field strength of the wave at the receiving antenna. In Appendix B, E is given by Eq. (B.12) in terms of power density P_D and wave impedance Z_0 as

$$E = \sqrt{Z_0 P_D} \quad (15.2.10)$$

Also, Z_0 is given by Eq. (B.10) as

$$Z_0 = \sqrt{\frac{\mu}{\epsilon}} \quad (15.2.11)$$

The free-space values are as follows: $\mu = \mu_0 = 4\pi \times 10^{-7} \text{ H/m}$, and $\epsilon = \epsilon_0 = 8.854 \times 10^{-12} \text{ F/m}$. Substituting these in Eq. (15.2.11) gives

$$Z_0 = 120\pi \Omega \quad (15.2.12)$$

The field strength may now be found by substituting Eqs. (15.2.2) and (15.2.12) in Eq. (15.2.10)

$$E = \frac{\sqrt{30P_T G_T}}{d} \text{ V/m} \quad (15.2.13)$$

This is the fundamental equation that gives the field strength at the receiving antenna, for free-space propagation conditions. A receiving antenna has an effective length ℓ_{eff} (analogous to effective area) such that the open circuit EMF of the antenna V_s is given by

$$V_s = E\ell_{\text{eff}} \quad (15.2.14)$$

Effective length is discussed in Section 16.9.

EXAMPLE 15.2.2

Calculate the open-circuit voltage induced in a $\frac{1}{2}\lambda$ dipole when 10 W at 150 MHz is radiated from another $\frac{1}{2}\lambda$ dipole 50 km distant. The antennas are positioned for optimum transmission and reception.

SOLUTION

$$\begin{aligned}\lambda &= \frac{3 \times 10^8}{150 \times 10^6} \\ &= 2 \text{ m}\end{aligned}$$

In Chapter 16 it is shown that, for a $\frac{1}{2}\lambda$ dipole, the maximum gain is 1.64 : 1, and the effective length is λ/π . Therefore,

$$\begin{aligned}V_s &= \frac{\sqrt{30 \times 10 \times 1.64}}{50 \times 10^3} \cdot \frac{2}{\pi} \\ &= 282 \mu\text{V}\end{aligned}$$

Equation (15.2.13) is sometimes expressed in terms of the field strength at unit distance, E_0 . Thus, at $d = 1 \text{ m}$,

$$E = E_0 = \frac{\sqrt{30P_TG_T}}{1} \text{ V/m} \quad (15.2.15)$$

and therefore Eq. (15.2.13) can be written as

$$E = \frac{E_0}{d} \text{ V/m} \quad (15.2.16)$$

Note that both E and E_0 are in units of V/m, although Eq. (15.2.16) may tend to suggest that E was in units of $(\text{V}/\text{m})/\text{m}$, or V/m^2 . Equation (15.2.16) really expresses a proportionality, and written in full it would be

$$E = E_0 \times \frac{1 \text{ (m)}}{d \text{ (m)}} \text{ V/m} \quad (15.2.17)$$

Microwave Systems

Microwave radio systems operating at frequencies above 1 GHz propagate mainly in a line-of-sight or free-space mode, whether they are on the ground or in satellite systems. Since the 1950s, microwave radio systems have become the workhorses of long-distance telephone communications systems. These systems provide the needed transmission bandwidth and reliability to allow the transmission of many thousands of telephone channels as well as several television channels over the same route and using the same facilities. Carrier frequencies in the 3- to 12-GHz range are used. Since microwaves travel only on line-of-sight paths, it is necessary to provide repeater stations at about 50-km intervals. This makes the equipment costs for such a system very large, but this is more than made up for by the increased channel capacity. Transmitter output powers are low (they may be less than 1 W), because highly directional high-gain antennas are used.

Figure 15.2.1(a) shows the equipment needed to provide one channel of a microwave system. It consists of two terminal stations and one or more repeater stations. At the sending terminal, the inputs comprising several hundred telephone channels and/or a television channel are frequency-multiplexed within the baseband band pass of 6 MHz. The baseband frequency modulates a 70-MHz IF signal, which is then up-converted to the microwave output frequency f_1 within the 4-GHz band. This signal is amplified and fed through a directional antenna toward a repeater station some 50 km distant. At the repeater station, the signal at f_1 is received on one antenna pointed toward the originating station, down-converted to the IF, amplified, and up-converted to a new frequency f_2 retransmission toward the receiving terminal station. When the signal is passed through a chain of several repeaters, alternate links in the chain use alternate frequencies so that retransmitted energy at a repeater station does not feed back into its own receiver.

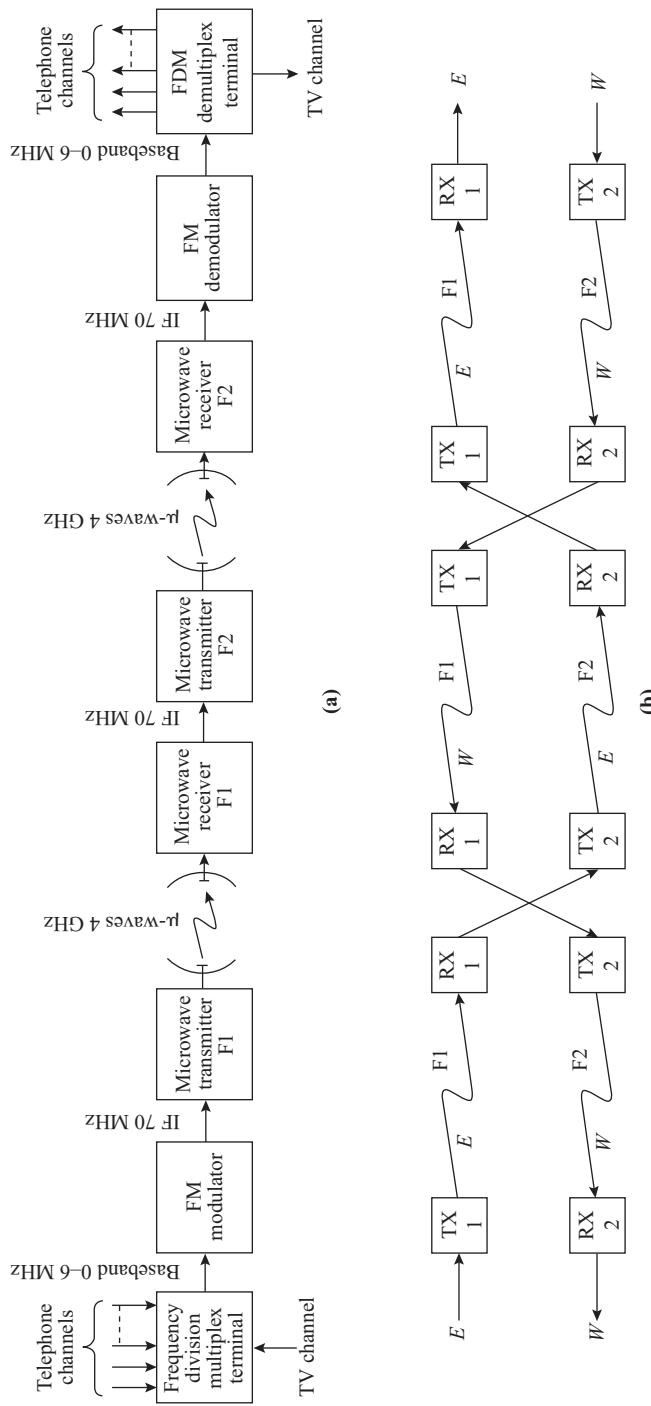


Figure 15.2.1 Microwave relay system. (a) One-way channel showing the equipment used on the route. (b) Two-way channel pair showing frequency interchanges at intermediate repeater stations.

With two frequencies in use, alternating at each repeater, two channels, one in each direction, can be provided, as illustrated in Fig. 15.2.1(b). In some microwave systems several two-way channel pairs are provided, and a more complex system of frequency switching is used at the repeater stations.

At the receiving terminal station, the signal is down-converted to the IF and then demodulated to recover the baseband signal. This baseband signal is then demultiplexed to recover the individual telephone or television channel signals.

Terminal stations use two antennas, one for receiving and one for transmitting. The system may have several transmitters and receivers, but they all use the same antennas. Repeater stations are provided with two antennas pointing in each direction, for a total of four. Repeater station sites are chosen at high points, such as on hilltops or tall buildings, and sturdy towers provide additional elevation to maximize the distance between stations.

One-way single-channel microwave systems using portable terminal and repeater stations are frequently used for remote television pickup of special events and for other temporary installations, such as the testing of new microwave routes. This equipment is often mounted in mobile vans outfitted with telescopic towers that can be quickly erected.

15.3 Tropospheric Propagation

Mode of Propagation

The troposphere is the region of the earth's atmosphere immediately adjacent to the earth's surface and extending upward for some tens of kilometers (the region in which we normally live and travel, including high-flying jet planes). In this region, the free-space conditions are modified by (1) the surface of the earth and (2) the earth's atmosphere.

Considering first the effect of the earth's surface, a much-simplified model has been developed which successfully describes electromagnetic propagation over a wide range of practical circumstances and is illustrated in Fig. 15.3.1(a). In this simplified picture the earth is assumed to be flat, and the space wave reaching the receiver has two components: the direct wave, which follows a ray path s_d , and a ground-reflected wave, which follows a ray path s_i . The reflected wave travels a greater distance than the direct wave, and although this has negligible effect on the amplitude, it does introduce a phase difference which is highly significant. Let Δs be the path difference; then, since a phase angle of 2π radians corresponds to a path length of one wavelength (λ), the phase angle corresponding to Δs is

$$\phi_s = \frac{2\pi}{\lambda} \Delta s \quad (15.3.1)$$

[Note that $2\pi/\lambda$ is simply the phase-shift coefficient introduced in Eq. (13.5.4)]

The problem now is to find Δs , and this can be solved from the geometry of Fig. 15.3.1(b). h_T is the height of the transmitting antenna above ground, h_R is the height of the receiving antenna, and d is the distance along the flat earth between the two. From triangle FBD [Fig. 15.3.1(b)],

$$s_i^2 = (h_T + h_R)^2 + d^2$$

From triangle ABC ,

$$s_d^2 = (h_T - h_R)^2 + d^2$$

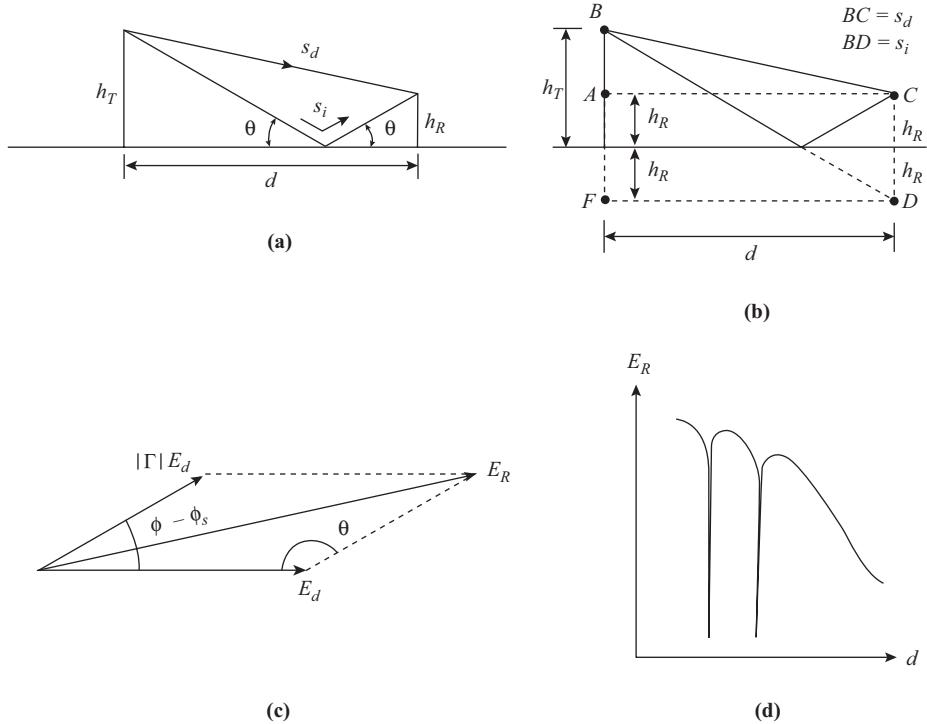


Figure 15.3.1 (a) Simplified model for tropospheric propagation ray paths. (b) Geometry used to determine phase difference. (c) Field strength phasor diagram at the receiver. (d) Sketch of field strength modulus versus distance.

Therefore,

$$\begin{aligned} s_i^2 - s_d^2 &= (h_T + h_R)^2 - (h_T - h_R)^2 \\ &= 4h_T h_R \end{aligned} \quad (15.3.2)$$

Also, applying the formula for the difference of two squares,

$$\begin{aligned} s_i^2 - s_d^2 &= (s_i + s_d)(s_i - s_d) \\ &= (s_i + s_d) \Delta s \end{aligned}$$

For most practical purposes, $s_i \approx s_d \approx d$. Therefore,

$$s_i^2 - s_d^2 \approx (2d) \Delta s \quad (15.3.3)$$

Equating Eqs. (15.3.2) and (15.3.3), the result is

$$2d \Delta s = 4h_T h_R$$

Therefore,

$$\Delta s = \frac{2h_T h_R}{d} \quad (15.3.4)$$

Substituting this in Eq. (15.3.1) gives

$$\phi_s = \frac{4\pi h_T h_R}{\lambda d} \quad (15.3.5)$$

This is not the only phase change that takes place. Reflection at the earth's surface also affects the amplitude and phase of the reflected wave relative to the direct wave. The nature of the reflection depends in a complicated way on the constitution of the reflection surface, the angle of incidence of the wave, and whether the wave is horizontally or vertically polarized (meaning whether the electric field of the wave is horizontal or vertical, respectively, with respect to the earth's surface).

Let the electric field reflection coefficient at the point of reflection be represented by $\Gamma = |\Gamma|/\phi$, analogously to the voltage reflection coefficient introduced in Section 13.8. The amplitude of the reflected wave at the receiving point will be $|\Gamma|E_d$, where E_d is the amplitude of the direct wave. The phase of the reflected wave relative to the direct wave is $(\phi - \phi_s)$, since it is known that ϕ_s is a phase lag because of the longer path length for the reflected wave, and a positive value of ϕ is assumed to be a phase lead. The phasor diagram for the field strength at the receiving point is therefore as shown in Figure 15.3.1(c). The angle θ is equal to $180^\circ - (\phi - \phi_s)$, and applying the cosine rule to the triangle to find the resultant field strength gives

$$\begin{aligned} E_R^2 &= E_d^2 + |\Gamma|^2 E_d^2 - 2|\Gamma|E_d^2 \cos \theta \\ &= E_d^2(1 + |\Gamma|^2 + 2|\Gamma| \cos(\phi - \phi_s)) \end{aligned}$$

Therefore,

$$E_R = E_d \sqrt{1 + |\Gamma|^2 + 2|\Gamma| \cos(\phi - \phi_s)} \quad (15.3.6)$$

For a wide range of conditions it is found that the reflection coefficient is equal to -1 ; that is, the reflection occurs with negligible amplitude change but 180° phase change. For this particular condition, Eq. (15.3.6) reduces to

$$\begin{aligned} E_R &= E_d \sqrt{1 + 1 + 2 \cos(180 - \phi_s)} \\ &= E_d \sqrt{2(1 - \cos \phi_s)} \end{aligned} \quad (15.3.7)$$

Using the identity $2 \sin^2 A = 1 - \cos 2A$ allows this to be written as

$$E_R = 2E_d \sin\left(\frac{\phi_s}{2}\right) \quad (15.3.8)$$

Equation (15.2.6) gives E_d as

$$E_d = \frac{E_0}{d}$$

Thus, substituting Eqs. (15.2.16) and (15.3.5) in Eq. (15.3.8) results in

$$E_R = \frac{2E_0}{d} \sin\left(\frac{2\pi h_T h_R}{\lambda d}\right) \quad (15.3.9)$$

The variation of E_R with d is sketched in Fig. 15.3.1(d). The sharp dips toward zero occur where the sine term goes to zero. When d is large, so that the angle within the brackets of Eq. (15.3.8) is small (less than about 0.5 rad or about 30°), the approximate form

$$E_R \approx \frac{2E_0}{d} \left(\frac{2\pi h_T h_R}{\lambda d} \right)$$

or

$$E_R = E_0 \frac{4\pi h_T h_R}{\lambda d^2} \quad (15.3.10)$$

may be used.

EXAMPLE 15.3.1

In a VHF mobile radio system, the base station transmits 100 W at 150 MHz, and the antenna is 20 m above ground. The transmitting antenna is a $\frac{1}{2}\lambda$ dipole for which the gain is 1.64. Calculate the field strength at a receiving antenna of height 2 m at a distance of 40 km.

SOLUTION

$$\begin{aligned} \lambda &= \frac{300 \times 10^6}{150 \times 10^6} \\ &= 2 \text{ m} \\ E_0 &= \sqrt{30 \times 100 \times 1.64} \quad [\text{Eq.(15.2.15)}] \\ &= 70 \text{ V/m} \end{aligned}$$

The angle $2\pi h_T h_R / \lambda d$ is very much less than 0.5 rad, as an approximate calculation will show. Therefore, Eq. (15.3.10) can be used:

$$\begin{aligned} E_R &= \frac{70 \times 4 \times \pi \times 20 \times 2}{2 \times (40 \times 10^3)^2} \\ &= 11 \mu \text{V/m} \end{aligned}$$

Equation (15.3.10) shows the importance of antenna heights. For example, doubling the height of the base-station antenna in the previous example will double the field strength at the receiving point. Note also that the field strength is proportional to E_0 , which, in turn, is proportional to the square root of the transmitted power. Doubling the power results in only a $\sqrt{2}$ increase in field strength.

Equations (15.3.9) and (15.3.10) also apply when a transmission takes place from the h_R antenna and is received at the h_T antenna: it is only necessary to change E_0 to the new value determined by the gain of the antenna at h_R and the power transmitted from there.

As the distance d increases, it becomes necessary to take into account the curvature of the earth. Reflection from the curved surface reduces both the amplitude of the reflected wave and the phase difference. These two effects tend to offset each other, and the resultant amplitude does not vary rapidly as a result.

Radio Horizon

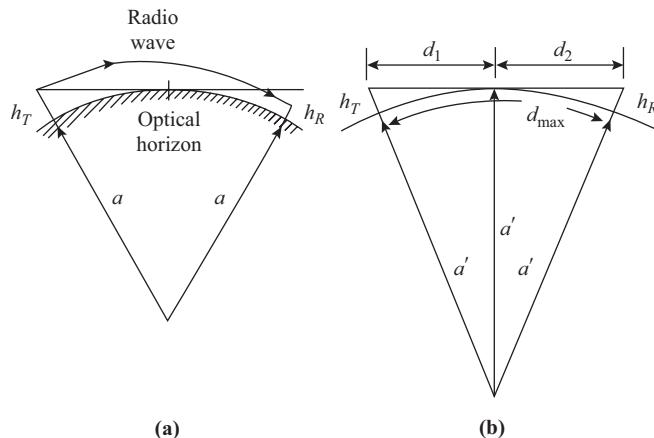
The curvature of the earth has a more important effect in that it presents a horizon that limits the range of transmission. This range is greater than the optical range because the effect of the earth's atmosphere is to cause a bending of the radio wave, which carries it beyond the optical horizon. Figure 15.3.2(a) shows a

typical radio-wave ray path, and Fig. 15.3.2(b) shows how the path can be considered straight by assigning a greater radius to the earth than it actually has. For standard atmospheric conditions the increase in radius has been worked out at $\frac{4}{3}$, so that

$$a' = \frac{4}{3}a \quad (15.3.11)$$

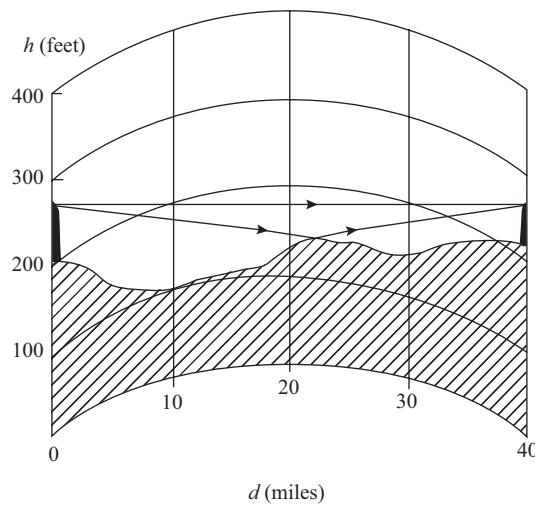
where a is the earth's actual radius and a' is the fictitious radius that accounts for refraction. From Fig. 15.3.2(b),

$$(a') + d_1^2 = (a' + h_T)^2$$



(a)

(b)



(c)

Figure 15.3.2 (a) Curvature of ray path resulting from change of refractive index of the air. (b) Equivalent straight line ray path for fictitious earth radius a' . (c) Example of a contour map for radio path planning.

Therefore,

$$d_1^2 = 2a'h_T + h_T^2 \quad (15.3.12)$$

But, since $a' \gg h_T$,

$$d_1^2 \approx 2a'h_T \quad (15.3.13)$$

Similarly,

$$d_2^2 \approx 2a'h_R \quad (15.3.14)$$

The maximum radio range d_{\max} is

$$\begin{aligned} d_{\max} &\approx d_1 + d_2 \\ &= \sqrt{2a'h_T} + \sqrt{2a'h_R} \end{aligned} \quad (15.3.15)$$

Substituting in the numerical values, $a' = \frac{4}{3} \times 3960$ miles, and expressing h_T and h_R in feet results in the useful expression

$$d_{\max}(\text{miles}) = \sqrt{2h_T(\text{ft})} + \sqrt{2h_R(\text{ft})} \quad (15.3.16)$$

Alternatively, in metric units,

$$d_{\max}(\text{km}) = \sqrt{17h_T(\text{m})} + \sqrt{17h_R(\text{m})} \quad (15.3.17)$$

EXAMPLE 15.3.2

Calculate the maximum range for a tropospheric transmission for which the antenna heights are 100 ft and 60 ft.

SOLUTION

$$\begin{aligned} d_{\max}(\text{miles}) &= \sqrt{200} + \sqrt{120} \\ &= 25.1 \text{ miles} \end{aligned}$$

The phenomenon of *diffraction* will extend the range in practice somewhat beyond the radio horizon.

Contour Maps

All the results derived so far are applicable only for smooth earth conditions (for example, transmission over water or over reasonably flat land). Where the earth's contour is rugged, a profile map is drawn to enable proposed transmission paths to be studied. Special graph paper is available for this purpose, in which the abscissa lines are curved to allow for the fictitious radius a' , and the graphs can be scaled for d in miles and heights in feet (or d in km and heights in meters). Figure 15.3.2(c) shows an example of a profile map.

Additional problems arise in built-up areas, where buildings and structures can cause multiple reflections and shielding, which are particularly troublesome with mobile radio equipment. The only satisfactory solution is to conduct field trials to determine an acceptable site for the base station.

Super- and Subrefractions

Irregularities in the earth's atmosphere also affect tropospheric transmissions. A condition known as *super-refraction* occurs when the refractive index of the air decreases with height much more rapidly than normal,

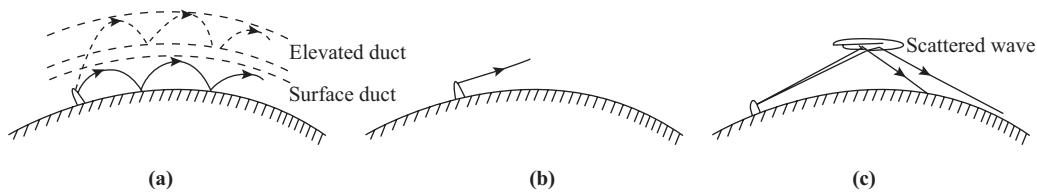


Figure 15.3.3 (a) Superrefraction. (b) Subrefraction. (c) Tropospheric scatter propagation.

so that the bending of the radio wave is much more pronounced than that shown in Fig. 15.3.2(a). The radio wave may then be reflected back from the earth to follow a path as sketched in Figure 15.3.3(a). In this way, the radio range is considerably increased. Unfortunately, the effect is not sufficiently reliable for it to be utilized for commercial communications systems, but it does account for some of the abnormally long distance interference that has been observed at VHF.

An increase of temperature with height (known as temperature inversion) gives rise to superrefraction, as does an increase of humidity with height. It is most noticeable when both of these effects occur together. The region in which superrefraction occurs is termed a *duct*, which can be formed both at the earth's surface and in elevated strata, as sketched in Fig. 15.3.3(a).

It is also possible for the opposite effects to occur, giving rise to *subrefraction*, which reduces signal strength by bending the ray away from the receiving point (Fig. 15.3.3[b]).

Inhomogeneities in the atmosphere can give rise to a scattering of radio signals, and by using highly directional high-gain antennas, and large transmitted power, reliable communication links well beyond the radio horizon can be established (Fig. 15.3.3[c]). The method is referred to as *tropospheric scatter propagation*. Ranges up to 400 miles, in the frequency band 40 to 4000 MHz, have been achieved.

Attenuation in the Atmosphere

The space wave is affected by atmospheric conditions, but only seriously at frequencies in the microwave region, above about 10 GHz. Heavy rain such as occurs in the tropics results in serious attenuation of electromagnetic waves at frequencies above about 10 GHz, while moderate rain, cloud, and fog will seriously attenuate electromagnetic waves at frequencies above about 30 GHz. Hail has little effect except at extra-high frequencies (above about 100 GHz), and the effect of snow is negligible at all frequencies.

The gas molecules in the air can result in attenuation of electromagnetic waves. Vibrational resonances in the water vapor molecule (H_2O) result in absorption peaks at wavelengths of 1.35 cm and 1.7 mm. The oxygen molecule (O_2) exhibits similar absorption peaks at 5 mm and 2.5 mm. Clearly, frequencies at which these resonances occur must be avoided in space-wave transmissions.

VHF/UHF Radio Systems

Propagation in the VHF and UHF bands between 30 MHz and 3 GHz takes place in the tropospheric mode. The major use of two-way radio communications in the VHF and UHF bands is communications between a fixed base station and several mobile units, located on vehicles, ships, or aircraft in the frequency band from 30 to 470 MHz. Typical applications are in control-tower-to-aircraft communication at airports, fire departments, ship control within harbors, police departments, armed-forces field operations, pipeline and transmission line maintenance, highway maintenance, taxicab and delivery vehicle dispatch, and personnel paging systems. Cellular telephone systems also use these frequencies. Since these systems operate in frequencies

above 30 MHz, their range of operation is limited to within the line-of-sight horizon of the base station (see Section 15.3), or that much further again if a repeater station is used. Large obstacles such as hills or tall buildings in an urban zone create shadows and odd reflection patterns that make complete coverage of the zone from a single base station difficult. For this reason, and to increase the horizon somewhat, it is usual practice to locate the base-station antenna on top of a high hill or building to gain additional height.

A limited number of channel assignments is available within the spectrum, mostly within the bands from 148 to 174 MHz and from 450 to 470 MHz. FM operation is preferred, and the maximum permissible channel spacing for this service has been progressively reduced from 120 kHz to the 15 kHz presently allowed, so that more channels can be assigned. Because of the narrow bandwidths used, the transmitters and receivers must be very stable and must maintain their operating frequency within ± 5 parts per million. Crystal control is a must if this type of stability is to be realized.

Dispatch systems for automobiles are usually required to cover as much area as possible, and omnidirectional vertically polarized antennas are usually used to accomplish this, both at the base station and in the mobile units. In some applications, such as pipeline and highway maintenance systems, the field of operations is strung out in a line over many miles, and for these systems vertically polarized multielement Yagi antennas aimed along the path are frequently used. This provides little coverage off the sides, but does provide better coverage along the line up to the horizon. The antennas used on the vehicles are nearly always short ground-whip antennas mounted on top of the vehicle. The longer whips used for the 50-MHz VHF band are not as popular.

Transmitter power in both the mobile units and the base-station units is usually limited to about 150 W, mainly because of the limited power available from the vehicle system. Voltage supplies for mobile equipment range from 12 V nominal for automobiles, 28 and 48 V for aircraft, and 48 V for railway locomotives. The base stations are usually operated directly from the 110-V, 60-Hz power mains, although for some applications backup battery power is also provided in case of power failures.

The transceivers are designed to alternately transmit and receive on the same frequency. For aircraft and ship control use and such systems as police and fire operation, the units may be designed to operate on one of several channels, with manual switching between channels provided. Each mobile unit is provided with a control head, which is usually separate from the main chassis, conveniently located near the operator. The control head provides a power on/off switch, audio volume control, muting threshold control, and a handset containing a microphone, a telephone receiver, and a push-to-talk switch. The base station may be self-contained and directly connected to an antenna near the operator's location, but usually the base-station transceiver unit is located with the antenna at a convenient high point and the operator is provided with a remote-control console. Connection between the base station and the control console is made by means of a pair of wires if the distance is short or a leased telephone line if the distance is considerable. A typical dispatch system is shown in Fig. 15.3.4(a).

Often it is found that it is impossible to cover the desired area from the single base-station location. In this case one or more additional base stations can be established. These may be connected back to the operator's console over wire lines and operated independently of each other by the operator. Alternatively, a radio repeater link can be established. This requires the use of a second frequency for the link between the main base station and the repeater station. Figure 15.3.4(b) shows an arrangement that might be used. Under normal operation, the base operator can communicate over the local base station on frequency 1 with any mobile unit within coverage area 1. When it is necessary to reach a vehicle in coverage area 2, he may turn off the local-area base station and turn on the repeater link on frequency 2. Now when the base operator transmits, he transmits on f₂ toward the repeater. The repeater receives control on f₂ and retransmits on f₁ to the extended-coverage area. When a mobile unit in the extended area transmits, it is received at the repeater on f₁ and retransmitted on f₂ toward the base station.

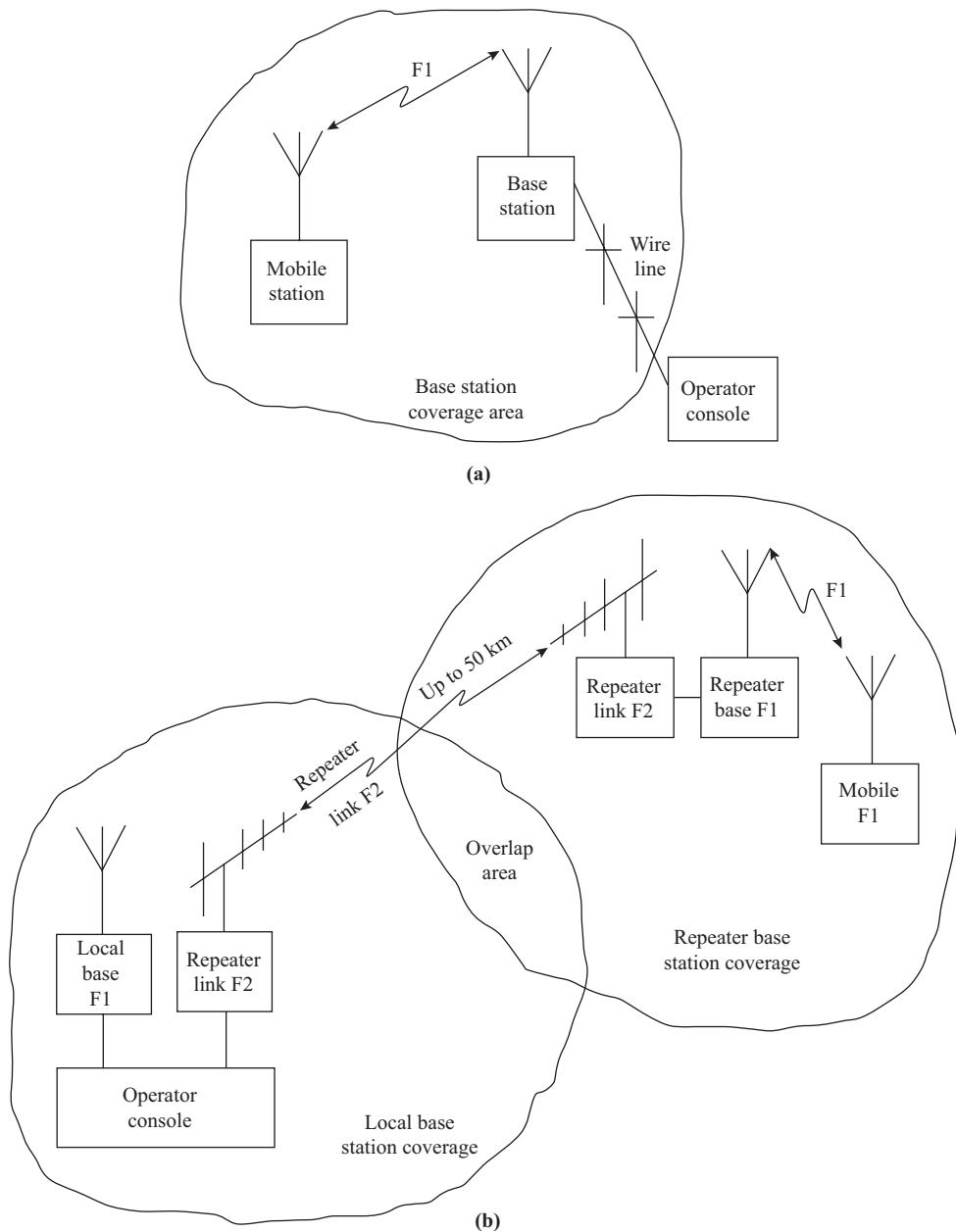


Figure 15.3.4 (a) Simple mobile dispatch system. (b) Dispatch system with a repeater station to extend coverage area.

When a system with several fixed stations is being operated in a repeater mode, the base operator must continuously monitor all the incoming signals from all the repeaters. If a mobile unit should be in an overlap area between two repeater stations, he may “key” both repeaters, causing interference at the control console. The operator must be able to purposely disable all but one repeater during a conversation. This can be

accomplished by sending a coded tone signal to turn off the desired repeater stations and then turn them back on when the conversation is complete. The detection characteristics of FM are such that a receiver located in the presence of two co-channel (same frequency) transmissions [such as f_1 base and f_1 repeater in Fig. 15.3.4(b)] will suppress the weaker signal. This is known as the *capture effect*.

15.4 Ionospheric Propagation

Ionospheric Layers

The upper reaches of the earth's atmosphere are ionized (that is, electrons are detached from atmospheric gas atoms), mainly as a result of receiving ultraviolet radiation from the sun, although other sources, such as cosmic rays, also contribute.

A state is reached where the free-electron density is maintained almost constant, the ionization rate being balanced by the recombination rate of electrons with positive ions. Clearly, the electron density will vary between day and night conditions and will also show a variation between winter and summer, as the ionization rate is dependent on solar radiation.

Various peaks are observed in electron density corresponding to the heights at which various gases settle in the upper atmosphere. The layers follow a meteorological classification, being known as the *C layer*, *D layer*, *E layer*, and *F₁* and *F₂* layers. Figure 15.4.1(a) shows some recently published results for typical electron-density distribution with height. It will be seen that for nighttime conditions only the F₂ layer remains. This is because in the lower layers collision processes in the denser gas atmosphere cause a more rapid recombination rate compared to the less dense F₂ region.

Figure 15.4.1(a) also shows the main ionizing radiations, and along the top of the graph is shown the plasma frequency f_N , an important parameter in radio communications via the ionosphere.

Plasma Frequency and Critical Frequency

When an electromagnetic wave enters an ionized region at vertical incidence, as shown in Fig. 15.4.1(b), the electric field acts as a force on the charged particles (electrons and ions), resulting in charge movement and hence current flow. Although a positive ion will carry the same magnitude of charge as an electron, it is more than 1000 times as massive and, therefore, its velocity will be correspondingly smaller and the ionic contribution to the current can be neglected.

The electron cloud will oscillate in the electric field of the wave, but with a phase retardation of 90° (for a sinusoidal wave) because of the electron mass inertia. This motion of the electron cloud produces a space current; in addition, the electric field has its own capacitive displacement current, which leads the field by 90°. The space current is therefore in phase opposition to the displacement current, and it appears to reduce the relative permittivity (dielectric constant) of the ionized medium, which is then given by

$$\epsilon_r = 1 - \frac{Nq_e^2}{\epsilon_0 m \omega^2} \quad (15.4.1)$$

where N = electron density, m^{-3}

m = electron rest mass = 9.11×10^{-31} kg

q_e = electron charge magnitude = 1.6×10^{-19} C

ω = angular frequency of wave, rad/s

ϵ_0 = permittivity of free space = 8.854×10^{-12} F/m

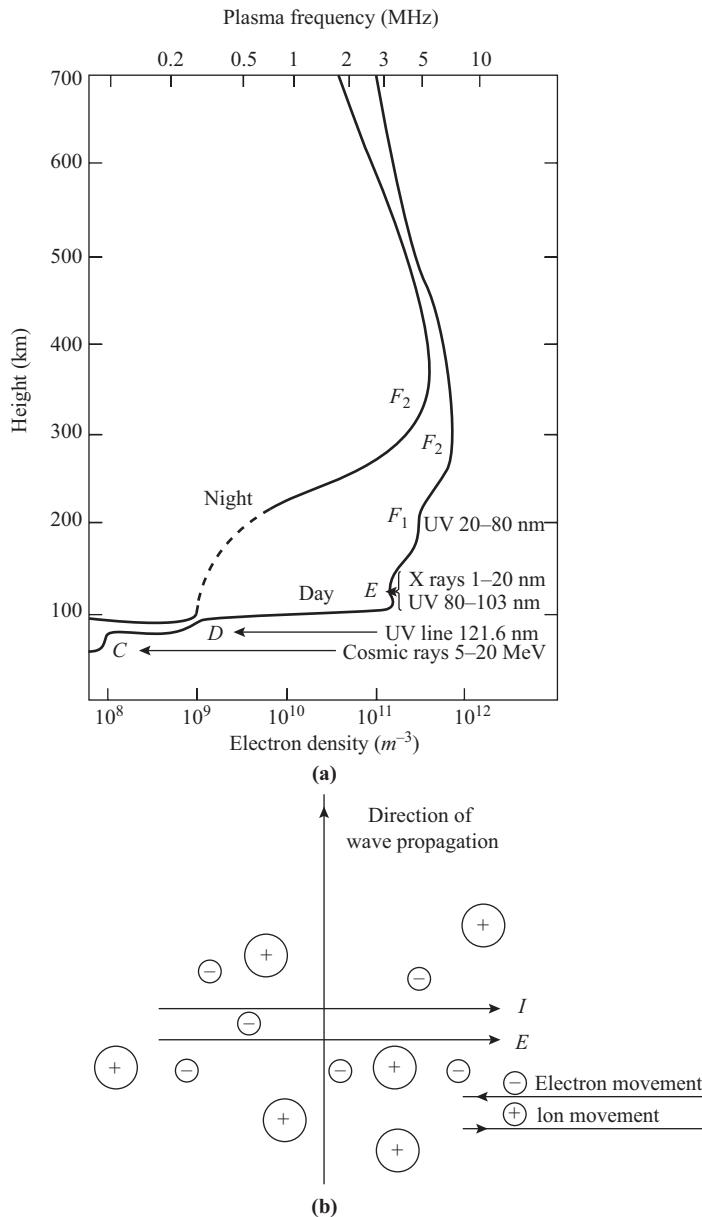


Figure 15.4.1 (a) Typical electron density distributions for summer noon and midnight conditions at mid-latitudes and the principal ionizing radiations at each level. Plasma frequency f_N is related to electron density m^{-3} by $f_N = 9\sqrt{N}$. The plasma frequency at the peak of a layer is called the critical frequency f_o . (Reprinted by permission of the IEEE.) (b) Electric field E and electric current density J vectors in an ionospheric layer.

The angular velocity of the wave can have a value that makes ϵ_r equal to zero, and this is termed the *plasma angular velocity*, ω_N . From Eq. (15.4.1), ω_N is seen to be

$$\omega_N^2 = \frac{Nq_e^2}{m\epsilon_0} \quad (15.4.2)$$

or

$$f_N^2 = \frac{Nq_e^2}{(2\pi)^2 m \epsilon_0} \quad (15.4.3)$$

Putting in the numerical values for the constants gives

$$f_N = 9\sqrt{N} \quad (15.4.4)$$

Equation (15.4.1) can now be rewritten as

$$\epsilon_r = 1 - \frac{f_N^2}{f^2} \quad (15.4.5)$$

The significance of f_N is that when a wave of this frequency reaches a region of electron density N , as given by Eq. (15.4.4), the relative permittivity, as given by Eq. (15.4.5), is seen to be zero. This, in turn, means that the total displacement current density is zero, and hence the *effective* electric field is zero. This can be accounted for in terms of a reflected wave that exactly cancels the incident wave at the point of reflection. Of course, it should be possible to receive this reflected wave, which is exactly what happens in short-wave radio communications via the ionosphere.

The highest-frequency wave that will be reflected from a given layer will be determined by the maximum electron density of that layer and will be given by

$$f_0 = 9\sqrt{N_{\max}} \quad (15.4.6)$$

Here f_0 is known as the *critical frequency*.

Phase and Group Velocities

An interesting point arises in connection with the velocities associated with the wave. The phase velocity has already been shown [Eq. (13.3.2)] to be

$$v_p = \frac{c}{\sqrt{\epsilon_r}} \quad (15.4.7)$$

Hence, in the ionosphere, when the wave reaches a height such that ϵ_r is zero, v_p becomes infinite! Now, as stated in Section 13.6, the energy in a wave travels at the group velocity v_g , and it can be shown that for an ionized layer

$$v_p v_g = c^2 \quad (15.4.8)$$

[This was shown in Eq. (14.2.10) to hold for the special case of a waveguide.] The wavefront in the ionosphere is a step function that will propagate energy at the group velocity, and hence, from Eq. (15.4.8), when the phase velocity is infinite, the group velocity is zero, so that the energy ceases to be propagated upward.

Secant Law and Maximum Usable Frequency

When a wave enters an ionized layer at an oblique angle of incidence i , it follows a curved ray path [Fig. 15.4.2(a)]. The phase velocity v_p at any given height can be determined by application of Snell's law of refraction for optics, the details of which will be omitted here. The result is that

$$\frac{\sin i}{c} = \frac{\sin r}{v_p} \quad (15.4.9)$$

Here r is the angle of refraction at the height where v_p occurs, as shown in Fig. 15.4.2(a).

At the apex of the path, $r = 90^\circ$, and therefore

$$v_p = \frac{c}{\sin i} \quad (15.4.10)$$

It can be seen that as the angle of incidence i approaches zero the phase velocity approaches infinity, in agreement with the results in the previous section.

As already shown,

$$v_p = \frac{c}{\sqrt{\epsilon_r}}$$

It follows, therefore, that

$$\sqrt{\epsilon_r} = \sin i$$

Substituting Eq. (15.4.5) for ϵ_r ,

$$\left(1 - \frac{f_N^2}{f^2}\right) = \sin^2 i$$

from which the reader should be able to derive that

$$f = f_N \sec i \quad (15.4.11)$$

This is known as the *secant law*. The highest frequency that can be used will be determined by N_{\max} , and hence f_0 ; this is known as the *maximum usable frequency* (MUF):

$$\text{MUF} = f_0 \sec i \quad (15.4.12)$$

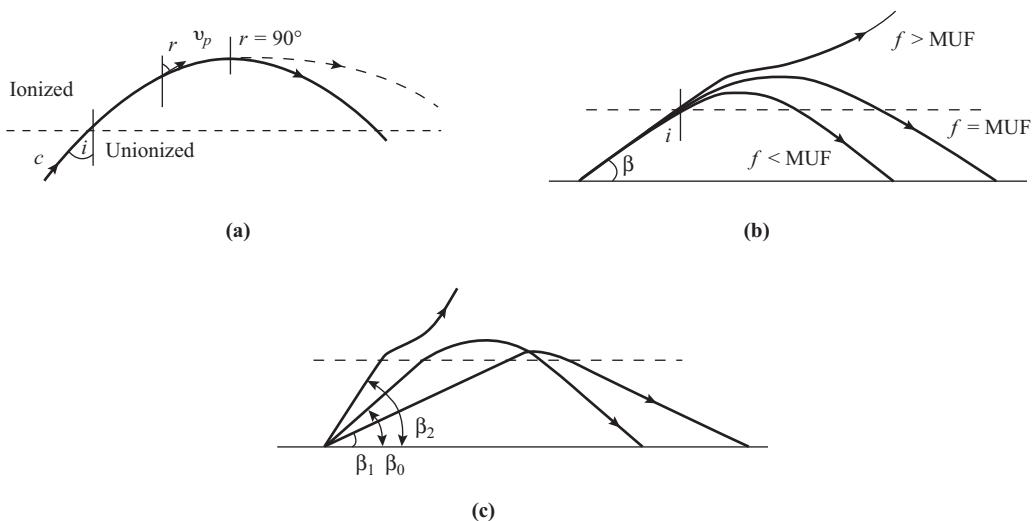


Figure 15.4.2 (a) Curved path of a wave in the ionosphere entering at oblique incidence, (b) Ray paths for fixed angle of incidence and varying frequency, (c) Ray paths for fixed frequency and varying angle of incidence.

It is possible for the wave to travel a considerable distance in the horizontal direction in the layer, as shown by the dashed ray path in Fig. 15.4.2(a), but irregularities in the ionized layer will eventually deflect the wave.

Figure 15.4.2(b) shows what happens when the angle of incidence is kept fixed and the frequency is varied. At frequencies lower than the MUF, the wave is reflected from a lower point in the layer. At higher frequencies than MUF, refraction is insufficient to return the wave to earth and it escapes through the layer (it may also be reflected from a more dense layer higher up).

Figure 15.4.2(c) shows what happens when the frequency is held constant while the angle of incidence is varied. Since in practice it would be the angle of elevation of the antenna beam (sometimes referred to as the takeoff angle) that would be varied, this is shown. The critical angle of elevation is shown as β_0 , and at this angle, f becomes the MUF. At angles less than critical, the wave is reflected from a lower region than N_{\max} , and at angles greater than critical, the wave escapes.

In the discussion so far, the curvature of the earth and the ionosphere have been ignored. This introduces little error where the ground distance between transmitter and receiver is less than about 1000 km. Beyond this, a correction factor has to be introduced in order to apply the secant law and the MUF equation.

Optimum Working Frequency

The frequency normally used for ionospheric transmissions is known as the *optimum working frequency* (OWF) and is chosen to be about 15% less than the MUF. It is desirable to use as high a frequency as possible, as the attenuation of the wave as it passes through the lower ionospheric layers is inversely proportional to the square of the frequency. This arises because, as the electrons collide with the gas molecules, they lose kinetic energy that they gained from the passing wave. (Electrons that do not collide return the energy periodically through reradiation in correct phase.) The kinetic energy of an electron of mass m is $\frac{1}{2}mv^2$, where v is the velocity. The velocity will be proportional to the time that the electric field acts in any given direction, and this will be proportional to the periodic time of the wave. But the periodic time is equal to l/f [Eq. (B.3)], where f is the frequency. Therefore, the kinetic energy lost is proportional to (l/f^2) .

The argument then is to use the highest possible frequency, which of course is the MUF. Irregularities in the ionosphere may, however, result in a MUF wave being occasionally deflected upward to escape through the layer. Practical experience has shown that frequencies about 15% lower than the MUF should be used.

Virtual Height

A wave traveling in a curved path has a horizontal component of group velocity v_h given by [see Fig. 15.4.3(a)]

$$v_h = v_g \sin r \quad (15.4.13)$$

From Eq. (15.4.8),

$$v_g = \frac{c^2}{v_p}$$

Therefore,

$$v_h = c^2 \frac{\sin r}{v_p}$$

Substituting for $(\sin r)/v_p$ from Snell's law, Eq. (15.4.9),

$$\begin{aligned} v_h &= c^2 \frac{\sin i}{c} \\ &= c \sin i \end{aligned} \quad (15.4.14)$$

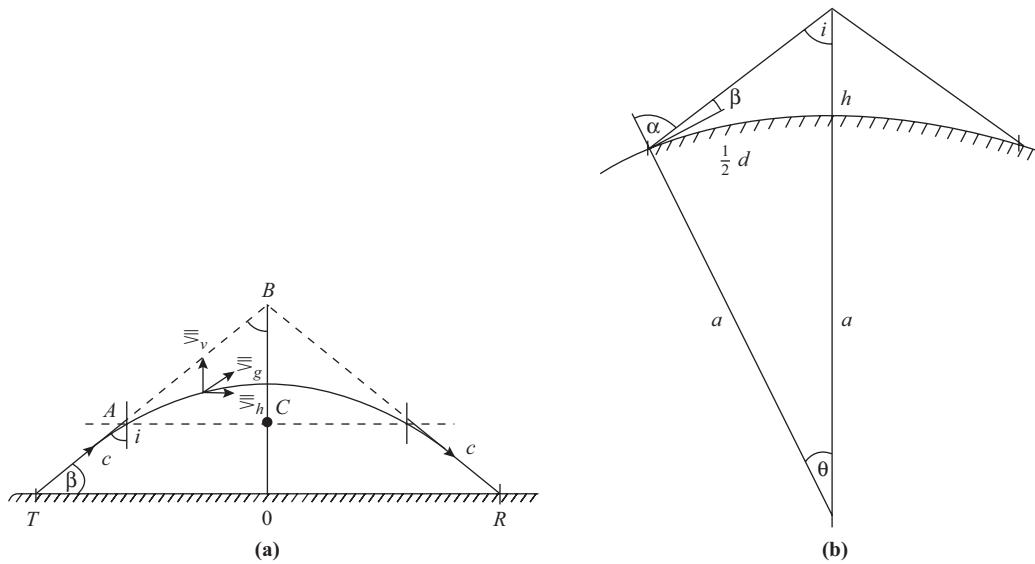


Figure 15.4.3 Determination of (a) virtual height and (b) path distance.

This shows that the horizontal component of group velocity is constant and independent of height in the ionized layer. Therefore, the time t required for the wavefront to reach the highest point in the path will be

$$t = \frac{AC}{v_h} \quad (15.4.15)$$

where AC is the horizontal distance from point of entry to the vertical dropped from the highest point [Fig. 15.4.3(a)].

Substituting from Eq. (15.4.14) for v_h , Eq. (15.4.15) becomes

$$t = \frac{AC}{c \sin i}$$

But from Fig. 15.4.3(a), $AC = AB \sin i$, and therefore

$$t = \frac{AB}{c} \quad (15.4.16)$$

Thus the wave can be considered to have traveled distance AB at constant velocity c in time t . The virtual height h of a layer is then OB [Fig. 15.4.3(a)], which is the height the wave appears to be reflected from had it been traveling at constant velocity c . The virtual height has the great advantage of being easily measured, and it is very useful in transmission-path calculations.

For flat-earth assumptions and assuming that the ionospheric conditions are symmetrical for the incident and reflected waves, the transmission-path distance TR , Fig 15.4.3(a), is

$$TR = \frac{2h}{\tan \beta} \quad (15.4.17)$$

Where the curvature of the earth is taken into account, the transmission-path distance can be determined from the geometry of Fig. 15.4.3(b):

$$\begin{aligned}\frac{\sin i}{a} &= \frac{\sin(180 - \alpha)}{a + h} \\ &= \frac{\sin \alpha}{a + h}\end{aligned}\quad (15.4.18)$$

Also,

$$180 - \alpha = 180 - (i + \theta)$$

Therefore,

$$i = \alpha - \theta \quad (15.4.19)$$

Substituting Eq. (15.4.19) in Eq. (15.4.18),

$$\frac{\sin(\alpha - \theta)}{a} = \frac{\sin \alpha}{a + h}$$

Therefore,

$$\theta = \alpha - \sin^{-1} \left(\frac{a}{a + h} \sin \alpha \right) \quad (15.4.20)$$

In terms of the angle of elevation, Eq. (15.4.20) becomes

$$\theta = (90 - \beta) - \sin^{-1} \left(\frac{a}{a + h} \cos \beta \right) \quad (15.4.21)$$

Also, from Fig. 15.4.3(b), the arc length

$$\frac{d}{2} = a\theta \quad (15.4.22)$$

where the angle θ is in radians. Substituting for θ from Eq. (15.4.21) and using radian measure for all angles,

$$d = 2a \left[\left(\frac{\pi}{2} - \beta \right) - \sin^{-1} \left(\frac{a}{a + h} \cos \beta \right) \right] \quad (15.4.23)$$

EXAMPLE 15.4.1

Calculate the transmission-path distance for an ionospheric transmission that utilizes a layer of virtual height 200 km. The angle of elevation of the antenna beam is 20° .

SOLUTION The flat-earth approximation [Eq. (15.4.17)], gives

$$\begin{aligned}d &= \frac{2 \times 200}{\tan 20^\circ} \\ &= 1100 \text{ km}\end{aligned}$$

Using Eq. (15.4.23),

$$d = 2 \times 6370 \left[(1.57 - 0.349) - \sin^{-1} \left(\frac{6370}{6570} \cos 20^\circ \right) \right]$$

= 966 km

Measurement of virtual height is usually carried out by means of an instrument known as an *ionosonde*. The basic method is to transmit vertically upward a pulse-modulated radio wave with a pulse duration of about 150 μs . The reflected signal is received close to the transmission point, and the time T required for the round trip is measured. The virtual height is then

$$h = \frac{cT}{2}, \quad \text{where } c = \text{speed of light} \quad (15.4.24)$$

The ionosonde will have facilities for sweeping over the radio-frequency range; typically, it will sweep from 1 to 20 MHz in 3 min. It will also have facilities for automatic plotting of virtual height against frequency, the resultant graph being known as an *ionogram* (Figure 15.4.4). The ionogram shows two critical frequencies, $f_o F_2$ and $f_x F_2$, which will be explained in the next section.

Effects of Earth's Magnetic Field

When a charged particle is displaced in a magnetic field, it experiences a force that causes it to move in a curved path. The magnetic field of the earth exerts such a force on the electrons in an ionized layer that they are displaced by the electric field of a radio wave. (There is also the magnetic field of the radio wave, but the force exerted by it is negligible in this situation.) In general, the electron paths will be helixes, as sketched in Fig. 15.4.5(a). At one particular wave frequency, known as the *gyrofrequency*, where the periodic time of the wave is equal to the time required for one complete revolution about the magnetic field axis, the electron path becomes a very wide single loop [Fig. 15.4.5(b)]. The gyrofrequency can be shown to be equal to

$$f_g = \frac{q_e}{2\pi m} B \quad (15.4.25)$$

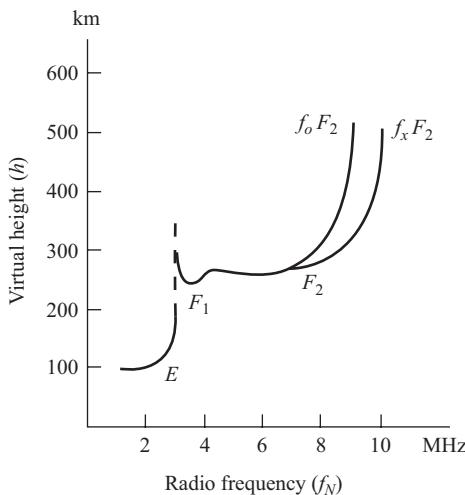


Figure 15.4.4 Sketch of an ionogram.

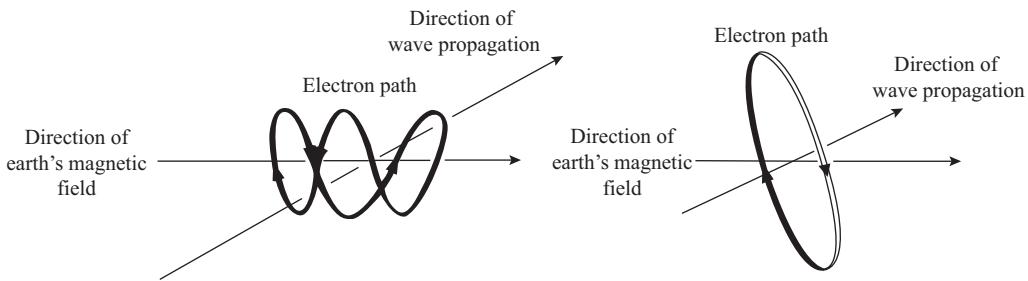


Figure 15.4.5 (a) Helical path of an electron moving in the earth's magnetic field. (b) Electron path at gyrofrequency.

where q_e is the charge of the electron, m is its mass, and B is the magnetic field strength. Substituting values for q_e and m and using an average value of 0.5×10^{-4} T for the earth's magnetic field results in a gyrofrequency of 1.4 MHz.

The significance of the gyrofrequency is that, because of the wide path followed by the electrons, the number of collisions between electrons and molecules in the D layer is increased, resulting in increased attenuation in the reflected wave at frequencies in the vicinity of f_g . Thus most of the medium-wave broadcast band suffers high attenuation of the reflected-wave component during the day, when the D layer is present.

Another effect of the earth's magnetic field is that the relative permittivity of the ionized layer develops two components. The details are complicated and will not be shown here, but the result is that two critical frequencies occur for an ionized layer. In practice, this only shows up for the F₂ layer, the two critical frequencies being known as the critical frequency for the ordinary ray, denoted by f_0F_2 , and the critical frequency for the extraordinary ray, denoted by f_xF_2 . These are shown in Fig. 15.4.4.

Service Range

The service range of a transmission of a given frequency is determined by the critical ray at the nearest point and the glancing ray at the farthest point, as shown in Fig. 15.4.6(a). Rays from the transmitting antenna at angles greater than the critical angle of elevation (or what is equivalent, at angles of incidence less than the critical value) will escape, creating a skip distance in which no signal is received.

The maximum possible range is reached when the critical ray coincides with the glancing ray [Fig. 15.4.6(b)], when the glancing ray is returned from the greatest virtual height h_m . The geometry of Fig. 15.4.6(b) is similar to that of Fig. 15.4.3(b) with β equal to zero and h being replaced by h_m . The angle θ is

$$\theta = \cos^{-1} \frac{a}{a + h_m}$$

and, from Eq. (15.4.22),

$$\begin{aligned} d &= 2a\theta \\ &= 2a \cos^{-1} \frac{a}{a + h_m} \end{aligned} \tag{15.4.26}$$

Equation (15.4.26) gives the maximum possible distance for a “single hop,” that is, one reflection involving the ionosphere. The range can be increased by using multiple hops, in which the wave is reflected from the earth after the first hop, to be reflected from the ionosphere once again farther on.

Of course, the ionospheric conditions at the point of reflection must be used in calculations to determine the transmission path. With multiple-hop transmissions the situation becomes more complicated, since

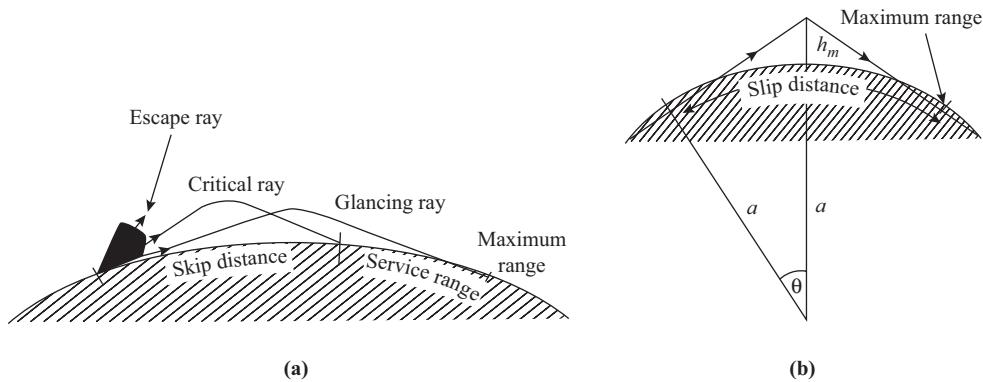


Figure 15.4.6 (a) Service range showing skip distance. (b) Maximum range.

two or more points must be taken into consideration. This is particularly so for paths following lines of latitude, since conditions may change from day to night over the path.

Irregularities in the ionosphere have been used successfully for *scatter propagation*, similar to that which occurs in the troposphere. With ionospheric scatter, distances on the order of 1000 miles are achieved at frequencies in the low end of the VHF band (between about 20 and 60 MHz). Fading effects are more troublesome in ionospheric scatter systems than in tropospheric scatter systems, which tends to limit *ionoscatter* propagation to narrow-band (for example, 3-kHz) signals.

Ionospheric Irregularities, Ionic Disturbances, and Fading

In previous sections, a very simple model of the ionosphere was assumed in which electron densities varied smoothly and uniformly, and changes resulting from diurnal and seasonal variations were assumed to be predictable. Calculations of MUF, virtual height, and so on, based on this model provide only an estimate for average conditions. In practice, the ionosphere exhibits irregularities in electron densities in the various layers, the extent of which ranges from less than 100 km to many hundreds of kilometers. It is known that some irregularities travel through the ionosphere with a horizontal component of velocity ranging from a few meters per second to greater than 1 km/s. These are known as *traveling ionospheric disturbances* (TIDs), which can seriously affect the accuracy of high-frequency direction finders.

Many of the causes of ionospheric disturbances are not well understood. Some of the factors thought to be involved are large-scale gravity waves in the atmosphere, electric currents and plasma instabilities in the ionosphere, and, in particular, solar activity. It has been observed that severe attenuation, or even complete loss of radio signals, accompanies solar flares. Intense x-ray radiation from the flares increases the ionization of the D layer, resulting in increased absorption, as described in Section 14.4. The fadeouts occur very suddenly and are known as *sudden ionospheric disturbances* (SIDs). They are also known as *Dellinger fadeouts* and *Mogel-Dellinger fadeouts*, having been reported by Mogel in Germany in 1930 and by Dellinger in the United States in 1937.

Protons are also discharged from the sun during solar flares. They affect the outer reaches of the earth's magnetic field about 30 hours after a flare has been observed and give rise to what is termed a *magnetic storm*. These storms adversely affect radio communications, particularly at high latitudes (near the magnetic poles). Magnetic storms not connected with solar flares, but associated with sunspots, have been recorded; yet others

have been observed apparently not connected with solar flares or sunspots, but which have a periodicity of recurrence of about 27 days, the rotational period of the sun. These appear to originate in well-defined regions of the sun (labeled M regions) and, in fact, are more noticeable at sunspot minima. They rise much more gradually than those associated with sunspots or solar flares, and although they are weaker, they persist much longer, sometimes up to about 10 days. Weak sporadic magnetic storms apparently not directly related to solar activity have also been observed.

Another form of ionospheric disturbances that can seriously affect radio communications is known as *sporadic E*. Thin, highly ionized layers occur in the E layer from time to time; the extent, position, and timing of these are all irregular, hence the term "sporadic E layer." Because of its high electron density, sporadic E can often take over a transmission normally beamed on the F layer; VHF reflections have also been observed from sporadic E, resulting in interference in the VHF television channels (for example, the "freak" reception of distant stations, which sometimes occurs).

Fluctuations in electron densities occur continuously in the ionosphere, giving rise to fluctuating phase differences in the various ray paths of a signal. Since the waves following these ray paths combine as phasors at the receiver, fluctuations in the received signal strength, called *interference fading*, will be observed. Another form of fading, known as *polarization fading*, occurs where the ordinary and extraordinary rays combine phasorally. The ordinary and extraordinary waves are always perpendicularly polarized with reference to each other, and because of the random variations in amplitude and phase of each, when they combine, the net polarization also varies in a random manner. The receiving antenna will of course be arranged for reception of a fixed-polarization wave (for example, one that is vertically polarized).

With modulated signals, fading can affect a very narrow range of the frequency spectrum independently of other parts of the spectrum, and this gives rise to what is termed *selective fading*. Selective fading comes about essentially because the ray paths in the ionosphere will be different for different frequencies, and they will not necessarily all experience a disturbance in a given region. Selective fading limits ionospheric transmissions to narrowband signals (for example, 3-kHz bandwidth).

Summary of Layers

In spite of the difficulties and uncertainties inherent in ionospheric transmission, it does provide an inexpensive and relatively quick means for setting up medium- to long-distance radio communications. Of course, it has been superseded on many circuits by satellite systems or submarine cables, both of which provide much better service, but at increased cost.

Methods are being developed for predicting ionospheric conditions that may be utilized in the planning of radio circuits. Equipment is also being developed that will automatically select optimum frequency for transmission in an effort to combat fading.

The main features of the various layers are summarized here, with values obtained from Fig. 15.4.1(a). Frequencies, heights, and so on, will of course vary considerably, as already discussed, and the values are presented simply to give an order-of-magnitude comparison for the various layers.

C and D layers. Virtual height, 60 to 80 km. Reflect low and very low frequencies (see Section 14.6), but for HF communications the D layer, in particular, introduces attenuation.

E layer. Virtual height, ~ 110 km. Critical frequency, ~ 4 MHz. Maximum single-hop range, ~ 2350 km.

F₁ layer. Virtual height, ~ 180 km. Critical frequency, ~ 5 MHz. Maximum single-hop range, ~ 3000 km.

F₂ layer. Virtual height, ~ 300 km daytime, ~ 350 km nighttime. Critical frequency, ~ 8 MHz daytime, ~ 6 MHz nighttime. Maximum single-hop range, ~ 3840 km daytime, ~ 4130 km nighttime.

HF Radio Systems

High-frequency radio systems using the bands from 1.6 to 30 MHz have been popular since the early days of radio in applications where relatively long distances must be covered at a low cost and under relatively light traffic conditions. Recent advances in satellite communications have usurped many of the traditional applications, but communications by satellite tend to be costly in terms of equipment, and HF radio will continue to be used for many applications.

The typical HF radio system is a two-point system with a transmitter and a receiver located at each end of the link. The transmitter and receiver used are usually a general-purpose communications set, although for some fixed applications single-frequency units can be used. Transceiver units are popular for temporary systems where one of the terminal points may be frequently moved to different locations.

The mode of transmission most often used for HF systems is single sideband, either with or without pilot carrier operated in the simplex mode. Independent sideband suppressed carrier transmission is useful where full duplex operation is desired with only one frequency assignment available. Separate transmitting and receiving antennas separated by some distance are required to prevent overloading of the local receiver by its associated transmitter. Dipoles are favored because they give some directional gain while being structurally simple. Fixed telephone stations may use more complex arrays, such as broadsides, to provide greater directivity. Transmitter powers ranging from a few watts to several kilowatts are commonly used.

The main advantages of HF systems lie in their ability to provide communication over great distances at relatively low cost. This makes them ideal for communications between remote settlements or camps, where the density of calls is low. HF systems do have several severe disadvantages, which have resulted in a search for other means of communications, stemming mainly from the overall unreliability of the systems resulting from variable propagation conditions (see Section 15.4). The problem of variable propagation conditions can be partially overcome by using frequency diversity, in which a given system is provided with several frequency assignments spanning the HF band of frequencies, so that the operator may choose the channel that gives the best results at any given time. However, this gives no protection against the total blackouts that periodically occur because of magnetic storms and other solar-induced phenomena.

Skip is also a problem. If HF stations are located closely enough together to use ground-wave propagation, distant stations frequently interfere through sky-wave propagation. Fading and distortion because of the changing ionospheric conditions cause annoying interference, especially in the longer circuits. Sometimes space diversity can be used to ease the problem, but since the diversity stations must be separated by several wavelengths, this becomes very expensive. In regions above the Arctic circle, auroral phenomena frequently make HF communications impossible by disturbing the necessary ionospheric layers. HF systems in equatorial regions are much more reliable.

Perhaps the most limiting disadvantage of the HF system is the fact that the bandwidth allowed for a channel is very narrow, sufficient for only one voice channel. This means that for every voice circuit a separate pair of radio circuits is required. Furthermore, since transmission is unreliable, the use of HF in public telephone systems is limited. In the past, HF radio was used extensively for overseas telephone communications. Recent advances in cable telephony and in satellite communications has just about eliminated this application except for some remote areas. It is still being used extensively in the South Pacific area.

15.5 Surface Wave

Mode of Propagation

A radio wave propagating close to the surface of the earth will follow the curvature of the earth as a result of the phenomenon of *diffraction*. This is the same phenomenon that causes sound waves, for example, to

travel around an obstacle. Diffraction effects depend on the wavelength in relation to the size of the obstacle and are greater the longer the wavelength. In the case of radio waves, the “obstacle” around which the wave must travel is the earth itself, and the surface wave is of importance at frequencies below about 2 MHz.

The conductivity and permittivity of the surface play an important part in the propagation of the surface wave, as the wave will introduce both displacement and conduction currents in the surface. These currents may penetrate to depths of from about 1 m at the highest frequency to tens of meters at the lowest, so the actual conditions on top of the surface are relatively unimportant.

The energy lost in the surface comes from the radio wave, which is therefore attenuated as it passes over the surface. The attenuation increases with increase in frequency, which is another factor that limits the usefulness of the surface wave to frequencies below about 2 MHz.

The surface wave in practice is always vertically polarized, as the conductivity of the ground would effectively short-circuit any horizontal electric field component.

The electric field strength as given by Eq. (15.2.13) can be modified to take into account the various factors affecting propagation by introducing an attenuating factor A , such that

$$E = \frac{A\sqrt{30P_TG_T}}{d} \text{ V/m} \quad (15.5.1)$$

It is normal practice to evaluate E for standard conditions of 1 kW radiated from a short unipole antenna, which has a power gain of 3. Equation (15.5.1) then becomes

$$E = 300 \frac{A}{d} \text{ V/m} \quad (15.5.2)$$

Graphs are available that present the factor A as a function of distance for various values of permittivity, conductivity, and frequency. At large distances (for example, $d > 100\lambda$), the factor A becomes inversely proportional to distance; the field strength plotted against distance is shown in Fig. 15.5.1(a) for a frequency of 500 kHz.

The change of refractive index with height of the atmosphere also causes diffraction of the surface wave, which can be allowed for by adopting an earth's radius of $\frac{4}{3}$ times the actual radius a . The main low-frequency diffraction is then determined for an obstacle of radius $4a/3$.

Ground Wave

At low enough frequencies that the height of the transmitting antenna above ground, in terms of wavelength, is small, the direct wave and the ground-reflected wave (both of which go to make up the space wave; see Section 15.3) effectively cancel each other, leaving only the surface wave. At higher frequencies the height of the antenna may be such that the space wave is comparable in magnitude to the surface wave, the resultant wave being the phasor sum. The resultant wave is termed the *ground wave* (not to be confused with surface wave alone). The situation is shown in Fig. 15.5.1(b). Normally, the ground-wave field strength is greater than that of the surface wave alone, and this is taken into account by introducing a multiplying factor, known as the *height-gain* factor in Eq. (15.4.26). The height-gain factor depends on the physical heights of the transmitting and receiving antennas and also on the factors that are used in the determination of the attenuation factor A . Like A , the height-gain factor is available in graphical form for a wide range of practical conditions.

Broadcast Fading Zone

The medium-wave (550 to 1600 kHz) broadcast service normally utilizes the surface wave. Some energy will, however, be transmitted into the ionosphere, where during daytime conditions it is almost completely absorbed in the D layer.

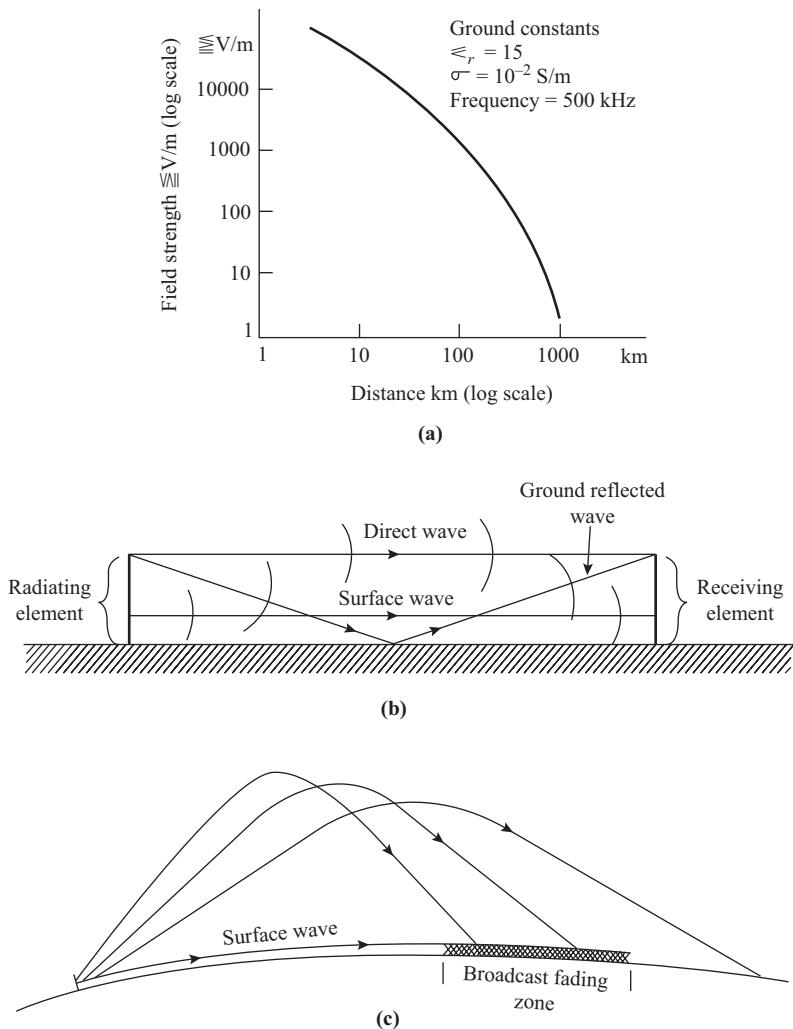


Figure 15.5.1 (a) Example of variation of surface wave field strength with distance, (b) Ground wave. (c) Broadcast fading zone.

During nighttime conditions, an appreciable portion of the ionospheric wave is returned to earth, extending the service area of the station well beyond that covered by the surface wave. There will also be a zone in which both the surface wave and the ionospheric wave are of the same order of magnitude, and the resultant signal strength is the phasor sum of the two. Unfortunately, fluctuations in the ionosphere produce fluctuations in the phase of the reflected wave relative to the surface wave, resulting in severe fading of the combined wave. The area in which this occurs is known as the *broadcast fading zone* and is shown in Fig. 15.5.1(c).

15.6 Low Frequency Propagation and Very Low Frequency Propagation

At frequencies below about 300 kHz it is convenient to consider propagation in two bands: 3 to 30 kHz, called the very low frequency (VLF) band, and 30 to 300 kHz, called the low frequency (LF) band. It is an

interesting point that the transatlantic radio links engineered by Marconi were at these low frequencies in the year 1901.

At frequencies in these bands, propagation is by means of surface waves up to distances of about 1000 km; beyond this, the sky wave plays an increasingly important part. The sky wave is propagated by means of multiple hops between the ionosphere and the earth. In the LF part of the band, theoretical treatment based on multiple hops of a TEM wave gives satisfactory results, while in the VLF band, the propagation is best considered to take place as a TM waveguide mode (see Chapter 14), the earth and the ionosphere forming the walls of a spherical waveguide.

Because of the low carrier frequencies and consequent narrow band-widths available, communications channels are limited to slow data rates. There are broadcast services of standard frequencies and time transmitted from various countries. The main characteristic of these low and very low frequencies is that they provide highly reliable radio links. Of course, the antenna structures have to be very large, and they are unfortunately inherently inefficient.

Radio navigational systems make extensive use of the low and very low frequency bands. The highly stable phase characteristics of the propagation allow the phase delay in a wave as a function of distance to be accurately known in advance, so that it can be measured by receiving equipment on the ship, aircraft, and so on, to determine position. The most common method is to use two or more transmissions to generate a pattern of *lines of position*, which can then be superimposed upon a map. Referring to Fig. 15.6.1(a), the signals received at position X , as given by Eq. (B.1), will be

$$A \sin(\omega_a t - \beta_a |s_1|) \quad \text{from transmitter } A \quad (15.6.1)$$

$$B \sin(\omega_b t - \beta_b |s_2|) \quad \text{from transmitter } B \quad (15.6.2)$$

By making ω_a and ω_b multiples of a common generating frequency ω_0 to which both are synchronized, it becomes possible to compare the phase of the two signals. The phase angles are $\beta_a |s_1|$ and $\beta_b |s_2|$, where the phase-shift coefficients β are as defined by Eq. (13.5.5). Let $\omega_a = 2\pi a f_0$ and $\omega_b = 2\pi b f_0$; then, by putting both signals through frequency dividers at the receiver, as shown in Fig. 15.6.1(b), and comparing phases, a measure of $|s_1| - |s_2|$ is obtained. The phase of the signals after frequency division will be

$$\text{phase of signal } A = \frac{\beta_a |s_1|}{a} \quad (15.6.3)$$

$$\text{phase of signal } B = \frac{\beta_b |s_2|}{b} \quad (15.6.4)$$

Now, $\beta_a = 2\pi/\lambda_a$ and $\lambda_a = \lambda_0/a$, so the phase angle of signals A and B , after frequency division, become

$$\text{phase of signal } A = \frac{2\pi}{\lambda_0} |s_1| \quad (15.6.5)$$

$$\text{phase of signal } B = \frac{2\pi}{\lambda_0} |s_2| \quad (15.6.6)$$

Therefore, the phase difference Δ is

$$\Delta = \frac{2\pi}{\lambda_0} (|s_1| - |s_2|) \quad (15.6.7)$$

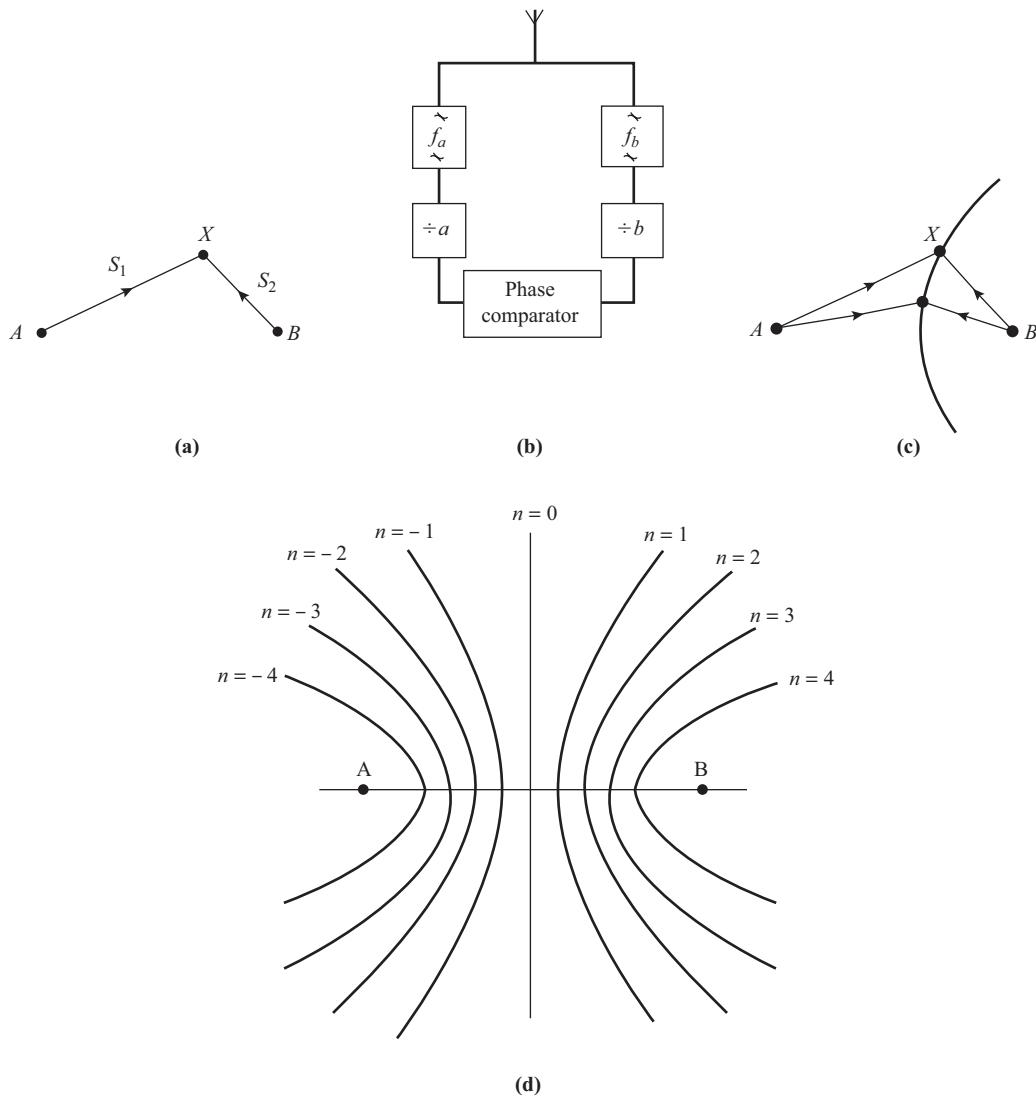


Figure 15.6.1 (a) Ray paths in a radio navigation system, (b) Comparison of signal phases, (c) Hyperbolic line of position, (d) Family of lines of position.

In effect, the phase difference goes through a series of maxima and minima with change in distance, just like that which occurs for standing waves on transmission lines (see Section 13.7). The two signals will be in phase when $\Delta = n2\pi$, where n is an integer. It follows, therefore, that

$$|s_1| - |s_2| = n\lambda_0 \quad (15.6.8)$$

For a given value of n , $|s_1| - |s_2|$ is a constant (since λ_0 is also fixed). Now, this is a property of a hyperbola [Fig. 15.6.1(c)], so that for a given value of n , a hyperbolic curve can be drawn along which $|s_1| - |s_2|$ is constant, and hence along which the signals are in phase. A whole family of curves can be

constructed, one curve for each value of n , as shown in Fig. 15.6.1(d). These curves are called *lines of position*. The space between lines of position are known, as *lanes*.

Along the base line joining A and B , the spacing is narrowest. Let L be the width of a lane on the base line. Then, on moving from line n to $n + 1$ along the base line, the total change in the path difference will be $2L$, since $|s_1|$ increases by L and $|s_2|$ decreases by L . In terms of wavelength, the change will be $(n + 1)\lambda_0 - n\lambda_0$. Hence

$$\begin{aligned} 2L &= \lambda_0 \\ L &= \frac{1}{2}\lambda_0 \end{aligned} \quad (15.6.9)$$

Equation (15.6.9) applies to the phase-comparison system shown in Fig. 15.6.1(b). In other systems the frequencies may be multiplied up to their lowest common multiple of f_0 , or the carrier may first be divided down to f_0 and then multiplied up to the slave frequency to which it is being compared. The term *master* is used to denote the transmitter to which the others are synchronized, while the others are known as *slaves*. In the *Decca navigational system*, the master is at $6f_0$, the red slave at $8f_0$, the green slave at $9f_0$, and the purple slave at f_0 . (The colors refer to the color of the grids superimposed on the navigation map.) By using more than one slave, two or more families of hyperbolae are made to intersect, enabling a position to be pinpointed, as in Fig. 15.6.2(a). The receiving equipment has to be able to integrate the phase change measured, in effect, to count the number of lines of position crossed as well as to indicate the position within a lane.

The Decca system works very well with surface-wave coverage, but phase variations resulting from the ionosphere reduce its accuracy where the sky wave has to be used. A worldwide radionavigational system named *Omega*, being developed by the U.S. Navy, operates in the VLF band. In this band, phase variations resulting from ionospheric changes are tolerable. Eight stations are involved, labeled A, B, \dots, H , and some of these are presently in operation, for example, A in Norway, B in Trinidad, and D in New York. The schedule of transmissions for the Omega System is shown in Fig. 15.6.2(b).

15.7 Extremely Low Frequency Propagation

Extremely low frequency (ELF) propagation is used to penetrate to great depths into the ground and the oceans. The impetus to its development has been the need for communications to submarines, especially those which form part of the nuclear deterrent force.

The ELF band extends from 30 to 300 Hz, the corresponding wavelength range being 10,000 (6210) to 1000 km (621 miles). Now, as pointed out in the introduction to Chapter 8, efficient radiation of electromagnetic waves requires that antenna dimensions be of the same order as the wavelength being radiated, and this condition cannot be met in the ELF range. Also, as seen in the studies of modulation, a modulated carrier wave requires a certain frequency bandwidth which is typically on the order of 1% of the carrier frequency. Thus, at ELF, this means that only 1 Hz of bandwidth is available for the information content. The rate of information transmittal is directly proportional to the bandwidth used, and therefore ELF does not offer the capacity for high rate of information transmission. The question then arises, why are ELF waves used for transmission purposes at all? The answer is that they provide the only practical means presently known of communications with submarines. Since this is primarily a matter of national defense, the problem of low efficiency antennas can be overcome by spending more money to increase power transmitted, and the problem of low communication rate can be overcome by just transmitting the code letters for a given message that is part of a list of standard messages aboard the submarine. For example, a bandwidth of 1 Hz

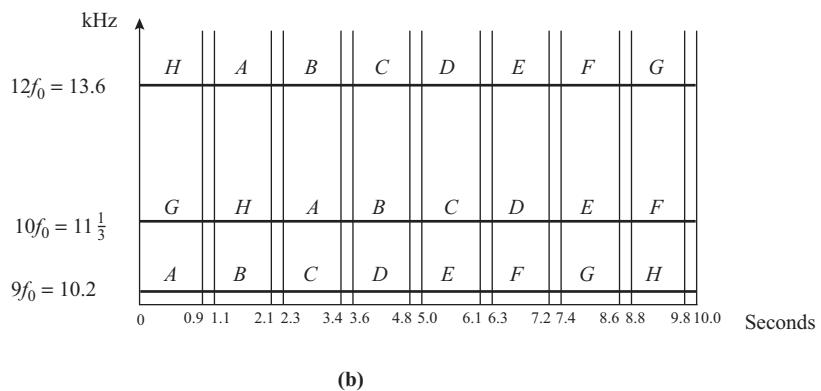
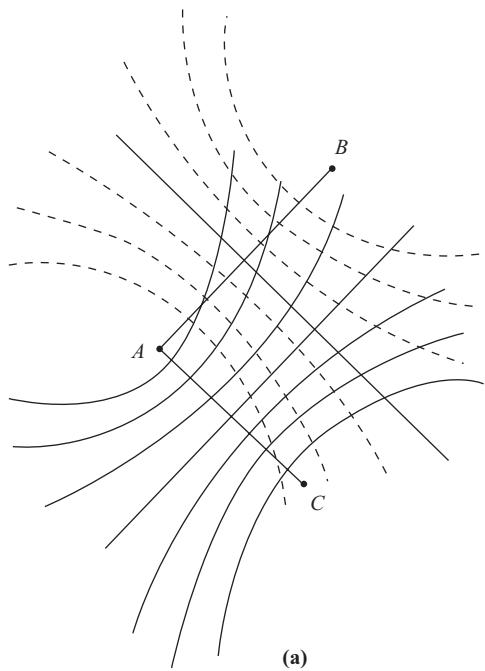


Figure 15.6.2 (a) Intersecting lines of position. (b) Chart showing the transmission schedule for the Omega navigation system.

allows a transmission rate of 1 bit/s. Ten bits allows $2^{10} = 1024$ messages to be stored, and the time required to transmit the necessary 10 bits in this example would be 10s.

As has been pointed out in the case of VLF propagation, the propagation around the earth may be considered to take place in a waveguide mode, the earth and the ionosphere forming the waveguide boundaries. Because of the very long wavelengths, the transmitting antenna has to be mounted horizontally. The length of the antenna is very much less than a wavelength. The situation is sketched in Fig. 15.7.1. It is found that the far field consists of a horizontal magnetic field and a vertical electric field, and propagation in the earth-ionosphere waveguide is by means of a quasi-TEM mode. A leakage field also occurs, directed into the surface of the earth, and subsurface communications utilize this leakage field. The leakage field is a plane TEM

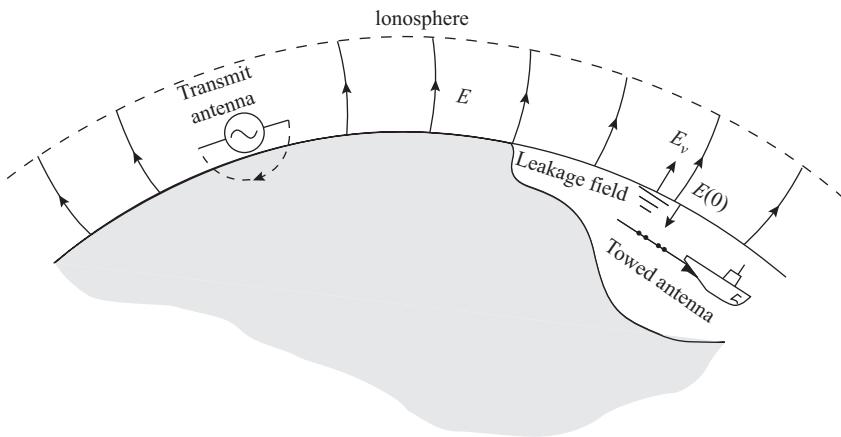


Figure 15.7.1. ELF propagation. (After Bernstein et al., "Long Range Communication at Extremely Low Frequencies," Proc. IEEE 62(3), 1974, pp. 292–312.)

wave that follows a propagation law similar to that of Eq. (13.5.2) for the voltage on a transmission line; that is, the electric field at depth z is related to $E(0)$, the electric field at the surface, by

$$E(z) = E(0)e^{-\gamma z} \quad (15.7.1)$$

The propagation coefficient γ is made up of the two components as given by Eq. (13.5.4), the attenuation coefficient α , and the phase shift coefficient β . These depend on the conductivity, permittivity, and permeability of the medium and on the frequency of the wave. The ratio $\sigma/\omega\epsilon$ is known as the loss tangent of the medium, and for $\sigma/\omega\epsilon \gg 1$, the attenuation coefficient is numerically equal to the phase shift coefficient, each being given by

$$\alpha = \beta = \sqrt{\frac{\omega\sigma\mu}{2}} \quad (15.7.2)$$

The permeability μ may be assumed to be equal to $4\pi \times 10^{-7}$ H/m.

EXAMPLE 15.7.1

Sea water has an average conductivity of 4 S/m and a relative permittivity of 80. Calculate the attenuation coefficient in dB/m for a signal of (a) 100 Hz and (b) 1 MHz.

SOLUTION (a) First it is necessary to evaluate the ratio $\sigma/\omega\epsilon$. At 100 Hz this is

$$\begin{aligned} \frac{\sigma}{\omega\epsilon} &= \frac{4}{2\pi \times 100 \times 80 \times 8.854 \times 10^{-12}} \\ &= 8.9 \times 10^7 \end{aligned}$$

Since this is very much greater than unity, Eq. (15.7.2) may be used to calculate the attenuation coefficient as

$$\alpha = \sqrt{\frac{2\pi \times 100 \times 4 \times 4\pi \times 10^{-7}}{2}}$$

$$= 3.974 \times 10^{-2} \text{ N/m}$$

Using the conversion factor [see Eq. (A.8)] $1\text{N} = 8.686 \text{ dB}$, we have

$$[\alpha] = 3.974 \times 10 \times 8.686 = \mathbf{0.345 \text{ dB/m}}$$

(b) At $f = 1 \text{ MHz}$, the loss tangent is $8.9 \times 10^7 \times 100/10^6 = 8.9 \times 10^3$, and since this is also very much greater than unity, Eq. (15.7.2) can be used in this case also. Thus

$$\alpha = 3.974 \times 10^{-2} \times \sqrt{\frac{10^6}{100}} = 3.974 \text{ N/m}$$

and

$$[\alpha] = 3.974 \times 8.686 = \mathbf{34.5 \text{ dB/m}}$$

An interesting graph of attenuation versus frequency for a plane TEM wave propagating in sea water is given in the book, *ELF Communications Antennas*, by Michael L. Burrows (Peter Peregrine Ltd., 1978). The graph is reproduced here in Fig. 15.7.2. It shows that the attenuation at ELF is a fraction of a decibel per meter, but it rises to the enormous attenuation of 1000 dB/m at about 1 GHz. This means, for example, that 3 mm of sea water will attenuate a signal at 1 GHz by 3 dB, or a factor of 2 : 1. At the visible end of the electromagnetic spectrum the attenuation decreases to a comparatively low value again, as shown in Fig. 15.7.2.

An important feature of ELF propagation is that noise at the surface will also be attenuated as it penetrates into the earth, so that the signal-to-noise ratio, as transmitted, is virtually independent of depth. Also,

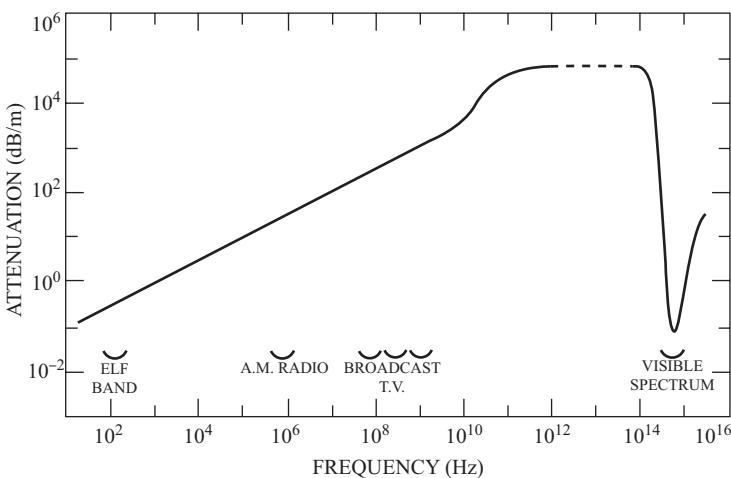


Figure 15.7.2 Attenuation of a plain electromagnetic wave in sea water as a function of frequency. (From Michael L. Burrows, *ELF Communications Antennas*, Peter Peregrine Ltd., 1978, Institute of Electrical Engineers, with permission.)

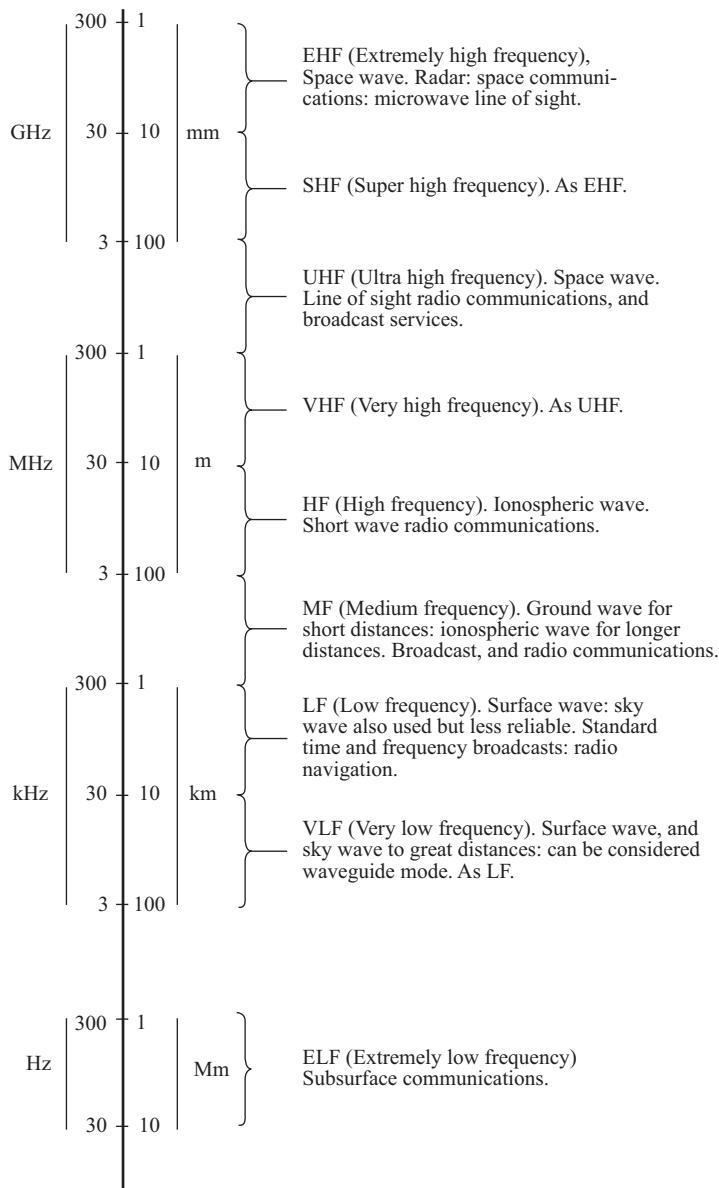


Figure 15.8.1 Classification of the various radio bands.

atmospheric noise is mostly lightning induced and consists of sharp spikes that are readily removed by simple peak limiters at the receiver. Sea water flattens out the peaks, and for such noise limiters to be fully effective it is necessary to include an “inverse ocean filter” in the receiver ahead of the limiters. Of course, in addition to atmospheric noise, noise will be generated in the receiver itself, and this will ultimately set a limit on receiver sensitivity.

As already mentioned, the transmitting antenna has to be very large, and such antennas have been built which cover many square miles of ground. It is found that the antenna wires, which are horizontally

mounted, may be buried as deep as 10 feet without noticeably affecting the radiation efficiency. The electric field increases with decreasing conductivity, and a low-conductivity ground is chosen for the antenna site. The signal strength is also influenced by the ionosphere height, and it must be kept in mind that the latter is only a small fraction of a wavelength at ELF. Transmitting antenna currents are on the order of 100 A, and some figures quoted in the paper by Bernstein et al. (Fig. 15.7.1) highlight the low efficiency of radiation. In one experiment quoted, the power supplied to the antenna was 3.88 MW. Of this, the percentages dissipated as heat were 43% in the conductors, 11% in the end-grounds for the antenna, and 46% in the ground return wire, the remainder, amounting to 69 W, being the power radiated.

At the frequency of operation (100 Hz), the attenuation was 1.5 dB per 1000 km, and at a depth of 10,000 km, a transmission rate of 1 bit/s was achieved with the 69 W radiated. Special coding and modulation techniques were used.

Although transmitting antennas have to be large, the same is not necessarily true for receiving antennas. Towed antennas consisting of two electrodes spaced 300 m along a towed cable have been used with submarines. The electrodes respond to the electric field gradient through the water. The cable itself is 600 m long, and the furthest electrode is 75 m from the end. Other types of receiving antennas are being experimented with, based on magnetometer principles. These are sensors that respond to the magnetic field of the wave. One of these has the acronym SQUID, for superconducting *quantum interference device*. The essential element in the SQUID is a small (2-mm diameter) loop of superconducting material. Any change in the incident magnetic field at the loop changes a superconducting current flowing in the loop, and in this way detection of the signal is achieved. Such an antenna requires a large amount of support equipment.

Because of the factors outlined, ELF transmission is really suitable only for one-way signaling, from a high-power ground station to the subsurface receiving station. The practical difficulties of employing large transmitting antennas and high-power transmitters at the subsurface station prevent transmissions being sent up to the surface. Also, atmospheric noise, which is attenuated in the downlink, increases toward the surface, adding to the difficulties of establishing a two-way communications link.

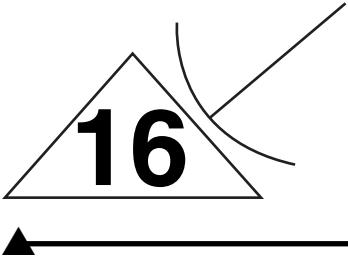
15.8 Summary of Radio-wave Propagation

The chart in Fig. 15.8.1 summarizes the classification of the various radio bands in terms of frequency and wavelength. It should be realized that sharp boundaries do not exist between the bands.

PROBLEMS

- 15.1.** Part of a microwave link can be approximated by free-space conditions. The antenna gains are each 40 dB, the frequency is 10 GHz, and the path length is 80 km. Calculate (a) the transmission path loss and (b) the received power for a transmitted power of 10 W.
- 15.2.** What feature of the microwave relay system makes it attractive for providing telephone trunk circuits? How many repeater stations would be required on a 6000-km route?
- 15.3.** A VHF radio link is set up between a shore station and an island in a lake 10 miles offshore. The antenna site on the island is atop a 100-ft cliff. Calculate the minimum height of the shore-station antenna if the minimum acceptable signal strength at either station is 10 $\mu\text{V}/\text{m}$. The frequency is 150 MHz and the transmit power is 1 W from a $\frac{1}{2}\lambda$ dipole from each station.
- 15.4.** Derive Eq. (15.2.14).

- 15.5. Why are VHF systems popular for dispatch work?
- 15.6. Why must a radio repeater station not retransmit on the same frequency on which it receives?
- 15.7. Explain the significance of the plasma frequency in connection with ionospheric radio communications. How is the critical frequency for an ionized layer related to the plasma frequency?
- 15.8. A radio communications link is to be established via the ionosphere. The maximum virtual height of the layer is 100 km at the midpoint of the path, and the critical frequency is 2 MHz. The distance between stations is 600 km. Determine suitable values for the optimum working frequency and the angle of elevation for the antenna main beam.
- 15.9. Explain what is meant by the gyro frequency and why frequencies in the region of the gyro frequency are not suitable for ionospheric transmissions.
- 15.10. Calculate the maximum range obtainable in a single-hop transmission utilizing the F₂ layer for which h_{\max} is 400 km. The earth's radius may be taken as 6370 km.
- 15.11. Discuss briefly the factors that give rise to fading in ionospheric radio transmissions.
- 15.12. List the advantages and disadvantages of HF radiotelephone circuits.
- 15.13. Explain the difference between the surface wave and the ground wave for radio transmissions in the frequency range from 300 kHz to 2 MHz.
- 15.14. Given that the free-space velocity of light c is 299,792.5 km/s and the ratio of phase velocity in the atmosphere to the free space values v_p/c is 0.9974, calculate the wavelength of the generating frequency f_0 in the Omega navigation system.
- 15.15. Compute and plot the variation in field strength, E , for a transmitting antenna with transmitted power $P_T = 10\text{ kW}$ and gain $G_T = 15\text{ dB}$. Let the distance d between transmitting and receiving antennae vary from 0.1 m to 500 m. (Hint: Plot equation 15.2.13 using MATLAB)
- 15.16. Using MATLAB/Mathematica, plot the *range of transmission* (d_{\max}) as a function of transmitter antenna height (h_T) and receiver antenna height (h_R) in meters. (Hint: Use the `surf()` function in MATLAB).
- 15.17. The *radio horizon* is expressed as: $D = \sqrt{2h}$, where D is the radio horizon in miles and h is the height of the antenna in feet. Plot, using MATLAB/Mathematica, the variation in D w.r.t. h , in SI units.
- 15.18. In the case of *space waves*, the *minimum usable frequency*, MUF , that leaves the atmosphere is stated as, $MUF = \frac{f_c}{\sin \theta}$, where θ is the angle of elevation above the horizon and f_c is the *critical frequency*. Plot MUF as a function of θ and f_c .



Antennas

16.1 Introduction

In a radio system, an electromagnetic wave travels from the transmitter to the receiver through space, and antennas (or aerials) are required at both ends for the purpose of coupling the transmitter and the receiver to the space link. Many of the important characteristics of a given antenna are identical for both transmitting and receiving functions, and the same antenna is often used for both.

Antennas may be constructed from conducting wires or rods, as, for example, the ordinary domestic TV antenna. At microwave frequencies apertures coupled to waveguides may be used; such antennas are naturally called *aperture antennas*. A horn antenna is an example of an aperture antenna. Antennas may be further classified as *resonant antennas*, in which the current distribution exists as a standing-wave pattern, and *nonresonant antennas*, in which the current exists as a traveling wave. Again, the ordinary TV antenna is an example of a resonant antenna, usually cut to one-half wave length, which gives it its resonant properties. Nonresonant antennas are used mainly for short-wave communications links and will be described later.

The types of structures used for antennas are many and varied, ranging from a simple length of wire suspended above the ground to the curtain arrays used for very low-frequency (VLF) broadcasting, from the insignificant-looking lens antenna on a traffic policeman's radar apparatus to the huge parabolic dish antennas of the astronomer's radio telescope. Several of these will be discussed in detail in this chapter.

16.2 Antenna Equivalent Circuits

In a radio communications link, the transmitting antenna is coupled to the receiving antenna through the electromagnetic wave. The arrangement is somewhat similar to the transformer coupling described in Section 1.7, except that with antennas the coupling is normally very weak, and an electromagnetic wave is involved rather than just the magnetic field, as in the case of transformers. Furthermore, the finite propagation time required for the wave to travel from the transmitting antenna to the receiving antenna can be significant. The antenna coupling system can, however, be represented as a four-terminal network as illustrated

in Fig. 16.2.1(a). The network representation is useful largely because it allows the well-known network theorems to be applied, and important general results can be obtained that are valid for any antenna.

Figure 16.2.1(a) shows antenna 1 transmitting. The input current is I_1 and the input voltage is V_1 . Thus the antenna impedance for transmitting is

$$Z_A = \frac{V_1}{I_1} \quad (16.2.1)$$

For simplicity, antenna terminals are shown. In the case of an aperture antenna fed by a waveguide, terminals do not of course exist in this sense, but impedance can still be measured in terms of the reflection coefficient as given by Eq. (13.11.2):

$$Z_A = Z_0 \frac{1 + \Gamma_A}{1 - \Gamma_A} \quad (16.2.2)$$

Here, Z_0 is the wave impedance of the waveguide, and the same equation can be used with transmission lines, with Z_0 the characteristic impedance of the line.

To find the impedance of antenna 1 in the receiving mode, Thévenin's theorem can be applied, and the Thévenin voltage equivalent generator found. The Thévenin voltage is the open-circuit voltage at antenna 1 terminals when antenna 2 is transmitting. The Thévenin impedance is found by shorting the emf source at antenna 2, applying a voltage at antenna 1's terminals, and measuring the resulting current. The Thévenin impedance is then the ratio of this voltage to current. Now, assuming that the coupling between the antennas is sufficiently weak that the short-circuited antenna 2 has no effect on the current in antenna 1, then the ratio of voltage to current at antenna 1 terminals in this case will be the same as given by Eq. (16.2.1). The antenna impedance is therefore the same for both transmitting and receiving. The equivalent circuit for transmitting is shown in Fig. 16.2.1(b), and that for receiving is shown in Fig. 16.2.1(c). In (b), the transmitter is shown as an equivalent voltage generator feeding an antenna impedance of Z_A , and in (c) the antenna is represented

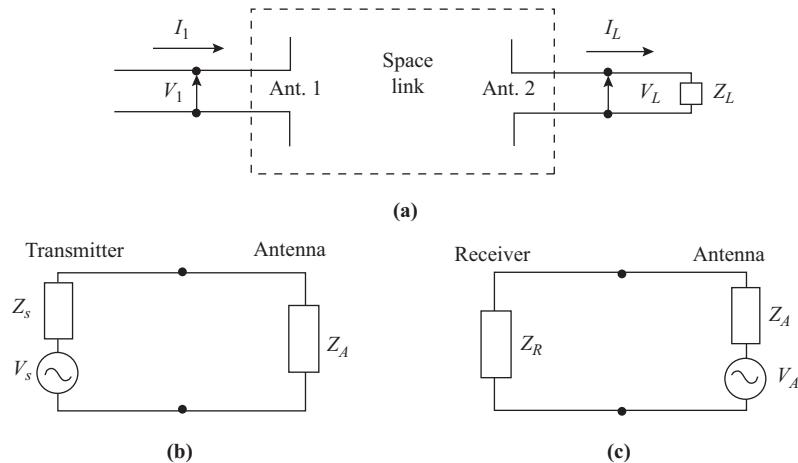


Figure 16.2.1 (a) Coupled antennas as a four-terminal network. (b) Equivalent circuit for the transmit antenna. (c) Equivalent circuit for the receive antenna.

by its Thévenin voltage generator equivalent circuit, feeding a receiver load impedance Z_R . The connection between antenna and transmitting or receiving equipment will be by means of feeder lines or guide, and the effect of mismatch on this is considered later.

A network theorem known as the *reciprocity theorem* can also be applied to the antenna system shown in Fig. 16.2.1(a). This theorem states, in essence, that if an emf E applied to the terminals of antenna 1 gives rise to a terminal current I at antenna 2, then applying E to the terminals of antenna 2 will give rise to the current I at the terminals of antenna 1. Now it is known that all practical antennas are directive; that is, they radiate better in some directions than others and receive better from some directions than others. A consequence of the reciprocity theorem is that the directive pattern for a given antenna will be the same for both the transmitting and the receiving modes of operating. Directivity is discussed in more detail in Section 16.7.

The antenna impedance Z_A is a complex quantity:

$$Z_A = R_A + jX_A \quad (16.2.3)$$

The reactive part X_A results from the reactive fields surrounding the antenna. As with any reactance, energy is stored in these fields and returned to source. Wherever possible, the reactance will be tuned out, so that the antenna presents a purely resistive load to the transmission line. The resistive part R_A is given by

$$R_A = R_{\text{loss}} + R_{\text{rad}} \quad (16.2.4)$$

The resistance R_{rad} is a fictitious resistance termed the *radiation resistance*, which, if it carried the same rms terminal current as the antenna, on transmission would dissipate the same amount of power as was radiated. A certain amount of power will be dissipated in the antenna as heat, and the power dissipated in R_{loss} , when carrying the same current as R_{rad} , gives the power lost in this way. The resistance R_{loss} therefore represents the losses in the antenna. The concepts of loss resistance and radiation resistance are most useful with wire antennas, for which the terminal currents are easily identified and the loss resistance is mainly the resistance of the antenna wire. For this type of antenna, let I be the rms terminal current. Then the total power supplied to the antenna is $I^2 R_A$, and the power radiated is $I^2 R_{\text{rad}}$. The antenna efficiency is therefore

$$\eta_A = \frac{I^2 R_{\text{rad}}}{I^2 R_A} = \frac{R_{\text{rad}}}{R_A} \quad (16.2.5)$$

In the receiving mode, the efficiency is defined as the ratio of power delivered to a matched load from the actual antenna to the power delivered to a matched load from the antenna with R_{loss} assumed equal to zero. Applying the maximum power transfer theorem (Section 1.11) to the receiving antenna circuit of Fig. 16.2.1(c), for the real antenna the maximum power is $V_s^2/4R_A$, and for the lossless antenna it is $V_s^2/4R_{\text{rad}}$. Thus the receiving efficiency is also given by Eq. (16.2.5).

Antenna matching to the feeder line is important for eliminating reflected waves and obtaining maximum power transfer. Examples of matching circuits are given in Section 13.12, and, in general, the matching network has to provide both reflectionless matching and matching for maximum power transfer. A matching network designed to meet one of these conditions automatically satisfies the other condition. Thus, as shown in Fig. 16.2.2(a), the impedance seen looking into the network at the feeder side is Z_0 , and at the antenna side Z_A^* , where $Z_A^* = R_A - jX_A$ is the complex conjugate of Z_A . This is required for maximum power transfer as described in Section 1.14.

In effect, the matching network tunes the antenna to resonance and then transforms the resistive part to Z_0 and can therefore be represented by the general arrangement shown in Fig. 16.2.2(b). Working from antenna to transmission line, the reactance $-jX_A$ tunes out the $+jX_A$ component of antenna impedance, and the transformer transforms the remaining R_A component to Z_0 . Now, assuming that the transmission line is

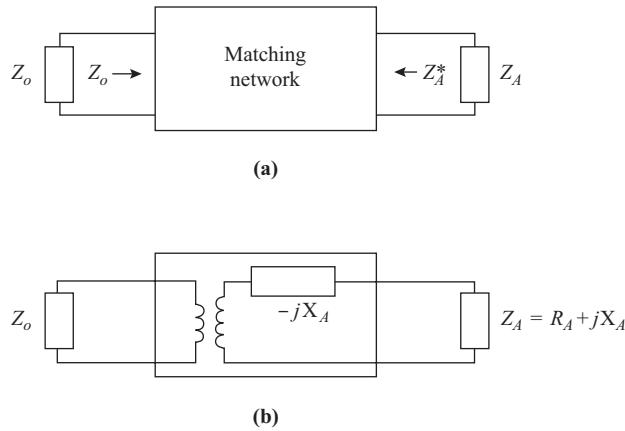


Figure 16.2.2 (a) Terminal impedances of a matching network providing reflectionless matching and maximum power transfer. (b) Circuit that meets the general impedance transformations of (a).

matched at its far end, the matching network sees an impedance Z_0 , and working from line to antenna, the transformer transforms this Z_0 back into R_A . This R_A is in series with the $-jX_A$ element, and therefore the output impedance is $R_A - jX_A$, or Z_A^* , as required for maximum power transfer to Z_A .

The antenna feeder should be matched at both ends as shown in Fig. 16.2.3. This provides maximum power transfer from the transmitter or into the receiver. It also provides reflectionless matching for the line, which prevents multiple reflections occurring should the line be mismatched at the antenna end.

Consider now the effect of a mismatch at the antenna end. Under transmitting conditions the transmitter and line appear as an emf source of internal resistance Z_0 feeding a load Z_A as shown in Fig. 16.2.4(a). The current flowing in this circuit is $V_0^2/|Z_0 + Z_A|^2$ and the power delivered to Z_A is $R_A V_0^2/|Z_0 + Z_A|^2$. The power delivered under matched conditions would be $V_0^2/4Z_0$, and therefore the matching efficiency can be written as

$$\eta_T = \frac{R_A V_0^2}{|Z_0 + Z_A|^2} \frac{4Z_0}{V_0^2}$$

$$= \frac{4R_A Z_0}{|Z_0 + Z_A|^2} \quad (16.2.6)$$

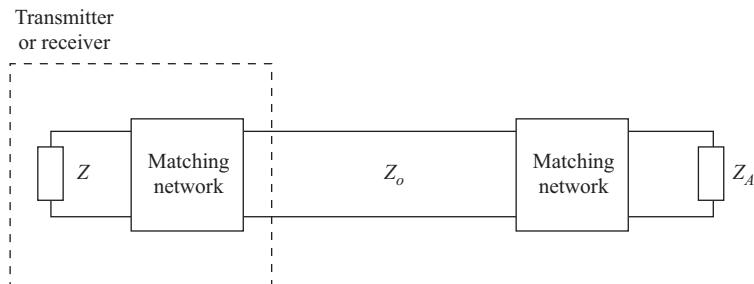


Figure 16.2.3 Antenna feeder matched at both ends.

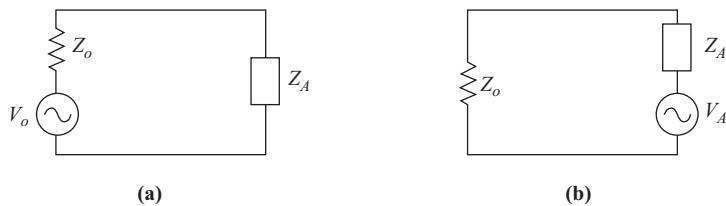


Figure 16.2.4 Mismatch conditions: (a) transmit and (b) receive.

Under receiving conditions [Fig. 16.2.4(b)], the power delivered to Z_0 is $Z_0 V_A^2 / |Z_0 + Z_A|^2$, and the power that would have been delivered under matched conditions is $V_A^2 / 4R_A$. Hence, the matching efficiency in the receiving case is also

$$\eta_\Gamma = \frac{Z_0 V_A^2}{|Z_0 + Z_A|^2} \frac{4R_A}{V_A^2}$$

$$= \frac{4R_A Z_0}{|Z_0 + Z_A|^2}$$

Thus the matching efficiency is the same for both transmitting and receiving conditions. Using the relationship given by Eq. (16.2.2) and the fact that $R_A = \frac{1}{2}(Z_A + Z_A^*)$, the matching efficiency is given by

$$\eta_{\Gamma} = 1 - |\Gamma_A|^2 \quad (16.2.7)$$

16.3 Coordinate System

The directional characteristics of an antenna are usually described in terms of spherical coordinates, shown in Fig. 16.3.1. The antenna is imagined to be at the center of a sphere, and any point P on the surface of the sphere can be defined in relation to the antenna by the radius d and the angles θ and ϕ . These are shown with reference to the rectangular coordinates x , y , and z . Also shown in Fig. 16.3.1 is the *equatorial plane*, which

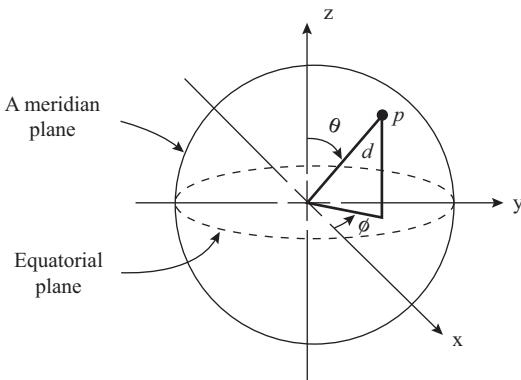


Figure 16.3.1 Spherical coordinates for point P on the surface of the sphere are radius d and angles θ and ϕ .

is the xy plane. Once the equatorial plane has been defined, any plane at right angles to it that contains the center of the sphere is known as a *meridian plane*. In practice, the equatorial plane and one of the meridian planes will be defined by the planes of symmetry for the antenna. Examples will be given later.

16.4 Radiation Fields

An electric current in a wire is always surrounded by a magnetic field. When the current is alternating, the free electric charges in the wire are accelerated, which gives rise to an alternating electromagnetic field, which travels away from the wire in the form of an electromagnetic wave. (An analogy may be made by considering a semirigid sheet being moved at constant velocity, which results in a steady air displacement. If the sheet is vibrated, which in effect imparts acceleration to some areas of it, a sound wave will be generated that travels through the air away from the sheet.)

The total field originating from an alternating current in a wire is complicated, consisting of (1) an electric field component that lags the current by 90° and that decreases in amplitude as the cube of the distance; (2) an electromagnetic field (a combined electric and magnetic field) that is in phase with the current and that decreases in amplitude as the square of the distance; and (3) an electromagnetic field that leads the current by 90° and that decreases in amplitude directly as the distance increases. Only the latter electromagnetic field reaches the receiver in a normal radio communications system, where it appears to the receiving antenna as a plane transverse electromagnetic (TEM) wave. The basic properties of a TEM wave are discussed in Appendix B. A useful rule of thumb is that for antennas for which the largest dimension D is very much greater than the wavelength being radiated ($D \gg \lambda$), the far-field zone becomes the only significant one for distances d greater than $2D^2/\lambda$:

$$d \gtrsim \frac{2D^2}{\lambda} \quad (16.4.1)$$

16.5 Polarization

In the far-field zone, the *polarization* of the wave is defined by the direction of the electric field vector in relation to the direction of propagation. *Linear polarization* is when the electric vector remains in the same plane, as shown in Fig. 16.5.1(a). A linear polarized wave that is propagated across the earth's surface is said to be *vertically polarized* when the electric field vector is vertical and *horizontally polarized* when it is parallel to the earth's surface. For example, in North America, television transmissions are horizontally polarized, and it will be observed that receiving antennas are also horizontally mounted, whereas in the United Kingdom vertical polarization is used, and there antennas are mounted vertically.

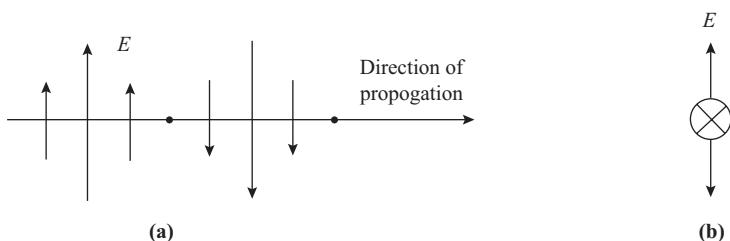


Figure 16.5.1 Linear polarization as viewed (a) on the axis of propagation and (b) along the direction of propagation.

In certain situations the electric vector may rotate about the line of propagation. This can be caused, for example, by the interaction of the wave with the earth's magnetic field in the F_2 layer of the ionosphere. Rotation of the electric vector can also be produced by the type of antenna used, and this effect is put to good use in satellite communications, as described in Chapter 19. The path traced out by the tip of the electric vector may be an ellipse, as illustrated in Fig. 16.5.2, in which case it is referred to as *elliptical polarization*. If the rotation is in a clockwise direction when looking along the direction of propagation, the polarization is referred to as *right-handed*; if it is anticlockwise, it is called *left-handed*. In Fig. 16.5.2(a) the direction of propagation is into the paper, so the polarization is right-handed. A special case of elliptical polarization is *circular polarization*, as illustrated in Fig. 16.5.3, and both right-handed and left-handed circular polarization are used in satellite communications systems as described later. Linear polarization can also be considered to be a special case of elliptical polarization. As shown in Fig. 16.5.2, elliptical polarization can be resolved into two linear vectors, E_x and E_y . Linear polarization results when one of these components is zero.

To receive a maximum signal, the polarization of the receiving antenna must be the same as that of the transmitting antenna, which is defined to be the same as that of the transmitted wave. For example, a wire dipole antenna, illustrated in Fig. 16.5.4(a), will radiate a linear polarized wave. A similar receiving dipole must be oriented parallel to the electric vector for maximum reception. If it is at some angle ψ , as illustrated in Fig. 16.5.4(b), then only the component of the electric field parallel to the receiving antenna will induce a signal in it. This component is $E \cos \psi$, and therefore the polarization loss factor is

$$\text{plf} = \cos \psi \quad (16.5.1)$$

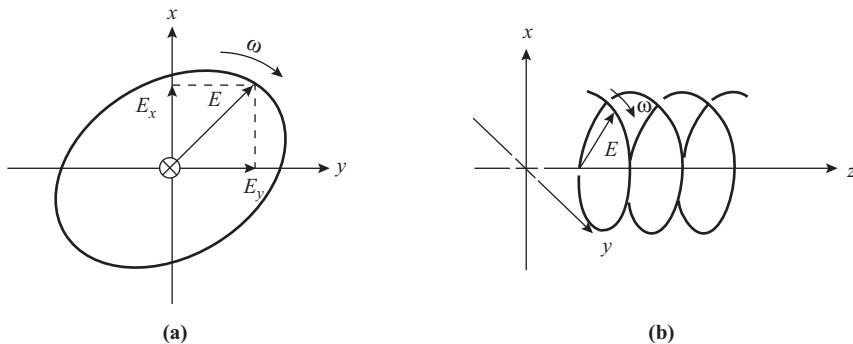


Figure 16.5.2 Elliptical polarization viewed (a) along the direction of propagation and (b) on the axis of propagation.

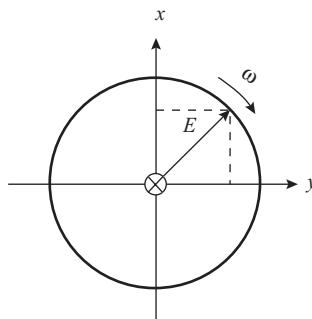


Figure 16.5.3 Circular propagation.

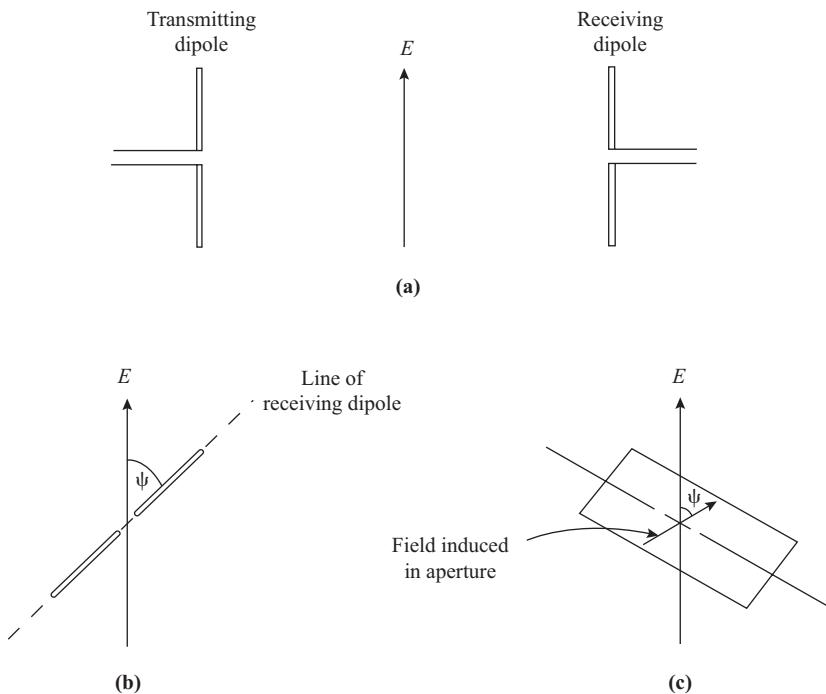


Figure 16.5.4 (a) Two dipoles aligned with the same polarization. (b) Receiving dipole in the same plane as E but with polarization misaligned. (c) Incoming wave E in same plane as aperture, but aperture polarization misaligned.

A similar situation can exist with an aperture antenna as illustrated in Fig. 16.5.4(c). The angle ψ is the angle between the induced field in the aperture and the incoming electric field, and the polarization loss factor is also $\cos \psi$ in this case. In both cases the direction of propagation is normal to the plane of the antenna.

16.6 Isotropic Radiator

The word *isotropic* means “equally in all directions,” so an isotropic radiator is one that radiates equally in all directions. A star is an example of an isotropic radiator of electromagnetic energy, but on a more practical level, all real antennas radiate better in some directions than others and cannot be isotropic. However, the concept of an isotropic radiator is a very useful one and provides a standard to which real antennas can be compared. Furthermore, since this is a hypothetical radiator, it may be assumed lossless; that is, its efficiency is unity. Let P_s represent the power input to a lossless isotropic radiator. Then since its efficiency is unity, this is also the power radiated. Consider this antenna at the center of the sphere shown in Fig. 16.3.1. Then, since any sphere has a solid angle of 4π steradians at its center, the power per unit solid angle is

$$P_i = \frac{P_s}{4\pi} \text{ W/sr} \quad (16.6.1)$$

This quantity is used as a standard to which real antennas can be compared. Another useful quantity is the power density. The surface area of a sphere of radius d is $4\pi d^2$, and therefore the power density for the lossless isotropic radiator is

$$P_{Di} = \frac{P_s}{4\pi d^2} \text{ W/m}^2 \quad (16.6.2)$$

It can be seen that the power density and the power per unit solid angle are related by

$$P_{Di} = \frac{P_i}{d^2} \quad (16.6.3)$$

16.7 Power Gain of an Antenna

For any practical antenna, the power per unit solid angle will vary depending on the direction in which it is measured, and therefore it may be written generally as a function of the angular coordinates θ and ϕ as $P(\theta, \phi)$. The power gain of the antenna is then defined as the ratio of $P(\theta, \phi)$ to the power per unit solid angle radiated by a lossless isotropic radiator. The gain function, denoted by $G(\theta, \phi)$, is

$$\begin{aligned} G(\theta, \phi) &= \frac{P(\theta, \phi)}{P_i} \\ &= \frac{4\pi P(\theta, \phi)}{P_s} \end{aligned} \quad (16.7.1)$$

The gain function is a very important antenna characteristic that can be measured or, in some cases, calculated, and some examples will be given later.

For most antennas, the gain function shows a well-defined maximum, which will be denoted by G_M , and the radiation pattern of the antenna is

$$g(\theta, \phi) = \frac{G(\theta, \phi)}{G_M} \quad (16.7.2)$$

The radiation pattern is seen to be simply the gain function normalized to its maximum value. The maximum value G_M is referred to as the *gain* of the antenna, but this is only a gain in the sense that the antenna concentrates or focuses the power in the maximum direction. It does not increase the total power radiated.

Closely associated with the power gain is the *directive gain* of the antenna. This is the ratio of $P(\theta, \phi)$ to the average power per unit solid angle radiated by the *actual* antenna and is denoted by $D(\theta, \phi)$. The average power per unit solid angle is $\eta_A P_s / 4\pi$, where η_A is the antenna efficiency and P_s is the power input, as before. Thus the average is seen to be equal to $\eta_A P_i$, and therefore the directivity is related to power gain by

$$D(\theta, \phi) = \frac{G(\theta, \phi)}{\eta_A} \quad (16.7.3)$$

In particular, the maximum value of $D(\theta, \phi)$ is termed the *directivity*, or *directive gain*, given by

$$D_M = \frac{G_M}{\eta_A} \quad (16.7.4)$$

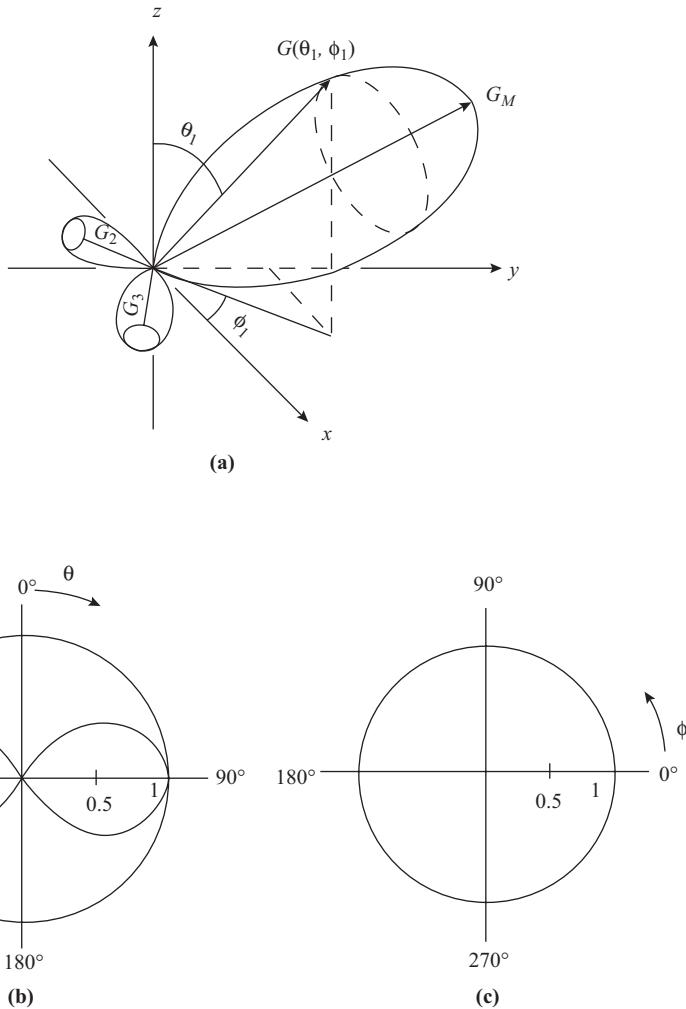


Figure 16.7.1 (a) Gain function $G(\theta, \phi)$. (b) Polar diagram of function $g(\theta)$. (c) Polar diagram for function $g(\phi)$ for Example 16.7.1.

When the gain function is plotted, a three-dimensional plot results, as sketched in Fig. 16.7.1 (a). The length of the line from the origin to any point on the surface of the figure gives the gain in the direction of the point. The maximum gain G_M is shown, as well as the gain $G(\theta_1, \phi_1)$, in the direction (θ_1, ϕ_1) . Minor lobes, as indicated by G_2 and G_3 , also occur in general.

In practice, two-dimensional plots are often used, one for the equatorial plane and one for the meridian plane, the function $g(\theta, \phi)$ usually being the one that is plotted. In the equatorial plane this is denoted by $g(\phi)$, since θ is constant, and in the meridian plane by $g(\theta)$, since ϕ is constant. This is illustrated in the following example.

EXAMPLE 16.7.1

For the Hertzian dipole (to be described later), the radiation pattern is described $g(\theta) = \sin^2\theta$ and $g(\phi) = 1$. Plot the polar diagrams.

SOLUTION The polar diagrams are plotted in Fig. 16.7.1(b) and (c).

The -3-dB beamwidth of an antenna is the angle subtended at the center of the polar diagram by the -3-dB gain lines. This is illustrated in Fig. 16.7.2, where the beamwidth is

$$\theta_3 = \theta_2 - \theta_1 \quad (16.7.5)$$

EXAMPLE 16.7.2

Determine the -3-dB beamwidth for the Hertzian dipole of Example 16.7.1.

SOLUTION $g(\theta_1) = \sin^2 \theta_1 = 0.5$, giving $\theta_1 = 45^\circ$. Also, $g(\theta_2) = \sin^2 \theta_2 = 0.5$, giving $\theta_2 = 135^\circ$. Therefore,

$$\theta_3 = \theta_2 - \theta_1 = 135^\circ - 45^\circ = 90^\circ$$

It will be observed from this example that the beamwidth applies only to the meridian plane for this antenna, since the equatorial plane polar diagram is a circle.

In certain cases the beamwidth may be specified for levels other than -3 dB , other common values being -10 and -60 dB . Obviously the beamwidth level must be specified along with the beamwidth.

16.8 Effective Area of an Antenna

A receiving antenna may be thought of as having an effective area that collects electromagnetic energy from the incident wave, rather as a solar collector collects energy from sunlight. Assuming that the antenna is in the far-field zone of the radiated wave, the wave incident on it will be a plane TEM wave having a power density of $P_D \text{ W/m}^2$ of wavefront. Let the receiving antenna be at the center of a spherical coordinate system, and let the incoming wave direction be specified by the angular coordinates (θ, ϕ) with reference to the antenna. The power delivered to a matched load (receiver) will be a function of direction, and this is taken into account by making the effective area a function of direction; that is, $A = A(\theta, \phi)$. Thus, if P_R is the power delivered to a matched load,

$$P_R = P_D A(\theta, \phi) \quad (16.8.1)$$

Equation (16.8.1) serves as a defining equation for effective area. The effective area will have some maximum value A_{eff} , called the *effective area of the antenna*, just as the maximum power gain is called the gain of the antenna. As a result of the reciprocity theorem, the effective area normalized to its maximum value has the same functional form as the normalized power gain [Eq. (16.7.2)]; that is,

$$\frac{A(\theta, \phi)}{A_{\text{eff}}} = g(\theta, \phi) \quad (16.8.2)$$

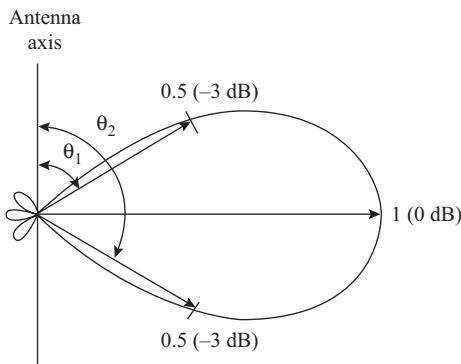


Figure 16.7.2 -3-dB Beamwidth of an antenna.

It follows that A_{eff} is proportional to G_M . It can also be shown from the reciprocity theorem that the constant of proportionality is the *same for all antennas*: $\lambda/4\pi$. Thus

$$\frac{A_{\text{eff}}}{G_M} = \frac{\lambda^2}{4\pi} \quad (16.8.3)$$

Thus, if the gain of an antenna is measured under transmitting conditions as G_T , its effective area under receiving conditions can be found from Eq. (16.8.3) to be

$$A_{\text{eff}} = \frac{\lambda^2 G_T}{4\pi} \quad (16.8.4)$$

This relationship was used, for example, in deriving Eq. (15.2.5) for free-space transmission. Note that G_T takes into account antenna efficiency, and therefore so does the effective area. Often, in theoretical calculations, the directivity D_T is used in Eq. (16.8.4) instead of G_T , and this will give a higher value of effective area since directivity excludes the antenna efficiency.

Another factor that can reduce the effective area is the mismatch factor, given by Eq. (16.2.7). The effective area is reduced directly by this factor, and, of course, if the antenna is matched to the line, no reduction occurs.

Previously, the effective area was shown to be a function of the angular coordinates θ and ϕ . Defined this way, $A(\theta, \phi)$ automatically takes into account any loss resulting from polarization misalignment. However, it is usually the maximum value A_{eff} of $A(\theta, \phi)$ that is known, as well as the polarization loss factor plf, as given by Eq. (16.5.1). Since the plf is defined for electric field strength, and A_{eff} for power, the reduction in A_{eff} as a result of polarization misalignment is $(\text{plf})^2$.

16.9 Effective Length of an Antenna

Although the concept of effective area can be used with any antenna, it is particularly useful with microwave antennas. At lower frequencies, where the physical structure of the antenna is of the form of a linear conductor or an array of conductors, an analogous concept, the *effective length*, proves to be more useful.

For a receiving antenna, the open circuit emf appearing at the terminals is V_A , the Thévenin equivalent voltage, as shown in Fig. 16.2.1(c). Now, this is produced by a wave having an electric field strength of E V/m sweeping over the antenna, and therefore an effective length ℓ_{eff} can be defined by

$$V_A = E\ell_{\text{eff}} \quad (16.9.1)$$

The effective length ℓ_{eff} as defined by Eq. (16.9.1) is the maximum value. The effective length will in general be a function of θ and ϕ , just as the effective area is in general, and with ℓ_{eff} it is to be understood that the antenna is oriented for maximum induced emf, so, for example, the polarization loss factor would be unity.

For the transmitting antenna, effective length is defined in a different manner, but it can be shown by the use of the reciprocity theorem that the effective length for transmitting is the same as that for receiving. The definition of ℓ_{eff} under transmitting conditions is in terms of the current distribution. The antenna current will vary as a function of physical length along the antenna. The effective length is defined such that the product of input terminal current and effective length is equal to the area under the actual current-length curve. Let I_0 represent the input terminal current; then

$$I_0\ell_{\text{eff}} = \text{area under current-length curve} \quad (16.9.2)$$

EXAMPLE 16.9.1

For the $\frac{1}{2}\lambda$ dipole described in Section 16.11, the current-length curve may be assumed to be $I = I_0 \cos \beta\ell$, where $\ell = 0$ at the input terminals. Find the effective length.

SOLUTION The physical length of the antenna is $\lambda/2$, and the average current for the cosine distribution is

$$I_{\text{av}} = \frac{2}{\pi} I_0$$

Thus

$$\begin{aligned} \text{area} &= I_{\text{av}} \times \text{physical length} \\ &= \frac{2}{\pi} I_0 \frac{\lambda}{2} = \frac{I_0 \lambda}{\pi} \end{aligned}$$

Hence, from Eq. (16.9.2),

$$\ell_{\text{eff}} = \frac{\lambda}{\pi}$$

For low- and medium-frequency antennas that are mounted vertically from the earth's surface, the effective length is usually referred to as the *effective height* h_{eff} . This is directly related to the physical height. Effective height must not be confused with the physical heights h_T and h_R introduced in Section 15.3. For example a $\frac{1}{2}\lambda$ dipole may be mounted at some mast height h above the ground, but its effective length is always λ/π .

16.10 Hertzian Dipole

The Hertzian dipole is a short linear antenna that, when radiating, is assumed to carry uniform current along its length. Such an antenna cannot be realized in practice, but longer antennas can be assumed to be made up of a number of Hertzian dipoles connected in series. The radiation properties of the Hertzian dipole are readily calculated. This is useful in itself in that it helps to illustrate the general properties discussed in the previous sections, but also the properties of longer antennas can often be deduced by superimposing the results of the chain of Hertzian dipoles making up the longer antenna.

An approximation to a Hertzian dipole can be achieved by capacitively loading the ends of a short center-fed dipole as shown in Fig. 16.10.1(a). The capacitive ends allow a nearly uniform charging current to be maintained in the wire if its length $\delta\ell$ is much shorter than a wavelength ($\delta\ell \ll \lambda$). Thus, over the entire length of the dipole, the current is assumed to be

$$i = I_0 \sin \omega t \quad (16.10.1)$$

The electric field in the far-field zone is directly proportional to the component of current that is parallel to the electric field, which, as shown in Fig. 16.10.1(b), is $I_0 \sin \theta$. From the physics of radiation, it is

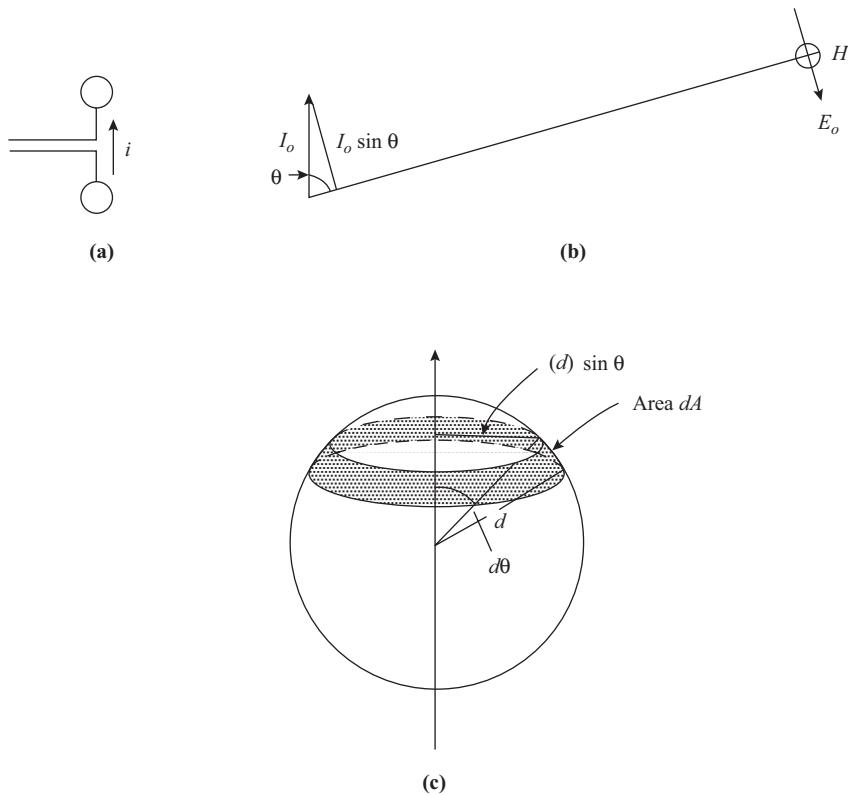


Figure 16.10.1 (a) Hertzian dipole. (b) Far zone field E_o is proportional to the current component $I_0 \sin (\theta)$. (c) Sphere used for determining radiated power.

also found that the instantaneous electric field is proportional to the rate of change of current and inversely proportional to the distance d , the full expression being

$$e_\theta = \frac{60\pi \delta\ell I_0 \sin \theta}{\lambda d} \cos \omega \left(t - \frac{d}{c} \right) \quad (16.10.2)$$

The subscript θ is used to show that the field is in the direction of the coordinate θ and at right angles to the direction of propagation. A delay time d/c is included, which is the time taken for a change in current to be effective at the point in the far-field zone.

The maximum value of the electric field, denoted by E_0 , is

$$E_0 = \frac{60\pi \delta\ell I_0 \sin \theta}{\lambda d} \quad (16.10.3)$$

The rms current I may be substituted for the maximum current I_0 to give the rms field strength E . In Appendix B it is shown that the power density in the far-field zone is given by $E^2/120\pi$, and using the rms field obtained from Eq. (16.10.3) gives, for the power density,

$$P_D = P_{DM} \sin^2 \theta \quad (16.10.4)$$

where

$$P_{DM} = 30\pi \left(\frac{\delta\ell I}{\lambda d} \right)^2 \quad (16.10.5)$$

The total power flow through a strip [Fig. 16.10.1(c)] is given by $dP = P_D dA$, where dA is the area of the strip. From the geometry of the figure, $dA = 2\pi d^2 \sin \theta d\theta$. The total power is the summation of all such elemental amounts dP , which in the limit becomes the integral

$$\begin{aligned} P_T &= \int_0^\pi (P_{DM} \sin^2 \theta) (2\pi d^2 \sin \theta d\theta) \\ &= 2\pi d^2 P_{DM} \int_0^\pi \sin^3 \theta d\theta \\ &= \frac{8\pi d^2 P_{DM}}{3} \end{aligned} \quad (16.10.6)$$

Substituting for P_{DM} and simplifying,

$$P_T = I^2 80 \left(\frac{\pi \delta\ell}{\lambda} \right)^2 \quad (16.10.7)$$

Now, the radiation resistance is defined by the relationship $P_T = I^2 R_{rad}$, and therefore, from Eq. (16.10.7),

$$R_{rad} = 80 \left(\frac{\pi \delta\ell}{\lambda} \right)^2 \quad (16.10.8)$$

The directive gain can also be determined. This is defined as the ratio of the power per unit solid angle to the average power per unit solid angle, as in the derivation of Eq. (16.7.4). The relationship between power per unit solid angle and power density is $P(\theta, \phi) = d^2 P_D$ [see, for example, Eq. (16.6.3)]. The average power per unit solid angle is $P_T/4\pi$, and therefore

$$\begin{aligned} D(\theta, \phi) &= \frac{4\pi P(\theta, \phi)}{P_T} \\ &= \frac{4\pi d^2 P_{DM} \sin^2 \theta}{8\pi d^2 P_{DM}/3} \\ &= 1.5 \sin^2 \theta \end{aligned} \quad (16.10.9)$$

From Eq. (16.10.9), the directivity is seen to be $D_M = 1.5$, and the normalized gain is $g(\theta, \phi) = \sin^2 \theta$. Because of the symmetry of the antenna, there is no variation of gain in the equatorial plane [that is, $g(\phi) = 1$], and the gain variation in the meridian plane is $g(\theta) = \sin^2 \theta$. These are the functions that were plotted in Example 16.7.1.

Lastly, the effective area of the Hertzian dipole is found, using Eq. (16.8.4), to be

$$A_{\text{eff}} = 1.5 \frac{\lambda^2}{4\pi} = 0.119\lambda^2 \quad (16.10.10)$$

This is the effective area for unity efficiency.

16.11 Half-wave Dipole

The *half-wave dipole* is a resonant antenna, the total length of which is nominally $\frac{1}{2}\lambda$ at the carrier frequency. Standing waves of voltage and current exist along the antenna, a good approximation to the distribution being obtained by assuming the antenna to be an opened-out $\frac{1}{4}\lambda$ section of an open-circuited transmission line. As shown in Section 13.7, the spacing between a standing-wave maximum and minimum is $\frac{1}{4}\lambda$, and since the current must be zero at the open circuit, it will be a maximum $\frac{1}{4}\lambda$ in from the end, while the voltage is a maximum at the end, going to a minimum at the $\frac{1}{4}\lambda$ point. Figure 16.11.1(a) shows how the magnitudes of voltage and current vary as a function of distance from the end, while Fig. 16.11.1(b) shows the line opened out to form the $\frac{1}{2}\lambda$ dipole, along with the voltage and current distributions, which are assumed unchanged. Because a 180° phase shift also occurs along the $\frac{1}{4}\lambda$ section ($\frac{1}{2}\pi$ radians for the incident wave, and $\frac{1}{2}\pi$ radians for the reflected wave in the opposite direction), it is convenient to show the voltage and current as in Fig. 16.11.1(b). The voltage is assumed to go through zero at the feed point. Of course, the voltage must be finite at the feed point, and the amplitude distribution cannot be identical to that of the transmission line section since the geometry of the antenna is different; however, results based on these assumptions agree very well with the measured results. The $\frac{1}{2}\lambda$ dipole may then be considered as consisting of a large number of Hertzian dipoles connected in series, the current in each being determined by the current distribution shown in Fig. 16.11.1(b).

There will also be a phase difference between radiation from different elements on the half-wave dipole as a result of the difference in distance ($d_0 - d$), as shown in Fig. 16.11.2(b). Applying the cosine rule to the triangle formed by ℓ , d_0 , and d results in $d^2 = d_0^2 + \ell^2 - 2d_0\ell \cos \theta$, and for $d_0 \gg \ell$, this gives $d_0^2 - d^2 \approx 2d_0\ell \cos \theta$. It is left as an exercise for the student to show that, applying the same mathematics as was used in deriving Eq. (15.3.4),

$$(d_0 - d) \approx \ell \cos \theta \quad (16.11.1)$$

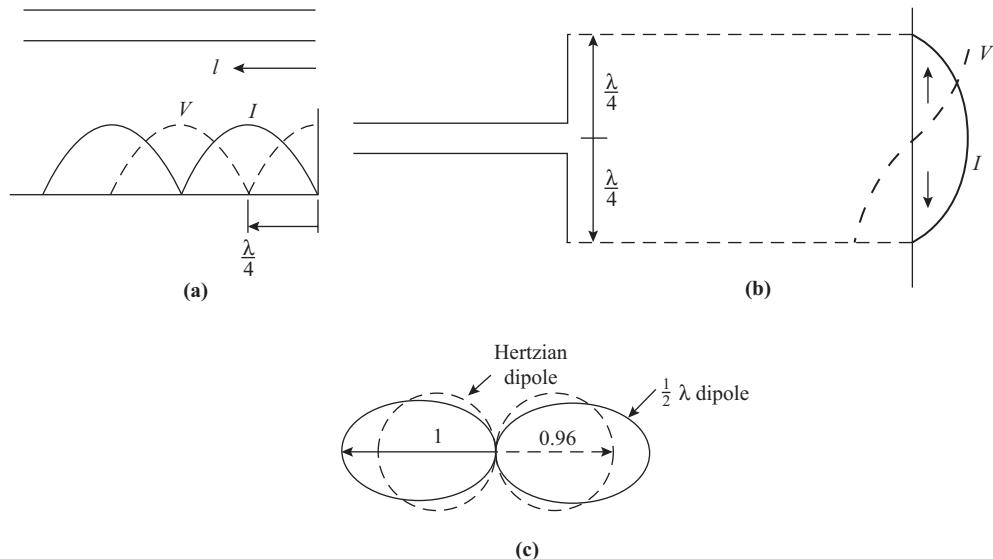


Figure 16.11.1 Half-wave dipole. (a) Current and voltage standing waves on an open-circuited line. (b) Current and voltage standing waves on a $\lambda/2$ dipole. (c) Radiation pattern of a $\lambda/2$ dipole compared to that of a Hertzian dipole.

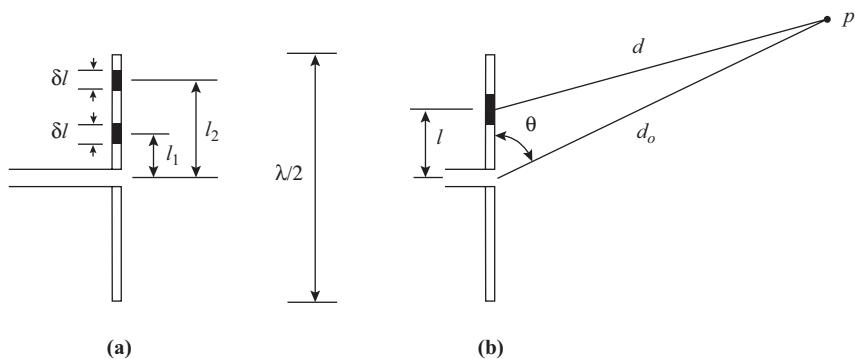


Figure 16.11.2 (a) Half-wave dipole made up of Hertzian dipoles. (b) Path difference introduces a phase difference.

The phase shift resulting from this is $(d_0 - d) 2\pi/\lambda$, and this is seen to depend on both ℓ and θ . Thus the response at some distant point P resulting from all Hertzian dipole elements making up the $\frac{1}{2}\lambda$ dipole must take into account this phase difference, as well as the current distribution on the $\frac{1}{2}\lambda$ dipole. The total response can be found by integrating the individual fields for the full length of the $\frac{1}{2}\lambda$ dipole. This is quite a difficult integration to perform, and only the results will be given here. Corresponding to Eq. (16.10.3), the peak field strength is

$$E_0 = \frac{60I_0}{d} F(\theta) \quad (16.11.2)$$

where

$$F(\theta) = \frac{\cos[(\pi/2) \cos \theta]}{\sin \theta} \quad (16.11.3)$$

The normalized power gain function is therefore

$$g(\theta) = F^2(\theta) \quad (16.11.4)$$

As with the Hertzian dipole, $g(\phi) = 1$ because of symmetry. It is left as an exercise for the student to show that the -3 -dB beamwidth for the $\frac{1}{2}\lambda$ dipole is

$$\theta_3 = 78^\circ \quad (16.11.5)$$

The *field strength* polar diagram $F(\theta)$ is shown in Fig. 16.11.1(c), along with that for the Hertzian dipole [$F(\theta) = \sin \theta$] for comparison, both normalized to unity for $\frac{1}{2}\lambda$ dipole. Other important results for the $\frac{1}{2}\lambda$ dipole are

$$\ell_{\text{eff}} = \lambda/\pi \quad (16.11.6)$$

$$D_M = 1.64 \quad (16.11.7)$$

$$A_{\text{eff}} = 0.13\lambda^2 \quad (16.11.8)$$

$$R_{\text{rad}} = 73 \Omega \quad (16.11.9)$$

The total impedance will be a function of frequency, being capacitive for frequencies just below the resonant value and inductive for frequencies above the resonant value, up to the next resonant value, which occurs when the physical length is approximately one wavelength. Because the velocity on the wire is slightly slower than that in free space, resonance does not occur at exactly the $\frac{1}{2}\lambda$ point, but at a slightly shorter length, in practice about 95% of the $\frac{1}{2}\lambda$ value. At $\ell = \frac{1}{2}\lambda$, the antenna impedance is $73 + j42.5 \Omega$, and at the 5% lower length it is 73Ω , as illustrated in Fig. 16.11.3.

Transmitted waves are rarely single-frequency sinusoids, but are modulated. All modulated waves are made up of a carrier and a number of sideband frequencies spread out on either side. Since the half-wave dipole (or any other tuned antenna) is resonant at only one frequency, and since it behaves like a frequency-dependent reactance at other frequencies, the sideband frequencies will be distorted somewhat. For narrow-band transmissions this distortion is not significant, but at higher bandwidths it can cause trouble. The characteristics can be improved slightly by spoiling the effective Q of the antenna by making the radiating

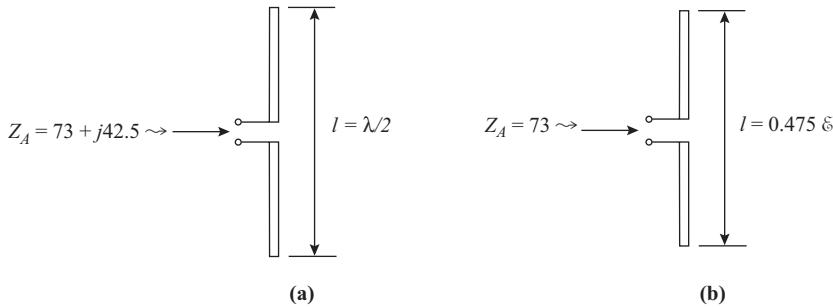


Figure 16.11.3 (a) Input impedance of a half-wave dipole cut to exactly $\lambda/2$ is $73 + j42.5 \Omega$. (b) Shortening its length by 5% reduces the reactive component to zero.

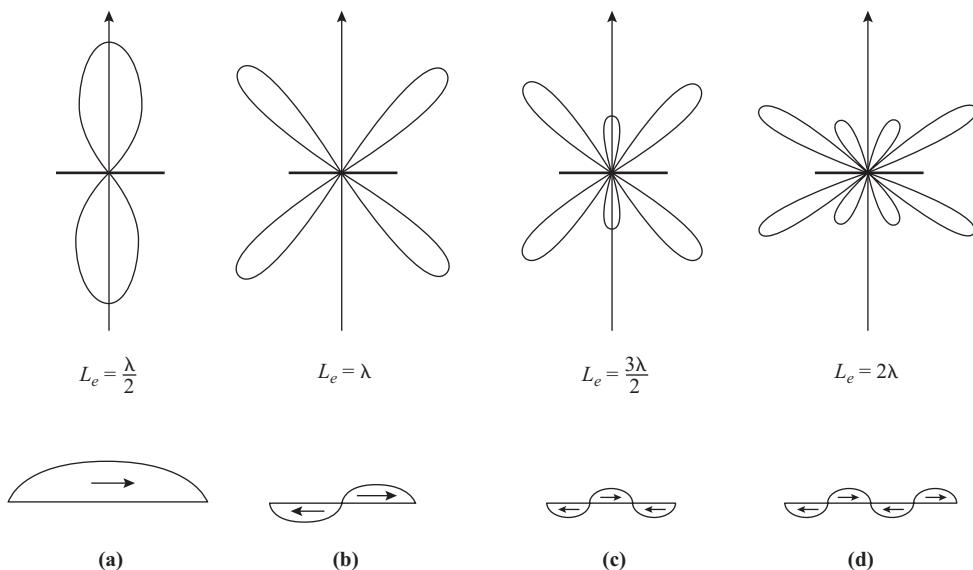


Figure 16.11.4 Resonant dipole radiation patterns with their current distributions for multiples of (a) one, (b) two, (c) three, and (d) four times the half-wavelength.

conductors large in diameter. This has the effect of increasing the capacitance and lowering the inductance, thus lowering the Q of the antenna. The result is a wider-frequency deviation between the 3-dB frequency points, or a wider bandwidth.

Resonance in the dipole is not limited to the half-wave frequency, but occurs also at all integer multiples of the half-wave frequency. The current distributions are different for each case, and the result is a different radiation pattern for each resonant frequency. This is illustrated in Fig. 16.11.4 for the cases of effective length equal to 1, 2, 3, and 4 times the half-wavelength. The drawings show that the number of lobes on each side of the radiator is equal to the multiple of half-wavelengths used. Current phasing changes by 180° from one $\frac{1}{2}\lambda$ section to another and thus is shown by the current arrows. Again, these patterns are the ones that would result if the antenna were mounted free in space, away from reflections. It must also be noted that the current node occurs at different positions on the antenna length, and if a resonant feed point is to be maintained, it must be located at one of the current nodes that occur; that is, it must be located at $\frac{1}{4}\lambda, \frac{3}{4}\lambda, \frac{5}{4}\lambda, \dots$ away from either end of the radiator.

With the increase in the number of lobes, the lobes nearest the antenna axis will always be larger than the others. These are the major lobes, and they get progressively closer to the axis with increasing number of lobes.

16.12 Vertical Antennas

Ground Reflections

The ground will act as an almost perfect reflecting plane for any antenna placed near its surface, and an apparent mirror image of the antenna will appear to be located immediately beneath the surface below the antenna. This is illustrated by Fig. 16.12.1, which shows a distant observer receiving a direct wave from a point on the antenna and a reflected wave that appears to come from the corresponding point on the image antenna.

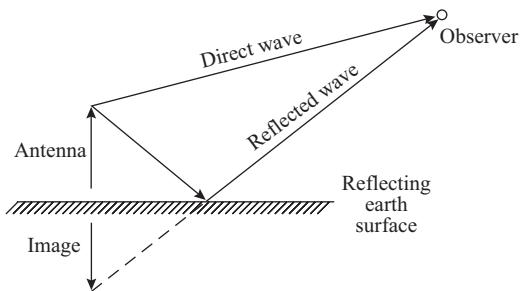


Figure 16.12.1 Vertical antenna and its reflected image.

Because of the interaction between the direct and reflected waves, the radiation polar pattern is drastically modified and appears to be the vector sum of radiation from two separate antennas, the real one and the image. The amount of interaction is dependent on how far above the ground the antenna is placed. If the effective height is several wavelengths, then practically no interaction will occur, and the antenna may be considered to be mounted in free space. For heights up to a few wavelengths, the reflections must be considered, and the antenna and its image act as a phased array of two antennas.

The image antenna is an exact mirror image of the real antenna, and the apparent currents within it are at each point the same as those within the real antenna, except for reversed polarity. This is true whether the antenna is grounded or not grounded, horizontally or vertically polarized.

Grounded Vertical Antennas

Most of the medium-frequency (MF) broadcast antennas fall into this category, as do the VHF mobile-whip antennas. This type of antenna, known as the *Marconi antenna*, is made up of a vertical mast, pole, or rod that forms the main radiating conductor. It may be free-standing or supported by insulated guy wires and is placed in a location where good electrical ground is available. Good locations for MF antennas include marshy fields and seacoast flats. If poor soil conditions exist, an artificial ground plane may be created by burying a mat of heavy conductors extending radially from the mast for up to at least a quarter-wavelength, and preferably at least a half-wavelength, from the mast. The ends of the buried radials are usually connected together, with deep grounding stakes driven at their extremities. The base of the mast is electrically insulated from ground, and the feed line from the transmitter is connected between the mast base and the ground. (In some special cases the feed point may be located at a current node farther up the mast.)

Since high-power transmitters may generate potentials of several hundred kilovolts on the antenna structure, high-quality insulators are a must. At lower frequencies special plastics such as polystyrene and Teflon are used.

The grounded vertical antenna combines with its image to act in exactly the same manner as a doublet or dipole, with the radiator vertical to the ground surface. Figure 16.12.2 shows grounded verticals of various lengths with the current distributions on the antenna and its image, and the resulting radiation pattern in the vertical plane. The radiation pattern in the horizontal plane is circular. For vertical radiators of $\frac{1}{4}\lambda$ or less in effective height, the current distributions on the combined antenna-image structure is identical to a doublet or dipole of twice the length of the radiator and produces a radiation pattern that is the same. However, half of the radiation pattern appears to be below the ground surface and in fact does not exist. All the power from the radiator is contained in that portion of the pattern that is above the surface. The $\frac{1}{4}\lambda$ vertical acts similarly to the half-wave dipole. The current distribution and radiation pattern are shown in Fig. 16.12.2(a). The structure behaves like two antennas being fed in parallel; the radiation resistance is reduced to half that of the dipole, or about 36.5Ω .

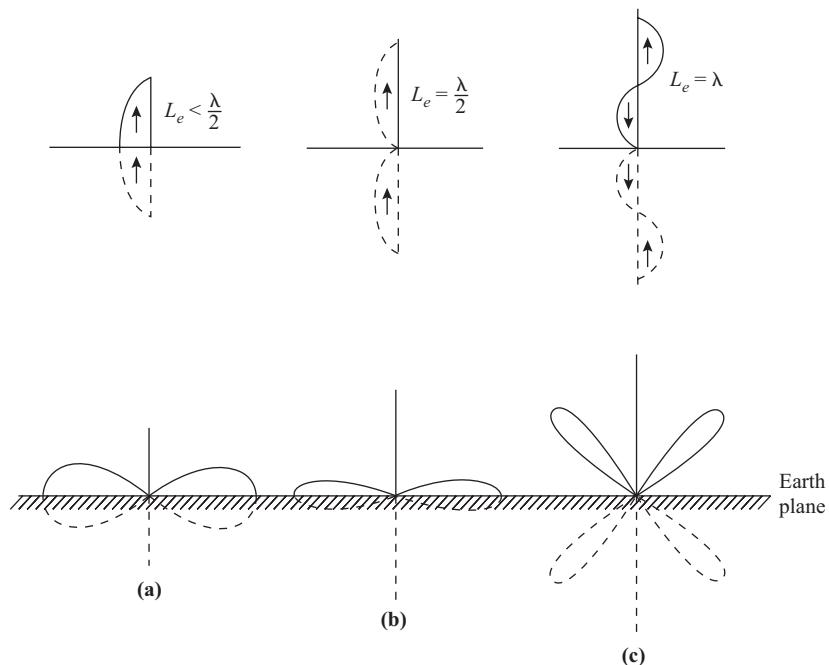


Figure 16.12.2 Grounded vertical radiators. Current distributions and radiation patterns for effective lengths (a) less than $\lambda/2$, (b) $\lambda/2$, and (c) twice $\lambda/2$.

Figure 16.12.2(b) shows the current distributions and pattern for a $\frac{1}{2}\lambda$ vertical radiator. The structure no longer behaves like a combined dipole but more like an array of two separate antennas of the same length fed 180° out of phase with each other. Each half is a $\frac{1}{2}\lambda$ antenna, and the two halves are aligned so that their radiation patterns are in phase along the ground surface. This makes the antenna appear to have the same pattern shape as the $\frac{1}{2}\lambda$ dipole, but with an additional gain of 3 dB.

When the length of the radiator is increased above $\frac{1}{2}\lambda$, side lobes begin to appear, and the main lobe lifts off the ground. Figure 16.12.2(c) shows the result for the 1λ radiator. This is identical in shape to the pattern for the 1λ dipole shown in Fig. 16.11.4(b), except that the bottom half of the pattern has been folded over and added to the top half. It is obvious from this that if a vertical is to radiate along the surface of the ground, it must not be higher than $\frac{1}{2}\lambda$ or too much power will be radiated into the sky. If sky waves are to be used, a longer radiator must be used to get the main lobe off the ground.

The $\frac{1}{4}\lambda$ verticals are favored because they may be fed directly with a cable at the low-impedance current node that occurs at the base of the mast. At low frequencies it is often not physically feasible to build a mast that is a full $\frac{1}{4}\lambda$ in height, so a form of “top-hat” loading is employed to make the radiator look electrically longer than it physically is. A horizontal wire is connected to the top of the mast to make up the missing length. This may be a simple wire, as in inverted L and T antennas, or it may be in the form of a disk. The length of the horizontal or the disk radius is adjusted so that the current node occurs at the base of the mast, and from the feed point the antenna appears to be a $\frac{1}{4}\lambda$ tuned vertical. The current density in the horizontal portion is low and does not radiate nearly as much power as does the vertical portion, with the result that little change occurs in the radiation pattern. Some minor lobes in the upward direction do occur, but the major lobe is caused to lie closer to the ground because of more even current distribution within the vertical mast. This type of antenna is sometimes referred to as a *unipole antenna*, because only one radiating element is fed.

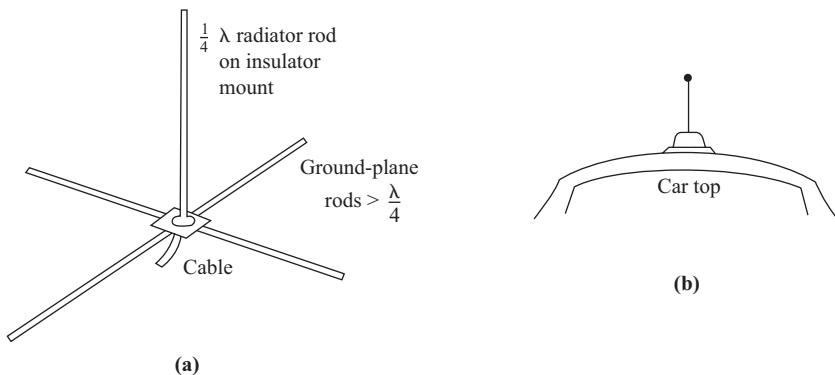


Figure 16.12.3 Whip antennas. (a) VHF vertical with a simulated ground plane. (b) UHF mobile whip mounted on a car top.

Grounded verticals are frequently used at VHF for mobile service because of their simple structure. In this case the radiator length varies from a few feet down to a few inches, and a real earth ground cannot be used. For fixed antennas a series of radial ground-plane rods are frequently used, while for the mobile antennas, the flat metal top of the vehicle is used to provide the ground plane. Figure 16.12.3 illustrates these grounds.

16.13 Folded Elements

The antennas discussed so far all use single-conductor radiators. However, if each conductor is twinned with a second conductor, insulated from, but closely parallel to it, and connected together at the voltage node points, then a similar current pattern will be induced in the second conductor. The radiation pattern will be exactly the same as for the single-conductor antenna, but the radiation resistance will be different, just four times that produced by the single radiator. Thus a folded vertical antenna as shown in Fig. 16.13.1(a) will have a radiation resistance four times that of the single vertical (4×36.5), or 146Ω . The folded dipole shown in Fig. 16.13.1(b) has a radiation resistance of 4×73 , or 292Ω .

The folded dipole antenna is favored as the driving element of VHF dipole arrays because it can be made very inexpensively with self-supporting tubing and provides a higher terminal impedance, which tends to offset the reduction of impedance resulting from the loading of the parasitic elements. Since the center of

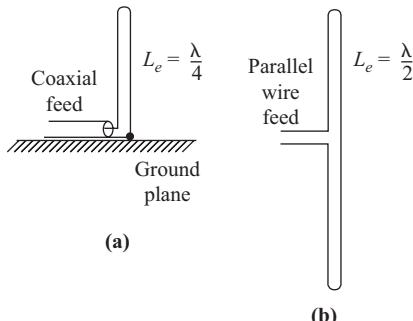


Figure 16.13.1 Folded dipoles. (a) Folded quarter-wave vertical and (b) folded half-wave dipole.

the second conductor is at a current node, it may be electrically connected to a grounded support rod without affecting the characteristics, making a very sturdy mechanical structure.

16.14 Loop and Ferrite-rod Receiving Antennas

Loop Antenna

The *loop antenna* is made up of one or more turns of wire on a frame, which may be rectangular or circular, and is very much smaller than one wavelength across. This antenna is popular for two reasons: (1) it is relatively compact, lending itself to use with portable receivers, and (2) it is quite directive, lending itself to use with direction-finding equipment.

A loop antenna is shown in Fig. 16.14.1(a), with its radiation pattern. The radiation pattern is the doughnut shape of the doublet antenna, except that the plane of the doughnut corresponds to the plane of the loop, so the loop will radiate equally well in all directions within its own plane. A distinct null response occurs along the axis of the loop, and it is this null that is used in direction finding. The physical shape of the loop does not drastically affect the radiation pattern except when the length and width of the loop are unequal, in which case a squashed doughnut shape occurs.

Loop antennas made of several turns of wire around a rectangular frame were popular for earlier model broadcast receivers, with the loop being mounted in the back of the cabinet. Recently, these have been almost entirely replaced by the smaller ferrite-rod antennas.

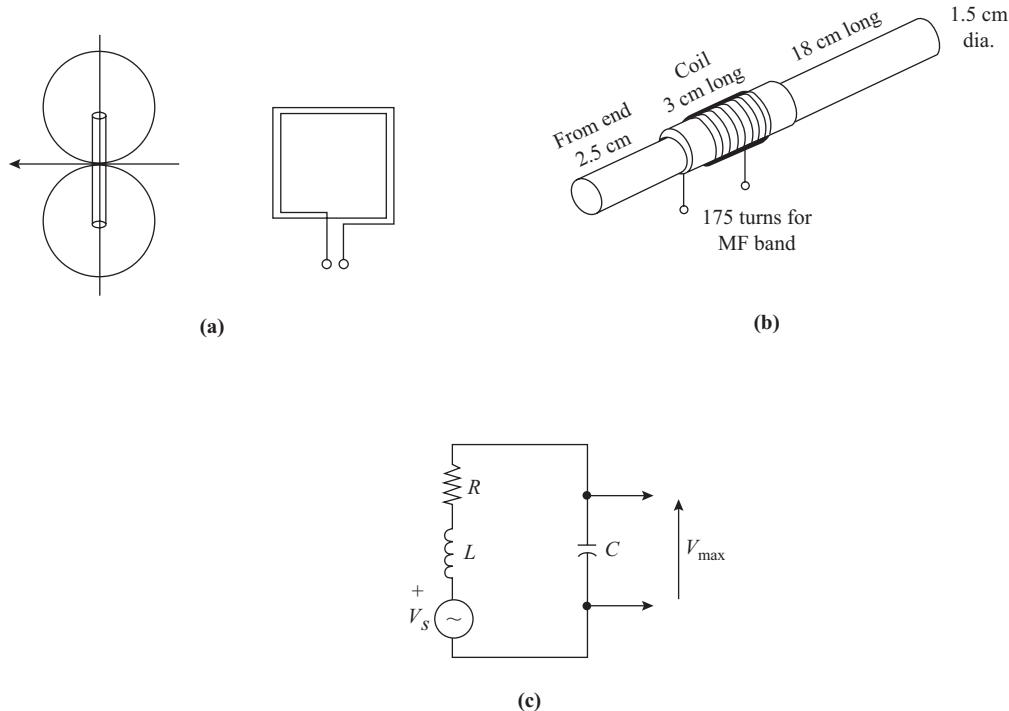


Figure 16.14.1 Loop antennas, (a) Square loop antenna with its radiation pattern. (b) Ferrite rod antenna. (c) Equivalent circuit for the ferrite rod antenna.

When the loop is aligned for maximum signal strength, the magnetic flux linkages are BAN , where B is the rms magnetic flux density in teslas, A is the physical loop area in square meters, and N is the number of turns in the loop. The induced emf is given by Faraday's law as the rate of change of flux linkages, which for a wave of angular frequency ω , gives

$$V_s = \omega BAN \quad (16.14.1)$$

When the loop is tuned by means of an external capacitor to the received frequency, the voltage at the capacitor terminal is magnified by the circuit Q to give

$$V_{\max} = V_s Q = \omega BAN Q \quad (16.14.2)$$

Since the loop is usually much smaller than the received wavelength, the induced voltage may be quite small. It can be increased by increasing any one of the factors in Eq. (16.14.2). The Q is determined by the desired selectivity. The area must be kept small: increasing the number of turns increases the coil inductance and changes the Q , and even changing the flux density B affects the Q . However, changing the flux density by using a magnetic core can be achieved with a minimal change of Q using ferrite cores. This alternative is now in widespread use.

Ferrite-rod Antenna

The *ferrite-rod antenna* is made by winding a coil of wire on a ferrite rod similar to the one illustrated in Fig. 16.14.1(b). Ferrites are materials that exhibit the properties of ferromagnetism. The materials exhibit a high relative permeability in the same manner as magnetic metals do, but unlike the ferromagnetic metals, they also have a high bulk resistivity. This means that at high frequencies, eddy currents induced within the materials are practically nonexistent, and high- Q coils can be made. Typical values for μ are around 100 and for resistivity 10,000 $\Omega\text{-cm}$. A high length-to-diameter ratio for the rod gives a high permeability, which is desirable.

The size of coil is a compromise among several factors. If the coil is too long compared to the rod length, the change of permeability with temperature will cause a noticeable change in the inductance. If it is too short, the Q will be low. Positioning the coil on the core is critical as well, since the effective permeability is a function of position on the rod, ranging from a maximum at the center to a minimum at either end. The coil is usually placed near the quarter-point, allowing adjustment in either direction to trim the coil inductance. When more than one coil is mounted on the same rod, they must be placed at opposite ends to minimize interaction between them.

The coil of wire on the ferrite rod is basically a modified loop antenna, so the induced maximum emf appearing at its terminals is given by

$$V_s = \omega BANF\mu_r \quad (16.14.3)$$

where F = modifying factor accounting for coil length, ranging from unity for short coils to about 0.7 for one that extends the full length of the rod

μ_r = effective relative permeability of the rod, as measured for the actual coil position

A = rod cross-sectional area

An expression for the effective length of a ferrite rod antenna can be derived by combining Eqs. (B.4), (B.8), (16.9.1), and (16.14.3) to give

$$\ell_{\text{eff}} = \frac{2\pi ANF\mu_r}{\lambda} \quad (16.14.4)$$

Since the voltage appearing at the terminals is of more importance in a receiving antenna, the factor $Q\ell_{\text{eff}}$ is often given as a figure of merit for rod antennas. The directional properties of the ferrite-rod antenna are similar to those of the loop antenna, although the null may not be quite so pronounced.

16.15 Nonresonant Antennas

Long-wire Antenna

The *long-wire antenna* is just that, a wire several wavelengths in length that is suspended at some height above the earth. The wire is driven at one end and has a resistive termination at the remote end that is matched to the characteristic impedance of the line at that end. This forms a transmission line with a ground return and a matched termination. When an alternating current wave is transmitted down this line toward the terminated end, about half of the energy is radiated into space. Since there is no reflection at the far end, no return wave exists, and no standing waves appear on the wire, regardless of its length-to-wavelength ratio. Of the energy not radiated, a small amount is dissipated in the wire, and the remainder is dissipated in the terminating resistance.

The long wire is illustrated in Fig. 16.15.1(a), with its horizontal radiation pattern. The radiation pattern shown would be true for any direction at right angles to the wire if the wire were mounted in free space.

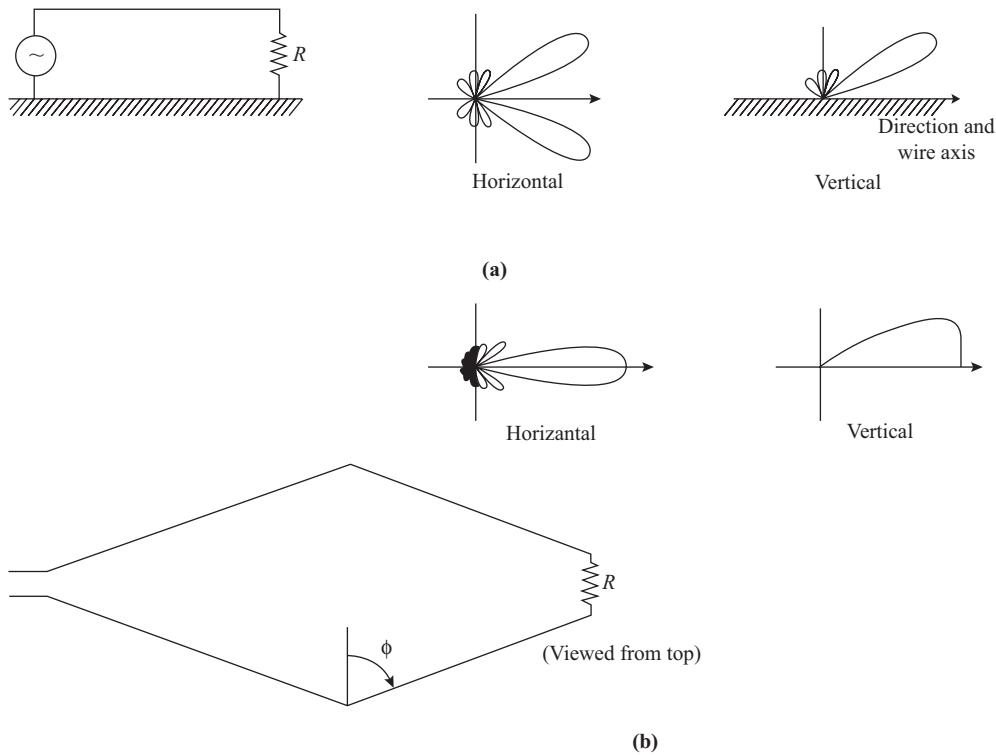


Figure 16.15.1 Nonresonant antennas, (a) Long-wire antenna with its horizontal and vertical radiation patterns. (b) Rhombic antenna with its horizontal and vertical radiation patterns.

Usually it is a fraction of a wavelength above the ground, and ground reflections cause most of the energy to be radiated upward so that the vertical pattern would be a single lobe of twice the strength of the horizontal lobes.

This antenna is not often used because it is not very efficient, has a comparatively low gain, and takes up a lot of space. Also, matching the transmitter to the line can be a problem. However, since no standing waves exist, the antenna has no resonances, and as long as the length of the wire lies in the range 2λ to 10λ , its characteristics remain relatively constant for all frequencies in that range. It is thus used as a broadband antenna for low-cost point-to-point communications, especially in the HF band from 3 to 30 MHz. The upward tilt of the pattern lends itself to skywave propagation in this band.

Rhombic Antenna

The *rhombic antenna* takes its name from its diamond-shaped layout. It is an array of four interconnected long-wire antennas, laid out in the manner shown in Fig. 16.15.1(b). Each of the four legs has the same length and lies in the range 2λ to 10λ . The transmission line feeds one end and transmits an unreflected current wave down each side toward the resistive termination at the far end. The lengths of the sides and the angle ϕ are interrelated and must be carefully chosen so that the side lobes cancel properly, leaving only a single main lobe lying along the main axis of the rhombus. Again, ground reflections cause the lobe to be tilted upward into the sky, and the amount of tilt is a function of the length of the legs.

The resistive termination is chosen so that no reflections occur, and the antenna is untuned, as is the long-wire antenna. Its frequency range is broad, almost 10 to 1, allowing a single structure to be used over most of the HF bands. It is highly directional and, if the tilt is chosen properly, is ideal for point-to-point skywave propagation.

The feed-point impedance falls in the range from 600 to 800 Ω , allowing direct feed with an open-wire parallel line, and the terminating resistor is in the same range. The angle ϕ falls between 40° and 75° and the leg length between 2λ and 10λ . The resulting directive gain obtained with the rhombic ranges from 15 to 60. The physical structure is relatively simple and inexpensive and usually takes the form of four poles placed at the apexes of the rhombus and wires supported on tension insulators for the radiators. The feed line is a parallel line and can be any length within reason, although it is usual practice to place the transmitter house near the feed end of the rhombus. For example, a 2λ rhombic for 3 MHz would have diagonals of about 320 m long by about 250 m wide. This antenna is not practical in an urban environment.

16.16 Driven Arrays

A *linear array* of antennas consists of a number of basic antenna elements, usually $\frac{1}{2}\lambda$ dipoles, equispaced out along a line referred to as the array axis. By suitably phasing the radiation from each element, the directivity can be altered in a number of ways. To illustrate this, arrays of identical elements operating in the transmitting mode will be considered, as sketched in Fig. 16.16.1(a). It is assumed that each element is fed with currents of equal amplitudes, but with successive phase shifts α . Thus the current to the first element is I_0 , to the second, I_1 , to the third, I_2 , and to the n th element, I_{n-1} . Figure 16.16.1(b) shows the plan view of the array, and for convenience this is taken to be the equatorial plane. The distance between successive wavefronts is $s \cos \phi$, and therefore the phase lead of element n with respect to element $(n - 1)$ is $(2\pi/\lambda)s \cos \phi$. Thus the total phase lead of element n with respect to $(n - 1)$ is

$$\psi = \frac{2\pi}{\lambda} s \cos \phi + \alpha \quad (16.16.1)$$

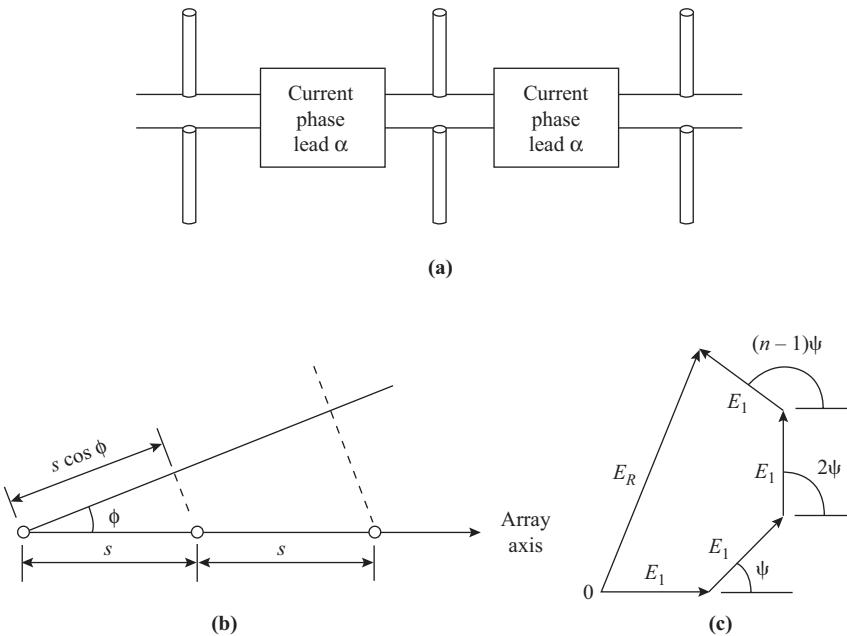


Figure 16.16.1 (a) Linear array of dipoles. (b) Plan view. (c) Phasor diagram for the field strength in the far field zone.

Figure 16.16.1(c) shows the phasor diagram for the individual field strengths at a point in the far-field zone. The individual field strengths are assumed to have equal amplitudes E_1 and to undergo successive phase shifts ψ . The resultant field strength amplitude E_R is the closing side of the polygon, and a geometrical result for such a polygon is

$$E_R = E_1 \frac{\sin(n\psi/2)}{\sin(\psi/2)} \quad (16.16.2)$$

Now, if the *same total power* had been radiated from a single element, the field strength would have been $\sqrt{n}E_1$. The *array factor* AF is the ratio of E_R to $\sqrt{n}E_1$, and therefore

$$AF = \left| \frac{\sin(n\psi/2)}{\sqrt{n} \sin(\psi/2)} \right| \quad (16.16.3)$$

This is the factor by which the array increases the field strength over that of a single element *radiating the same total power*.

The array factor has a maximum that occurs at $\psi = 0$. As ψ approaches zero, $\sin(n\psi/2) \rightarrow (n\psi/2)$ and $\sin(\psi/2) \rightarrow (\psi/2)$, so that, from Eq. (16.16.3),

$$AF_{\max} = \frac{(n\psi/2)}{\sqrt{n}(\psi/2)} = \sqrt{n} \quad (16.16.4)$$

The direction of the maximum can be selected by an appropriate choice of element spacing and current phasing, as illustrated in the following sections.

Broadside Array

The *broadside array*, as the name suggests, has its maximum directed along the normal to the plane of the array, when ψ is 90° in Fig. 16.16.1(b). This requires ψ to equal zero as shown in Eq. (16.16.4), and from Eq. (16.16.1), this in turn requires α to equal zero or the currents to be in phase. This can be achieved very conveniently by spacing the elements $\frac{1}{2}\lambda$ apart and alternately crossing the feed points as shown in Fig. 16.16.2(a). The field strength polar diagram in the equatorial plane is given by the normalized array factor. The diagram is sketched in Fig. 16.16.2(a). Increasing the number of elements sharpens the beam in the broadside direction, but will also introduce small sidelobes as sketched in Fig. 16.16.2(a). The polar diagram in the meridian plane will be that for a single element (normalized to the maximum value $\sqrt{nE_1}$), and for an array of half-wave dipoles, it would be similar to that shown in Fig. 16.11.1.

Quite often the broadside array is used in conjunction with a second array of reflectors, which is made the same size and mounted a half-wavelength behind the main array. The back lobe is now reflected forward and adds directly to the forward lobe, improving its gain and directivity and making the structure unidirectional. The reflectors may be driven, or they may be parasitic reflectors.

A variation of this array, called a *collinear broadside array*, is formed when a number of dipoles driven in phase are spaced in-line along the same axis. This array radiates equally well in all directions within the plane normal to the axis, but produces very little radiation off this plane and none along the array axis.

End-fire Array

The *end-fire array*, as the name suggests, has the main beam directed along the axis of the array, when ϕ is 0° in Fig. 16.16.1(b). As shown in Eq. (16.16.4), a maximum requires ψ to be zero, and hence, from Eq. (16.16.1),

$$0 = \frac{2\pi}{\lambda} s \cos(0) + \alpha$$

Thus

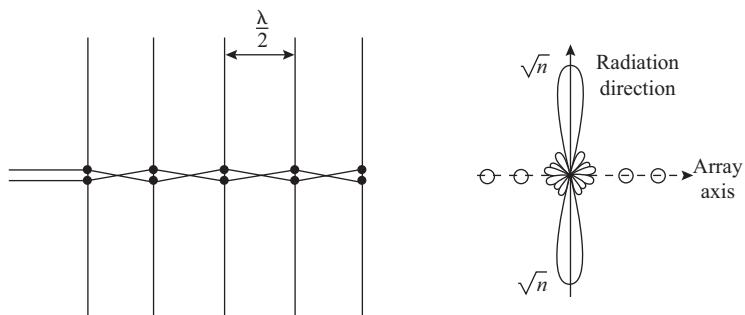
$$\alpha = -\frac{2\pi}{\lambda} s \quad (16.16.5)$$

The negative sign indicates that successive phase lags are required, proportional to the spacing s . Thus the phase angle ψ for the end-fire array becomes

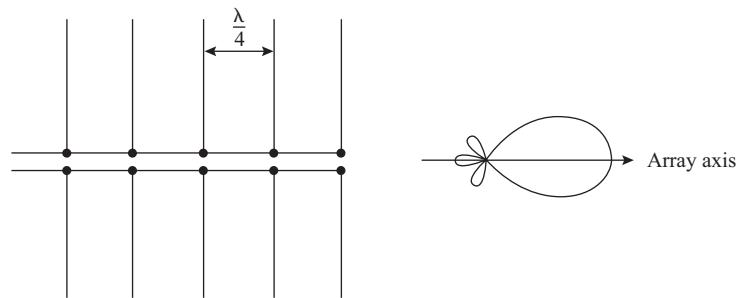
$$\psi = \frac{2\pi}{\lambda} s (\cos \phi - 1) \quad (16.16.6)$$

This shows that with ϕ equal to zero, ψ is also zero, and therefore the array factor becomes the maximum value as given by Eq. (16.16.4). When $\phi = 180^\circ$, the angle ψ becomes $-4\pi s/\lambda$, which, for $s = \lambda/2$, yields $\psi = -2\pi$. This would also give a maximum, and for proper end-fire operation with only a single beam, this spacing of $\pi/2$ must be avoided. A common arrangement used in practice is to space the elements by $\pi/4$ and directly feed them in parallel as shown in Fig. 16.16.2(b). Depending on the number of elements, some minor backlobes may exist, but these are generally much smaller than the forward lobe. For a given number of elements, the end-fire array does not produce as narrow a beam as the broadside array.

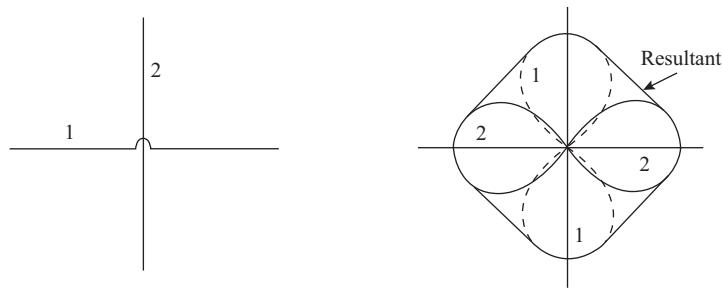
Both the broadside and the end-fire arrays may be used for any frequency band. Because of their physical size, they are usually limited to the HF bands and higher, but they have been used in the LF bands for point-to-point communications. Overseas broadcasting stations in the MF and HF bands frequently use them as well for repeated broadcasting to the same distant area.



(b)



(b)



(c)

Figure 16.16.2 Driven arrays. (a) Broadside array with radiation pattern. (b) End fire array with radiation pattern. (c) Turnstile antenna showing its radiation pattern as the sum of two dipole patterns rotated at right angles to one another.

Turnstile Antenna

Figure 16.16.2(c) shows a simple *turnstile antenna*, consisting of two half-wave dipoles placed at right angles to each other and fed 90° out of phase with each other. This results in the two dipole patterns combining in

the manner shown in the figure, producing an almost circular pattern in the plane of the turnstile. The pattern also has the feature that it is polarized in that same plane, so that if it were mounted in the horizontal plane, the antenna would radiate horizontally polarized waves about equally well in all directions along the ground. Several of these turnstiles may be stacked along a vertical axis and phased so as to improve the radiation directivity along the ground (that is, in the plane of polarization). This type of antenna is frequently used for television broadcasting in the VHF–UHF bands.

16.17 Parasitic Arrays

Parasitic Reflectors

Parasitic elements are secondary antennas that are placed in close proximity to the main or driven antenna. They are not directly fed, but have currents induced in them from the main element (or from the received wave in the case of a receiving antenna). The secondary antennas are tuned so as to cause a lagging or leading phase shift in energy that is reradiated from them, and this changes the radiation pattern of the main antenna, as shown in Fig. 16.17.1(a) and (b).

The reflector element is placed behind the driven dipole and is made about 5% longer than the driven ($\lambda/2$) dipole so that it is inductive. Spacing is adjusted until maximum radiation occurs along the normal in front of the dipole (array axis) as shown in Fig. 16.17.1(a). Optimum spacing, which is found by experiment, is usually about 0.15λ .

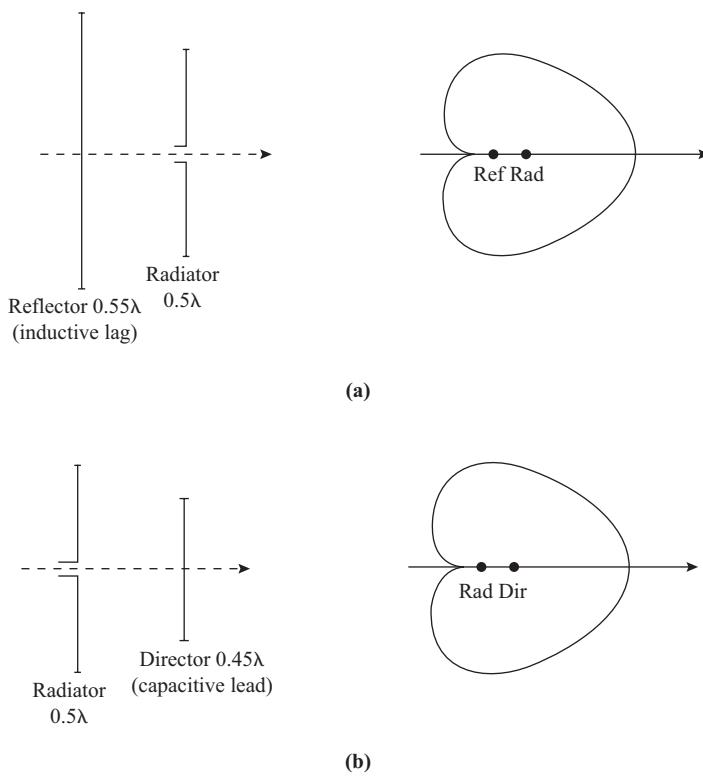


Figure 16.17.1 Effects of parasitic elements on the radiation pattern of a dipole. (a) Reflector. (b) Director.

Parasitic Directors

The director element, which is placed in front of the driven dipole, is made about 5% shorter than the dipole so that it is capacitive. It is spaced to provide maximum radiation in the forward direction, and optimum spacing is again found experimentally to be about 0.15λ . The pattern is shown in Fig. 16.17.1(b).

Yagi–Uda Array

The *Yagi–Uda* (or simply the *Yagi*) antenna is a parasitic array comprising a driven half-wave dipole antenna that is usually a folded dipole, a single parasitic reflector, and one or more (up to 13) director elements with each director cut to act as if the previous one were the driven element, so that the whole structure tapers in the direction of propagation. The structure is illustrated in Fig. 16.17.2(a). All the elements are electrically fastened to the conducting, grounded central support rod. This has no effect on the currents since the support point in the center of each element is at a current node.

Only one reflector need be used, since the addition of a second or third reflector adds practically nothing to the directivity of the structure. The directive gain is improved considerably by the addition of more directors to give directive gains from about 7 dB for a three-element Yagi to about 15 dB for a five-element Yagi. The pattern consists of one main lobe lying in the forward direction along the axis of the array, with several very minor lobes in other directions. Polarization is in the direction of the element axes.

A folded dipole is often used as the driven element to raise the antenna terminal impedance. The parasitic elements are quite closely coupled to the driven element, which results in a lowering of the effective radiation resistance. The fourfold multiplication of the folded dipole brings the impedance level back up to reasonable levels. A Yagi with a folded dipole would have a terminal resistance of 200 to 300 Ω .

This type of antenna would be very bulky at low frequencies; hence, it is used most often in the VHF range. Radio amateurs have built Yagis for the 20-m band, but the structure is large and cumbersome. Its high directional gain makes it ideal for point-to-point fixed-frequency communications networks, either at terminal

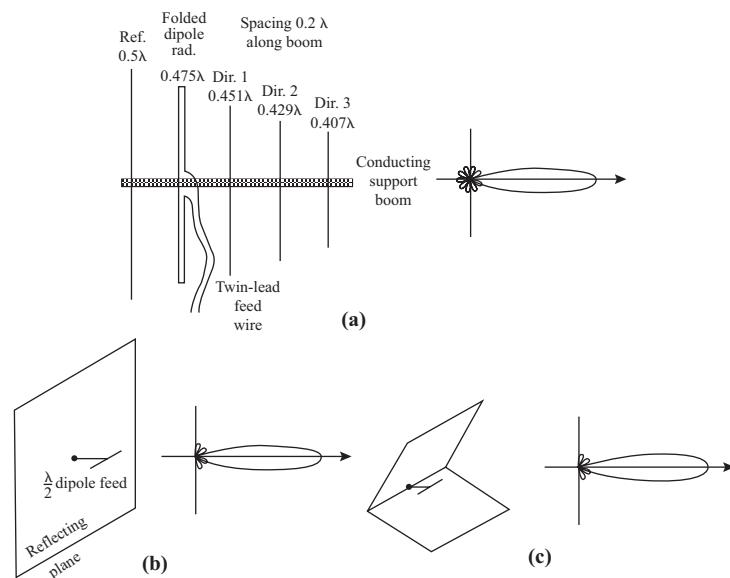


Figure 16.17.2 (a) Five-element Yagi–Uda array. The structure is shown with its dimensions and radiation pattern. (b) Plane reflector with a dipole feed. (c) Corner reflector with a dipole feed.

or at repeater stations. They have also been used for base stations on mobile communications systems where the working area is strung out along a line, such as railroads, highways, or pipelines.

Plane Reflector Arrays

At UHF it is common practice to use a *plane reflecting surface*, either a flat surface or a corner of two surfaces, in place of the single reflector element of the Yagi. The reflecting surface (see Fig. 16.17.2[b]) must be at least one wavelength across in each direction and can be much larger. It may be a solid metal surface, or it may be wire mesh or a network of interconnected metal rods. Somewhat sharper directivity is obtained with the corner reflector, shown in Fig. 16.17.2(c), but only in the plane across the fold.

The plane reflector is arranged so that the driven dipole (or directional antenna of any type) is mounted a quarter-wavelength ahead of the reflector surface. A mirror image of the driven antenna occurs a quarter-wavelength behind the surface and appears to radiate a wave that arrives at the driven antenna in phase with the driving antenna radiation, but delayed one period. The resulting radiation pattern appears to be the vector sum of the radiation from two in-phase antennas, one of which is the image.

16.18 VHF–UHF Antennas

Discone Omni

The *discone antenna* is designed to radiate an omnidirectional pattern in the horizontal plane, with vertical polarization. It is a broadband antenna with usable characteristics over a frequency range of nearly 10 : 1. It is usually designed to be fed directly from a $50\text{-}\Omega$ coaxial line and is mounted directly on the end of that line. The discone is illustrated in Fig. 16.18.1(a).

This type of antenna is ideal for base-station operation for urban mobile communications systems, since it gives a good omnidirectional pattern, is physically very compact and rugged, and is quite inexpensive to construct. Its directional gain along the horizontal plane is comparable to that of the dipole antenna.

Helical Antenna

The radiator element of a *helical antenna* is basically a coil of wire. If the helix diameter is much less than one wavelength, its length less than one wavelength, and it is center-fed, the whole structure will behave very much like a compact dipole antenna, radiating in the “normal” mode. Since the current wave must propagate along the helix conductor, its actual velocity along the axis will be much less than free-space velocity. The velocity will be determined by the diameter and pitch of the helix coil, and the half-wave resonant length of the helix will physically be very much less than the free-space half-wavelength. For most combinations of helix, the polarization is elliptical. This type of antenna is sometimes used in locations where it is not possible to mount a full-sized dipole, such as in urban areas or on rooftops.

If the helix diameter is made approximately one wavelength, and the helix made several wavelengths long and end-fed, the helix radiates in an end-fire mode, producing a narrow beam of circularly polarized waves. The 3-dB beam width obtainable with single helices is on the order of 15° to 30° . When used in this mode, the radiator is usually end-fed and a plane reflector is placed behind the feed end. The result is a highly directional antenna that is physically compact and that can be easily mounted on a movable table for tracking moving sources. Arrays of end-fire helices are often used for tracking satellites. These structures are much less expensive than the large parabolas used for radio astronomy and have been very popular with amateur satellite trackers. The end-fire helix is illustrated in Fig. 16.18.1(b).

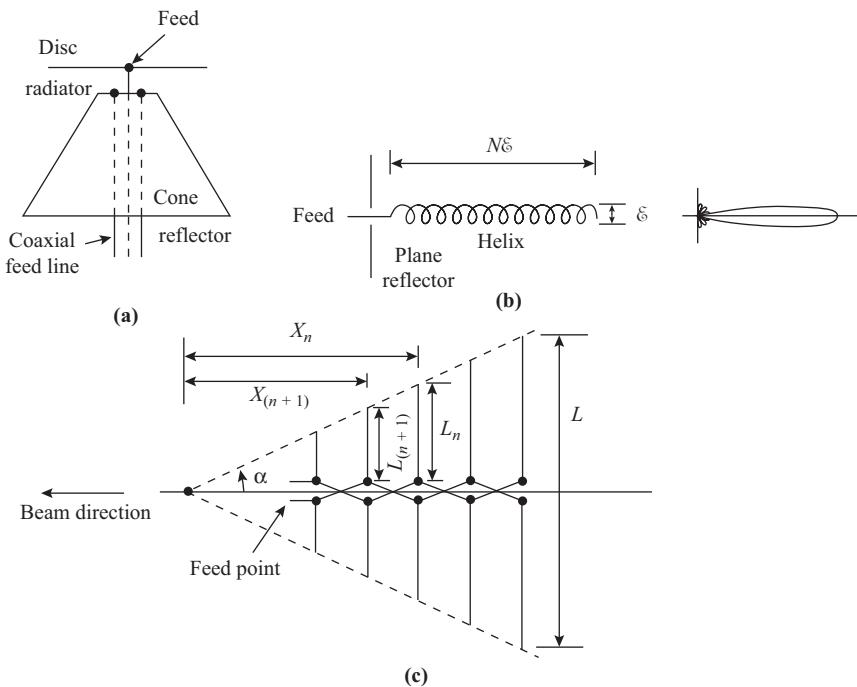


Figure 16.18.1 VHF–UHF antennas. (a) Discone omnidirectional antenna. (b) End-fire helix antenna. (c) Log-periodic dipole array.

Log Periodic Antenna

The *log periodic antenna* is basically an array of dipoles, fed with alternating phase, lined up along the axis of radiation. The structure is illustrated in Fig. 16.18.1(c). The element lengths and their spacing all conform to a ratio, given as

$$\tau = \frac{L_{(n+1)}}{L_n} = \frac{X_{(n+1)}}{X_n} \quad (16.18.1)$$

Also, the angle of divergence is given as

$$\alpha = \tan^{-1} \left(\frac{L_n}{X_n} \right) \quad (16.18.2)$$

The open-end length L must be larger than $\frac{1}{2}\lambda$ if high efficiency (90%) is to be obtained.

This antenna has the unique feature that its impedance is a periodic function of the logarithm of the frequency, hence its name. The antenna characteristics are broadband, and it has the directional characteristics of a dipole array. This type of antenna is often used for mobile-base-station operations, where many channels must be handled over a single antenna system with good directive characteristics.

16.19 Microwave Antennas

Horns

Radio waves can be radiated directly from the end of a waveguide in the same way as from the end of an open transmission line. The end of the waveguide represents an abrupt transition from the characteristic impedance of the waveguide into that of free space, and the radiation resulting is neither efficient nor very directive. This state of affairs can be improved considerably by flaring out the end of the waveguide to form a hornlike structure. A gradual transition can thus take place as the wave passes from the mouth of the horn.

Narrow-mouthed horns with long flare sections produce sharper beams than shallow, wide-mouthed ones. Also, the wider-mouthed horns tend to produce a wavefront with a distinct curvature, which is undesirable. The ideal would be for the waves to leave the horn with a completely planar wavefront, and to accomplish this a focusing mechanism, such as a curved reflector or a lens, may be used with the horn.

Three types of horns are shown in Fig. 16.19.1. The first is the sectoral horn, which is flared in only one plane [Fig. 16.19.1(a)]; the second is the pyramidal horn, which is flared in both planes [Fig. 16.19.1(b)]. Both of these are used with rectangular waveguides. The third type is conical [Fig. 16.19.1(c)] and is used with a circular waveguide to produce a circularly polarized beam. Horn-type antennas do not provide very high directivity but are of simple, rugged construction. This makes them ideal as primary feed antennas for parabolic reflectors and lenses.

The choice of horn dimensions is dependent on the desired beam angle and directive gain and involves specification of the ratio of flare length to wavelength L/λ and flare angle ϕ , shown in Fig. 16.19.1(d).

Paraboloidal Reflector Antenna

The most widely used antenna for microwaves is the *paraboloidal reflector antenna*, which consists of a primary antenna such as a dipole or horn situated at the focal point of a paraboloidal reflector, as shown in Fig. 16.19.2. The mouth, or physical aperture, of the reflector is circular, and the reflector contour, when projected onto any plane containing the focal point F and the vertex V , forms a parabola as shown in Fig. 16.19.2(b). The path length $FAB = FA'B'$ for this curve, where the line BB' is perpendicular to the reflector axis. The important practical implication of this property is that the reflector can focus parallel rays onto the focal point, and, conversely, it can produce a parallel beam from radiation emanating from the focal point. Figure 16.19.2(c) illustrates this. An isotropic point source is assumed to be situated at the focal point. In addition to the desired parallel beam being shown, it can be seen that some of the rays are not captured by the reflector, and these constitute *spillover*. In the receive mode, spillover increases noise pickup, which

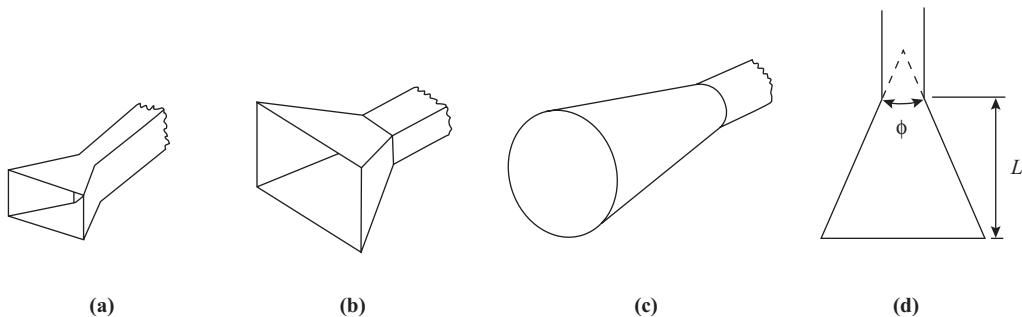


Figure 16.19.1 Microwave horn antennas. (a) Sectoral horn. (b) Pyramidal horn. (c) Conical horn. (d) Horn flare dimensions.

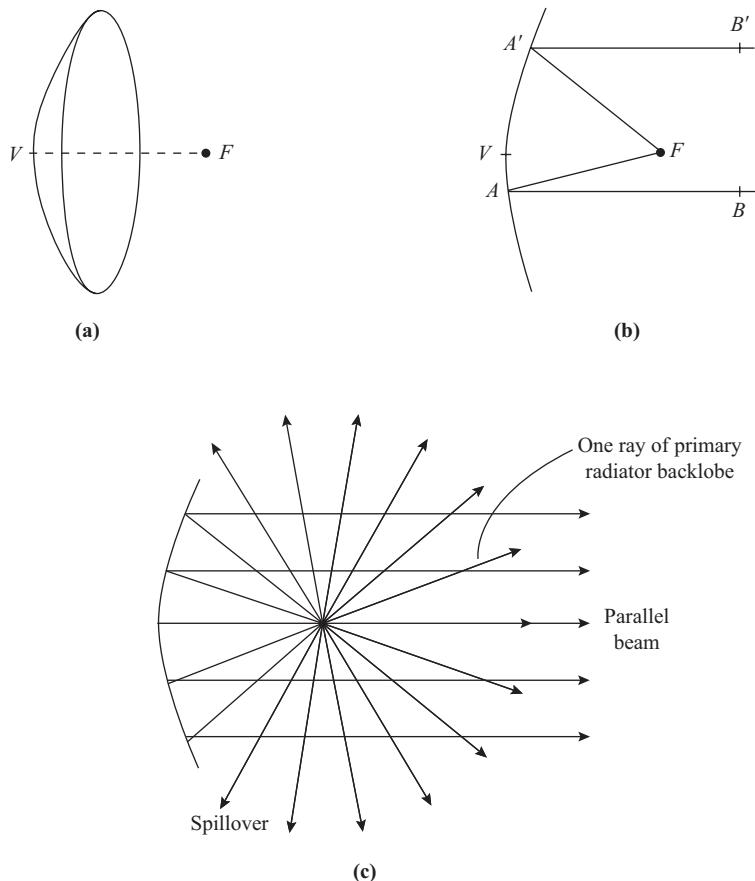


Figure 16.19.2 (a) Parabolic reflector. (b) Parabola. (c) Radiation from the paraboloid reflector and primary radiator.

can be particularly troublesome in satellite ground stations. Also, some radiation from the primary radiator occurs in the forward direction in addition to the desired parallel beam. This is termed *backlobe* radiation since it is from the backlobe of the primary radiator. Backlobe radiation is undesirable because it can interfere destructively with the reflected beam, and practical radiators are designed to eliminate or minimize this. The isotropic radiator at the focal point will radiate spherical waves, and the paraboloidal reflector converts these to plane waves. Thus, over the aperture of an ideal reflector, the wavefront is of constant amplitude and constant phase.

The directivity of the paraboloidal reflector is a function of the primary antenna directivity and the ratio of focal length to reflector diameter, f/D . This ratio, known as the aperture number, determines the angular aperture of the reflector, 2Ψ [Fig. 16.19.3(a)], which in turn determines how much of the primary radiation is intercepted by the reflector. Assuming that radiation from the primary antenna is circularly symmetric about the reflector axis ($F - V$) and is confined to angles ψ in the range $-\pi/2 < \psi < \pi/2$, it is found that the effective area is given by

$$A_{\text{eff}} = AI(\theta) \quad (16.19.1)$$

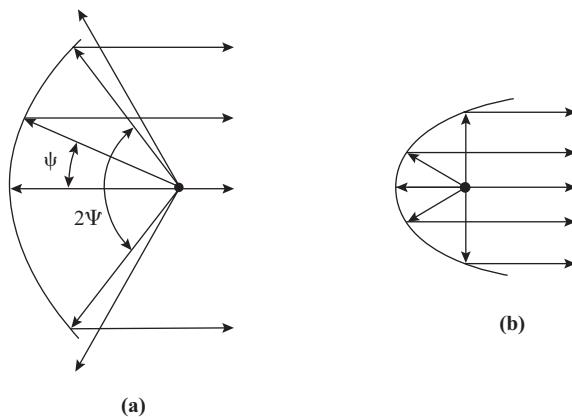


Figure 16.19.3 (a) Focal point outside the reflector. (b) Focal point inside the reflector.

where $A = \pi D^2/4$ is the physical area of the reflector aperture, and $I(\theta/2)$ is a function, termed the *aperture efficiency* (or *illumination efficiency*), which takes into account both the radiation pattern of the primary radiator and the effect of the angular aperture. With the focal point outside the reflector, as shown in Fig. 16.19.3(a) (which requires $f/D > 1/4$), the primary radiation at the perimeter of the reflector will not be much reduced from that at the center, and the reflector illumination approaches a uniform value. This increases the aperture efficiency, but at the expense of spillover occurring. Making f/D too large increases spillover to the extent that aperture efficiency then decreases. Reducing f/D to less than $1/4$ places the focal point inside the reflector, as shown in Fig. 16.19.3(b). Here, no spillover occurs, but the illumination of the reflector tapers from a maximum at the center to zero within the reflector region. This nonuniform illumination tends to reduce aperture efficiency. Also, placing the primary antenna too close to the reflector results in the reflector affecting the primary antenna impedance and radiation pattern, which is difficult to take into account. It can be shown that the aperture efficiency peaks at about 80%, with the angular aperture ranging from about 40° to 70° depending on the primary radiation pattern. The relationship between aperture number and angular aperture is

$$\frac{f}{D} = 0.25 \cot\left(\frac{\Psi}{2}\right) \quad (16.19.2)$$

Typically, for an angular aperture of 55° , the aperture number is

$$\frac{f}{D} = 0.25 \times 1.92 = 0.48$$

This shows that the focal point should lie outside the mouth of the reflector, since f/D is then greater than $1/4$. Satisfactory results are obtained in practice if the main lobe of the primary antenna intercepts the perimeter of the reflector at the -9 to -10 dB level as shown in Fig. 16.19.4.

On substituting $\pi D^2/4$ for A in Eq. (16.19.1) and using Eq. (16.8.3) for gain, we get

$$G = I(\theta) \left(\frac{\pi D}{\lambda} \right)^2 \quad (16.19.3)$$

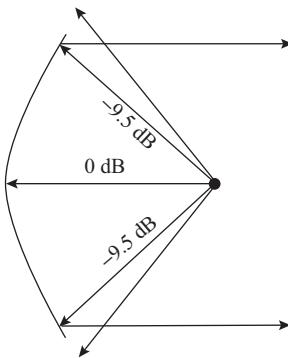


Figure 16.19.4 Edge illumination from primary antenna is -9 to -10 dB below that at the center.

The beamwidth also depends on the primary radiator and its position. In practice, it is found that for most types of feed the -3 -dB beamwidth is given approximately by

$$\text{BW}_{(-3-\text{dB})} \cong \frac{70\lambda}{D} \text{ degrees} \quad (16.19.4)$$

and the beamwidth between nulls by

$$\begin{aligned} \text{nulls BW} &= 2[\text{BW}_{(-3-\text{dB})}] \\ &= \frac{140\lambda}{D} \text{ degrees} \end{aligned} \quad (16.19.5)$$

EXAMPLE 16.19.1

Find the directivity, beamwidth, and effective area for a paraboloidal reflector antenna for which the reflector diameter is 6 m and the illumination efficiency is 0.65. The frequency of operation is 10 GHz.

SOLUTION

$$\lambda = \frac{c}{f} = \frac{300 \times 10^6}{10 \times 10^9} = 0.03 \text{ m} = 3 \text{ cm}$$

$$A = \frac{\pi D^2}{4} = \frac{3.14 \times 6^2}{4} = 28.26 \text{ m}^2$$

$$A_{\text{eff}} = 0.65A = 18.4 \text{ m}^2$$

$$D_0 = \frac{4\pi}{\lambda^2} A_{\text{eff}} = 257,000 \text{ (54.1 dB)}$$

$$\text{BW}_{(-3-\text{dB})} = \frac{70\lambda}{D} = \frac{70 \times 0.03}{6} = 0.35^\circ$$

$$\text{BW}_{(\text{null})} = 2 \times 0.35 = \mathbf{0.70^\circ}$$

Variations on the Parabolic Reflector

The *parabolic reflector* is a favorite antenna for fixed point-to-point microwave communications systems. It is relatively simple in construction, and unless large in size, it is quite inexpensive. Huge steerable parabolic dishes have been built for use with the radio telescopes, up to 200 ft in diameter, and mounted on a movable turret that allows rotation in both the horizontal and vertical directions to allow the tracking of moving targets such as satellites and radio stars.

Antennas used for radioastronomy must utilize all their area for reception to get the highest efficiency and the lowest noise figure. Special feed systems are used so that the feed antenna is reduced in size or physically located out of the path of the incoming radiation. Two types of feed are shown in Fig. 16.19.5. The first of these uses a dipole antenna, which normally radiates outside the parabola as well as onto it, but has a spherical reflector placed directly behind the dipole to prevent direct radiation. The backlobe radiation is reflected back at the parabola and is added to the main portion of the radiation. Some tuning is necessary since the reflector position is different for different frequencies.

The second method is known as the *Cassegrain feed system* [Fig. 16.19.5(b)]. The horn feed antenna, the paraboloid reflector, and the hyperboloid subreflector have a common axis of symmetry as shown, and the virtual focal point of the hyperboloid is coincident with the focal point of the paraboloid. Radiation reflected off the subreflector illuminates the main reflector approximately uniformly and, equally important, spillover at the edges is low. This is of particular advantage in low-noise receiving systems, where large spillover results in high noise levels.

Two types of modified horns also use parabolic surfaces. The first of these is the *Cass horn* [Fig. 16.19.5(c)]. Incoming radiation is reflected from a large lower parabolic section that has a rectangular sectional front and forms the bottom of the horn, up to a smaller parabolic surface that forms the top of the horn, and then to a horn antenna located at the focal point of the combined parabolas. This horn has the advantage that incoming radiation is not blocked by the feed structure, but it is much more complex to build.

The second horn is the *hog horn* [Figure 16.19.5(d)]. This is basically a large horn antenna with a parabolic reflector mounted directly in front of it and oriented to radiate at right angles to the axis of the horn. It is often used for microwave communication links, such as the Bell TD-2 system, because it is compact and simple in construction while giving the high directivity of the parabola and also high efficiency. It also has the advantage that it may be rotated at the neck on a sliding joint, allowing steering of the beam without undue manipulation.

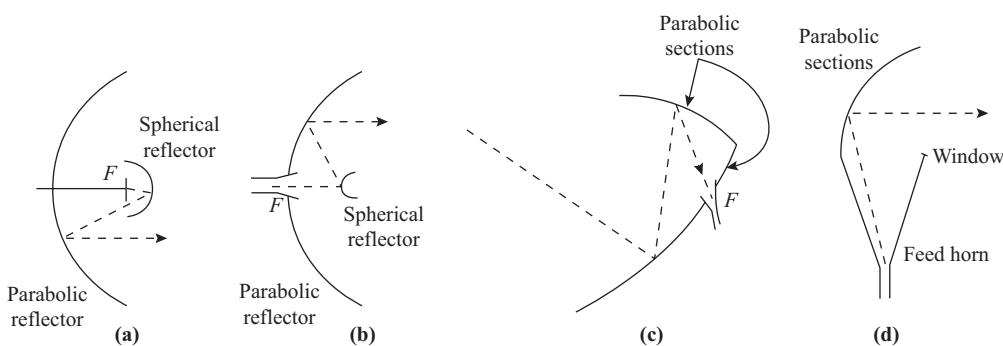


Figure 16.19.5 Methods of feeding microwave antennas. (a) Parabolic reflector fed from a dipole and a small spherical reflector. (b) Cassegrain-feed parabola fed from a horn and a small paraboloid reflector. (c) Cass horn, (d) Hog horn.

Dielectric Lens Antennas

Electromagnetic radiation is refracted when it passes through a surface separating a zone of lower dielectric constant from one of higher dielectric constant in exactly the same manner that light is refracted. The angles of incidence and refraction are related by the modified version of Snell's law, which states [referring to Fig. 16.19.6(a)] that

$$\frac{\sin \phi_r}{\sin \phi_i} = \sqrt{\frac{\epsilon_{ri}}{\epsilon_{rr}}} = \frac{1}{n} \quad (16.19.6)$$

for waves entering the region of high dielectric constant, where n is the refractive index of the material. The refractive property is reciprocal, and the same relationships obtain when the radiation passes from the higher dielectric medium into the lower dielectric medium, except that the subscripts i and r are interchanged.

The material used for the lens is usually one of the high-dielectric plastics, such as polystyrene or Teflon.

Figure 16.19.6(b) illustrates the principle of the collimating lens, which is used to make a diverging beam of radiation into one traveling in only one direction with a planar wavefront. The lens in this case is a convex one. Radiation along the axis of the lens passes through both surfaces at right angles, so no refraction takes place. Radiation at an angle from the axis is incident at the curved interface at an angle other than normal and is refracted toward the normal as it passes into the lens, at A' . The curvature of the lens is such that after refraction the rays are all parallel to the axis.

Radiation at the angle along FA' takes slightly longer to reach the refracting surface, arriving with a slight time lag compared to that along the axis FA . However, the velocity of propagation within the dielectric

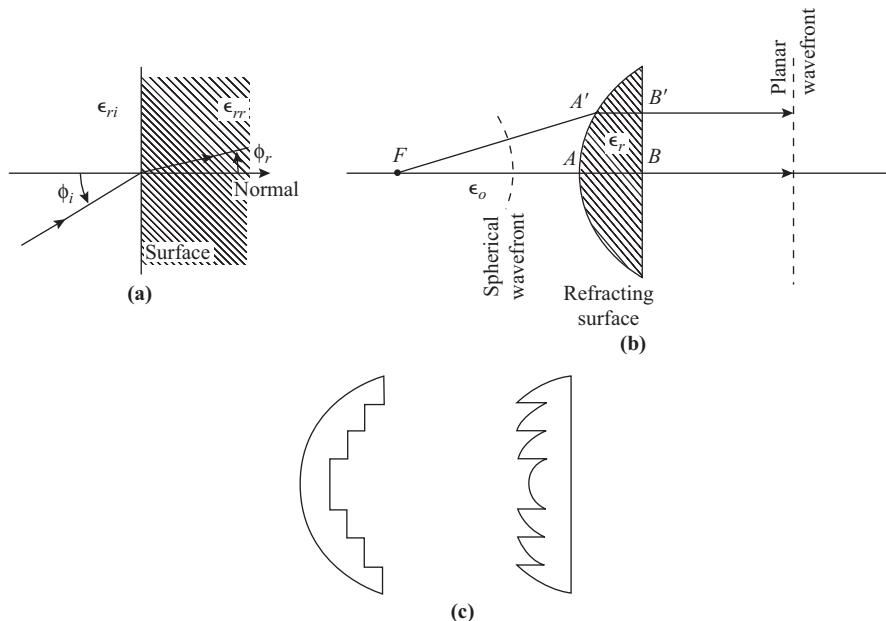


Figure 16.19.6 Dielectric lens antennas. (a) Snell's law of refraction. (b) Principle of the collimating lens. (c) Cross sections of two types of zoned lenses.

is slower than that in air, and a compensating delay occurs in the radiation along the axis AB as compared to that along $A'B'$, so that all the radiation arrives at the flat interface BB' in phase. The result is a planar wavefront leaving the lens.

Many types of lens systems can be used for the collimating or paralleling function. Two of these are shown in Fig. 16.19.6(c). These are *zoned lenses*, which are basically the same in function as the convex lens illustrated in Fig. 16.19.6(b). However, the zone structure allows a large reduction in the volume of dielectric material that must be used to make the lens, with a corresponding savings in cost and weight. This saving is made at a slight sacrifice of directivity.

Again, ideally, the lens should be fed with even illumination over its entire surface to achieve maximum efficiency and gain. The horn antenna is the most popular method of feed and most closely approximates the even-illumination requirement. For the higher microwave frequencies, the lens makes a very compact, highly directive narrow-beam antenna that is popular for applications such as portable communication links and mobile radar systems of the type used for vehicle-speed monitoring.

Slot Antennas

When a slot in a large metallic plane is coupled to an RF source, it behaves like a dipole antenna mounted over a reflecting surface. Figure 16.19.7(a) shows a rectangular slot in a large plane. The slot is coupled to a line or feed waveguide in such a manner that the E field lies along the short axis of the slot, as indicated. At microwave frequencies, the slot may be energized as described in Section 14.2. At lower frequencies, a transmission line may be connected directly across the slot as shown in Fig. 16.19.7(a).

The behavior of this *slot antenna* can be shown to be equivalent to a complementary antenna in which the slot is replaced by a sheet dipole with a reflector behind it in free space, as shown in Fig. 16.19.7(b). The dimensions of the slot are usually such that the long axis is approximately a half-wavelength at the operating frequency. The resulting radiation pattern is similar to that for the dipole with a reflector.

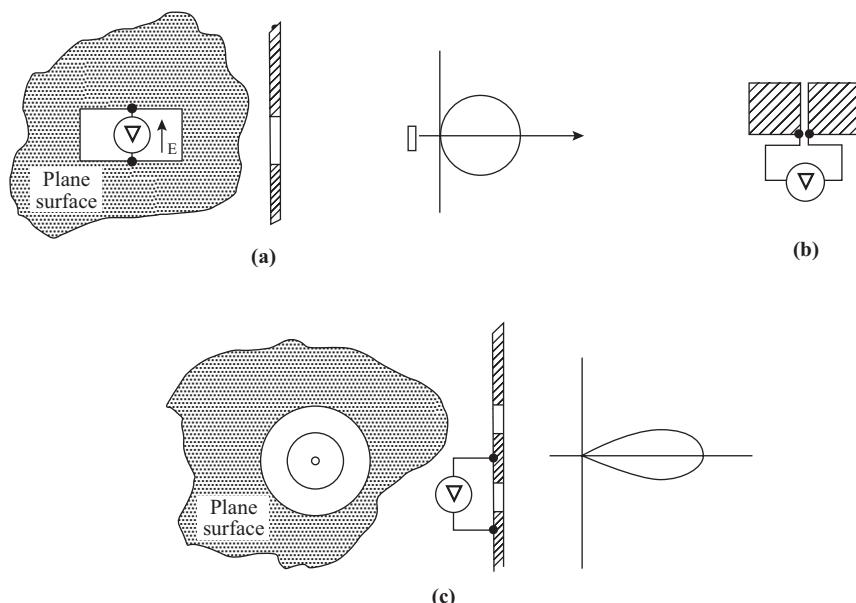


Figure 16.19.7 Slot antennas. (a) Rectangular slot. (b) Complementary dipole equivalent of a rectangular slot. (c) Annular slot.

The feed impedance of the slot can be calculated by first calculating the impedance for the complementary dipole. The actual slot impedance is related to the dipole impedance by the complementary characteristic, which states that

$$Z_r \times Z_c = \left(\frac{Z_0 \text{ space}}{2} \right)^2 = (60\pi)^2 \quad (16.19.7)$$

where Z_r is the radiation impedance of the slot and Z_c is the radiation impedance of the complementary (dipole) antenna. The shape of the slot does not have to be rectangular, but can be any convenient shape. Rectangular and circular (annular) slots are favored because they are easy to make and relatively easy to analyze. Figure 16.19.7(c) shows an annular slot, which produces a narrow beam of radiation.

Arrays of slots are ideal for use in aircraft. The slots can be formed directly in the metal skin of the aircraft and then windowed with a dielectric material such as polystyrene. The smooth surface produced does not interfere with the streamlining of the aircraft. Phasing the feed to the slots allows production of a beam that may be swept through a wide angle without physically moving the structure, thus allowing its use for mobile radar systems.

A variation of the slot antenna is the *notch antenna*, where an appropriately shaped notch is cut out of the edge of a large metal surface and connected to an RF source. Again, the chief use is on aircraft, where the notches can be made in the edges of the wing surfaces and filled with dielectric material to make them aerodynamically smooth.

PROBLEMS

- 16.1.** (a) The terminal input current to an antenna is $2/11^\circ$ A when the terminal voltage is $100 \angle 100/0^\circ$ V. Determine the antenna impedance. (b) The voltage reflection coefficient measured on a 50Ω transmission line feeding an antenna is $0.1/5^\circ$. Determine the antenna impedance.
- 16.2.** Assuming 100% efficiency for the antennas in Problem 16.1, determine the radiation resistance in each of cases (a) and (b), and the power transmitted in (a).
- 16.3.** Derive Eq. (16.2.7) of the text.
- 16.4.** An antenna of impedance $Z_A = 45 - j10 \Omega$ is connected directly to a 50Ω transmission line. Determine the matching efficiency.
- 16.5.** The (x, y, z) coordinates for a point in space are given as $(10, 5, -2)$. Determine the polar spherical coordinates (d, θ, ϕ) as shown in Fig. 16.3.1.
- 16.6.** A paraboloidal reflector antenna has a diameter equal to 100λ . Determine the distance at which the far-field zone is the only effective component of the total radiated field. The frequency of operation is 10 GHz.
- 16.7.** Two $\frac{1}{2}\lambda$ dipoles are arranged for transmission and reception as shown in Fig. 16.5.4. The receiving antenna is in the plane of the incoming wavefront, and the maximum induced emf is measured as $10 \mu\text{V}$. Calculate the induced emf when the receiving antenna is rotated (a) 30° , (b) 60° , and (c) 90° in the plane of the wavefront.
- 16.8.** A power of 100 W is radiated from an isotropic radiator. Calculate the power radiated per unit solid angle and the power density at a distance of 5 km from the antenna.
- 16.9.** An isotropic antenna radiates energy equally in all directions. The total power delivered to the radiator is 100,000 W. Calculate the power density and electric field intensity at distances of (a) 100 m, (b) 1 km, (c) 100 km, and (d) 1000 km. Plot the power density against distance on log-log paper.

- 16.10.** Calculate the directivity for the antennas for which the following specifications apply: (a) power gain 10³:1, efficiency 90%; (b) power gain 45 dB, efficiency 90%.
- 16.11.** The radiation pattern for an antenna in the meridian plane is given by $\sin^4 \theta$. Plot this function (a) on polar graph paper and (b) on linear graph paper. (c) Determine the -3-dB and the -10-dB beamwidths.
- 16.12.** An antenna has a gain of 35 dB at a frequency of 300 MHz. Calculate the effective area.
- 16.13.** Calculate the effective length of an antenna that has a directive gain over an isotropic antenna of 17 dB and a radiation resistance of 350 Ω at a frequency of 144 MHz. Use the relationship shown in Problem 16.14.
- 16.14.** By making use of Eq. (B.12) and the maximum power transfer theorem, show that the effective area and effective length of an antenna are related by

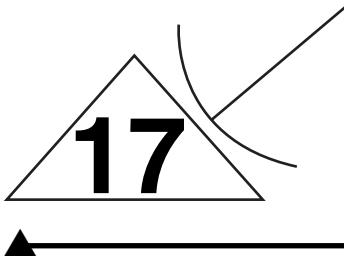
$$A_{\text{eff}} = \frac{30\pi\ell_{\text{eff}}^2}{R_{\text{rad}}}$$

- 16.15.** An elementary doublet has an electrical length of 0.0625λ and carries an RF current of 2.5 A rms. Calculate the field intensity at a point located 40 km from the doublet and at an angle of 25° from the main lobe of radiation.
- 16.16.** A directional antenna has an effective radiated power of 1.1 kW when it is fed with a terminal input power of 90 W. The radiation resistance is found to be 74 Ω at resonance, and the measured antenna current is 1.088 A rms. Find (a) the antenna efficiency; (b) the terminal resistance; (c) the antenna power loss; and (d) the antenna directive gain, in decibels, over an isotropic radiator.
- 16.17.** Calculate the current induced in the terminals of a vertical receiving antenna if it has a gain of 6 dB over an isotropic antenna, a terminal impedance of 35 Ω, a load impedance of 35 Ω, and it is in an electric field with an intensity of 10 μV/m at a frequency of 7 MHz.
- 16.18.** Calculate the capture area of the antenna in Problem 16.17.
- 16.19.** Plot the normalized radiation pattern for a $\frac{1}{2}\lambda$ -dipole antenna, using Eq. (16.11.4), on a polar graph for $0^\circ < \theta < 360^\circ$. Calculate values for each 10° displacement of θ.
- 16.20.** Calculate the 3-dB beamwidth of the $\frac{1}{2}\lambda$ antenna in Problem 16.19.
- 16.21.** Using the relationships given by Eqs. (16.11.8) and (16.11.9) and the results of Problem 16.14, calculate the effective length of a $\frac{1}{2}\lambda$ dipole.
- 16.22.** A loop antenna is made by winding 10 turns of wire on a 1-m² frame. It is located in a magnetic field of 0.015 μT, at 10 MHz and oriented for maximum signal strength. Find (a) the induced emf in the antenna, and (b) the terminal voltage if the antenna is tuned to resonance at 10 MHz, with a total resistance of 65 Ω in series with a 25-pF capacitor.
- 16.23.** A direction finder using a loop antenna is used to locate an illegal transmitter. The operator goes to location A, where he receives a good signal. He mounts his receiver so that the antenna faces due north and then rotates the antenna from 0° north clockwise through 48° to obtain a null. He moves the direction finder to a new location B, which he determines is exactly 2550 m due east of A, and obtains a new bearing by rotating the antenna 15° counterclockwise from 0° north. Compute the distance from point A and from point B to the transmitter.
- 16.24.** Calculate the effective length of a ferrite-rod receiving antenna that has 120 turns wound on a 1.40-cm-diameter ferrite rod that has a relative permeability of 160. Assume the length factor to be 0.75 and the frequency to be 1 MHz.
- 16.25.** Plot the array factor as a function of angle ϕ for a four-element broadside array for which the elements are spaced by $\frac{1}{2}\lambda$.

- 16.26.** Determine the current phasing required for an end-fire array for which the elements are spaced $\frac{1}{2}\lambda$. Plot the array factor as a function of angle ϕ for a four-element array with this spacing.
- 16.27.** Calculate the lengths of the elements and spacing for a five-element Yagi antenna for Channel 4 television (66 to 72 MHz). The effective length factor is to be 5%, and each element is to be 95% of the previous one in length.
- 16.28.** Calculate the element lengths for a 10-element log-periodic array if the smallest element is not less than 10% of the largest element, the angle of divergence $\alpha = 15^\circ$, and the longest element is cut for a frequency of 50 MHz.
- 16.29.** Calculate the angular aperture for a paraboloidal reflector antenna for which the aperture number is (a) 0.25, (b) 0.5, and (c) 0.6. Given that the diameter of the reflector mouth is 10 m, calculate the position of the focal point with reference to the reflector mouth in each case.
- 16.30.** (a) Specify the diameter of a parabolic reflector required to provide a gain of 75 dB at a frequency of 15 GHz. The area factor of the feed is 0.65. (b) Calculate the capture area of the antenna and its 3-dB beamwidth.
- 16.31.** For an antenna impedance of $Z_A = 45 + j15\Omega$, connected directly to transmission lines of impedances $50\Omega / 75\Omega / 120\Omega / 300\Omega$, and 600Ω , respectively. Obtain the *matching efficiency* in each case. (Hint: Write a MATLAB function file for this purpose). Repeat the above exercise, when the antenna impedance is $Z_A = 45 - j15\Omega$.
- 16.32.** A source has Unidirectional Cosine radiation intensity pattern given by: $U = U_{max} \cos\theta$, where U_{max} is the maximum radiation intensity. Plot, using MATLAB/Mathematica, the power pattern.
The total power radiated is given by: $P = U_{max} \int_0^{2\pi} \int_0^{\pi/2} \cos\theta \sin\theta d\theta d\phi$. Calculate the directivity, D of the source. Deduce that for a source with Bidirectional Cosine radiation intensity pattern, the directivity is half of the above case.
- 16.33.** A source has Sine (Doughnut) radiation intensity pattern given by: $U_{max} \sin\theta$, where U_{max} is the maximum radiation intensity. Plot, using MATLAB/Mathematica, the power pattern. The total power radiated is given by: $P = U_{max} \int_0^{2\pi} \int_0^{\pi} \sin^2\theta d\theta d\phi$. Calculate the directivity, D of the source.
- 16.34.** A short dipole coincident with the polar ($\theta = 0$) axis has a sine-squared radiation intensity pattern given by: $U = U_{max} \sin^2\theta$, where U_{max} is the maximum radiation intensity. Plot, using MATLAB/Mathematica, the power pattern. The total radiated power in the above case is given by:

$$P = U_{max} \int_0^{2\pi} \int_0^{\pi} \sin^3\theta d\theta d\phi$$
. Obtain the directivity, D of the dipole.
- 16.35.** A source has a unidirectional cosine-squared radiation intensity pattern given by: $U = U_{max} \cos^2\theta$, where U_{max} is the maximum radiation intensity. Plot, using MATLAB, the power pattern. The total radiated power in the above case is given by:

$$P = U_{max} \int_0^{2\pi} \int_0^{\pi/2} \cos^2\theta \sin\theta d\theta d\phi$$
. Obtain the directivity, D of the dipole.



Telephone Systems

17.1 Wire Telephony

The word *telephone* is derived from the Greek words *tele* meaning far and *phone* meaning sound. *Telephony* thus involves the conversion of sound signals into an audio frequency analog electrical signal, which can then be transmitted over an electric transmission system and then reconverted to sound pressure signals at the receiver end. The electrical signals may be transmitted by radio or by wire, and a system may well use both means to establish any given circuit. The wire telephone system was the earliest of such systems and still forms the backbone of modern communications.

The establishment of voice communications between customers at distant locations remains the main function of telephone systems. However, these systems are no longer just telephone systems. They must also provide many other services, primarily the transmission of a wide variety of computer data. Also, the process of transmitting the analog voice signals is being done in a digital manner, so the telephone systems are evolving into multipurpose digital transmission networks.

Telephone Circuits

Modern telephone systems operate on a duplex basis. A *duplex* transmission system allows simultaneous transmission in both directions. This may be accomplished by simply providing two separate circuits, one for each direction, resulting in an uneconomic duplication of facilities. Duplex in most telephone systems means simultaneous transmission over the same channel without the need for switching.

A telephone system must be able to transmit voice systems in both directions and it must provide a means of signaling from each terminal toward the other. This alert signaling originally only summoned the person at the far receiver, but this was rapidly changed to the process of signaling a central operator, who connected the desired line and then signaled the far end receiver. The same signal channel is used in automatic telephone systems to operate the automatic line switching apparatus at the central location so that an operator is no longer required. In modern digital switching systems, the signaling channel is provided over a completely separate digital network instead of using in-channel signaling.

Telephone systems that require switching at the central location to interconnect the calling and called party lines operate on a *loop* basis. The traditional telephone subscriber's loop consists of a pair of wires between the subscriber's location and the telephone switching center, a telephone *terminal set* (TS), and a *subscriber line interface circuit* (SLIC) at the switching location to supply battery current, ac signaling current, and a means of connecting the line to the switching machine.

Figure 17.1.1 shows a typical subscriber's loop circuit of the type used until recently. Signaling inward is done by dc switching, while signaling outward is done by an ac source to ring the subscriber's bell. The traditional subscriber's terminal consists of a handset containing a transmitter and an earphone receiver. The handset is connected to the line through an autotransformer (induction coil), which provides the necessary impedance matching between the line and the handset. The impedance ratios are adjusted so that the *sidetone* or signal feedback from the transmitter to the local receiver is minimized. A signal bell is connected across the line through the hook switch contact for incoming signaling, and a rotary interrupter dial switch pulses the dc loop current for outgoing signaling.

DC current to operate the transmitter is supplied from the central office battery over the loop wire pair. This current also passes through the pulsing contact *B* in the dial pulsing unit, which interrupts the current to transmit the coded information to select the proper address in the automated switching machine. Contact *A* closes to short-circuit the receiver while dialing is in progress so that the caller doesn't hear the clicks of the pulse train. It releases as the dial returns to its rest position.

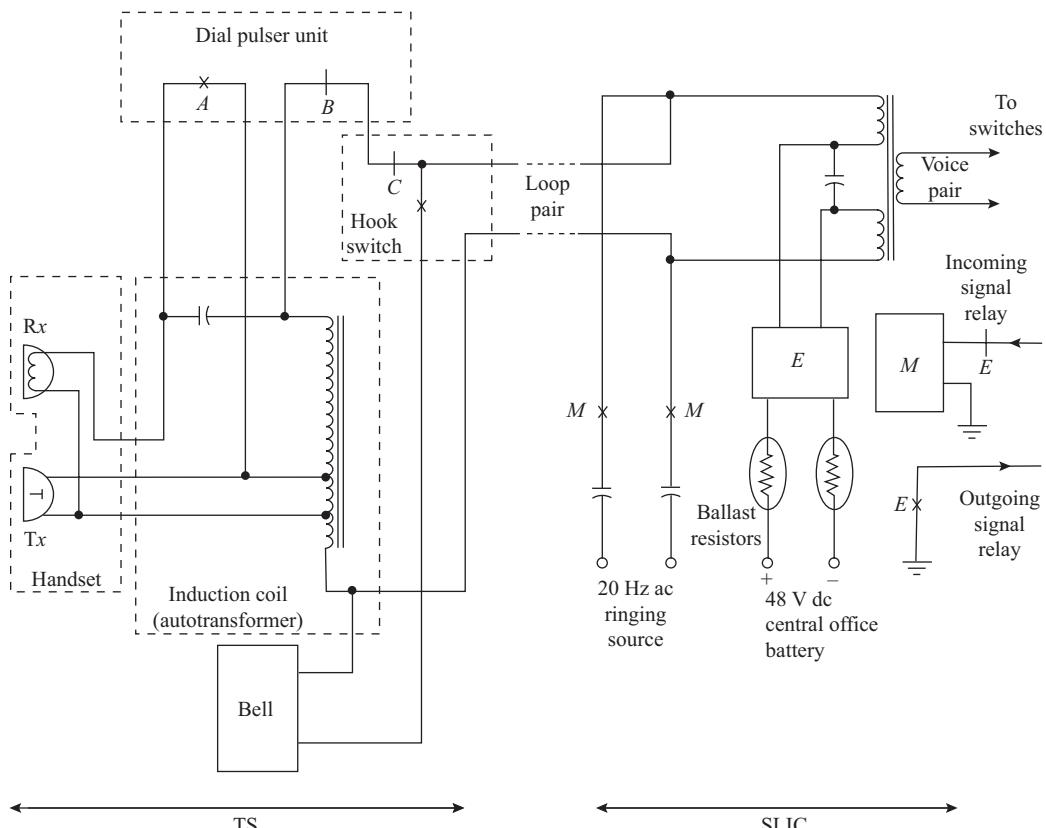


Figure 17.1.1 Telephone subscriber's loop circuit.

The SLIC at the switching location must provide several functions, commonly called *BORSCHT*, as follows:

- B: *Battery*. A 48-V central office battery supplies dc loop current (about 40 mA) to energize the voice transmitter and the outgoing signal circuit.
- O: *Ovvervoltage protection*. Bypassing or fusing or a combination to prevent damage from lightning, power line induction, or accidental power line contact.
- R: *Ringing*. Connection of the ac ringing signal to the loop for outgoing signaling (typically about 80 V at 20 Hz, interrupted for 4 s of every 6 s).
- S: *Supervision*. Detection of loop current to signal demand for service (or off-hook), termination of connection (or on-hook), and dial pulsing of routing codes for setting up the switches along the route (or detecting the same as outputs of a multifrequency Touch Tone signaling unit).
- C: *Coding/Decoding*. Provision of the PCM CODEC functions for connection to a digital switching system (only provided in recent digital switching machines).
- H: *Hybrid*. Two-wire to four-wire conversion before connection to a digital switching machine (as for the CODEC function).
- T: *Testing*. Provision to allow either automatic or manual testing of the subscriber's loop circuit from the central office.

The traditional SLIC shown in Fig. 17.1.1 provides only the BRST functions from the list above. The SLIC is located at the central office location and contains a bridging transformer that allows separation of the voice signal from the dc signaling current. The direct current for the loop is supplied from the common central office battery through the coils of the *E* relay and two ballast resistors when the book switch is off to complete the dc circuit at the telephone set. The relay passes the busy condition and dial pulse current interruptions to the switching machine. The ballast resistors compensate for different loop resistances to provide the same loop current regardless of loop length. A second *M* relay operates on signal from the switching machine to connect the ac bell ringing supply to the line. Ovvervoltage protection is provided separately by carbon block lightning arrestors and fuses on the cable entrance frame (not shown).

On originating a call the subscriber picks up the handset, which closes the hook switch to connect the handset to the line and disconnect the bells. This action closes the dc path in the loop, and the flow of direct current from the central office battery picks up the *E* relay. The *E* relay contacts signal the switching machine, indicating that the line is busy, and locks out the *M* ringing relay. The talking circuit is now energized. A slow operating relay also operates to hold the busy condition during dialing. Operating the dial sends several trains of 1 to 10 pulses at a rate of about 10 pulses per second through the *E* relay to select the proper switch combination in the switching machine. If the called line indicates a busy condition, a special busy tone is sent back on the voice circuit. If not busy, the *M* relay of the called loop is energized to send bursts of ac to ring the bell, and a ring tone is sent back on the voice circuit to let the caller know that ringing is proceeding.

When the demand signal from the switching machine energizes the *M* relay in the called loop, its contacts connect the ac ringing source to the loop line. This signal is an ac supply of 40 to 100 V at a low frequency (about 20 Hz) that is interrupted for about 4 s out of every 6 to ring the called party bell. When the called party lifts the receiver, the hook switch closes the dc path in the called loop to operate the *E* relay and disconnect the *M* relay to stop the ringing signal. The conversation can then proceed.

Touch Tone signaling has nearly completely replaced dial signaling since the advent of electronic switching offices. Touch Tone signaling uses a 2 out of 7 code to represent the 10 decimal digits and two other symbols. Each of the seven states is designated by the presence of a separate tone frequency within the voice band. The telephone dial unit is replaced by a key pad with 12 push-button switches. When one of the switches is depressed, two of the seven tone generators are energized and the two tones, one from the high

TABLE 17.1.1 Touch Tone Signaling Frequency Assignments

Low Group (Hz)	Codes ^a		
697	1	2	3
770	4	5	6
852	7	8	9
941	*	0	#
<i>High Group (Hz)</i>		1209	1336
		1477	

Source: *IT&T Reference Data for Radio Engineers*, 5th ed., Howard W. Sams & Co. Inc, Indianapolis, Indiana, 1969.

^aEach code designates one tone from the high group and one tone from the low group of frequencies.

group and one from the low group, are sent out audibly on the voice channel. Tone recognition equipment at the switching center decodes the called number and sets up the selection switches. Table 17.1.1 shows the tones used in the North American telephone system for tone dialing. The actual frequencies used have been carefully selected to minimize the possibility of accidental duplication by the voice signals.

Electronic Telephone

Electronic versions of the telephone set provide the same functions as the TS in Fig. 17.1.1, but the heavy electromagnetic components are replaced by a few integrated-circuit chips. Figure 17.1.2 shows the circuit for a complete electronic telephone set using three integrated-circuit chips and some discrete components. The electromagnetic ringer is replaced by a Motorola MC34017 ringer chip, which drives a piezoelectric sound transducer. It extracts ac ringing current from the telephone line through a coupling capacitor and impedance matching resistor. The chip contains a rectifier and regulator powered by the ring signal, so no

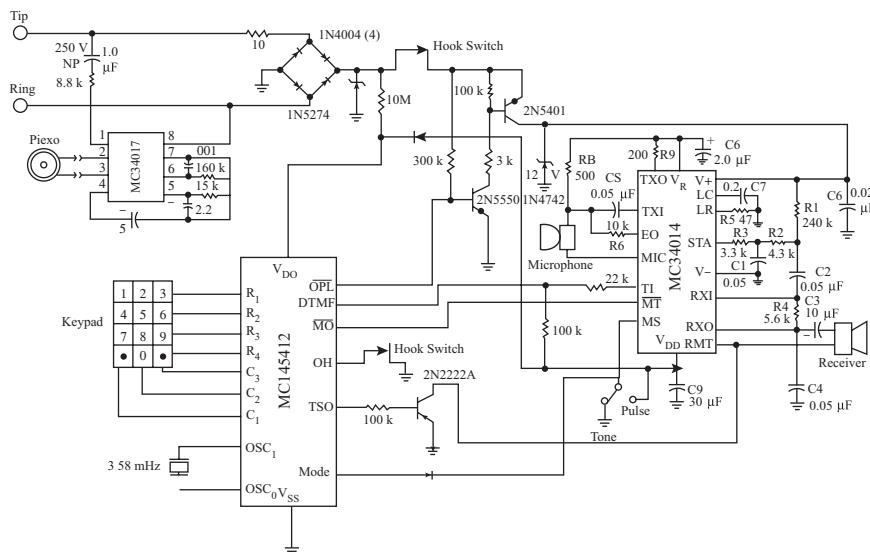


Figure 17.1.2 Complete electronic telephone set equipped with pulse/tone dialing. (Courtesy Motorola, Inc.)

external supply is needed except that signal. Three capacitors and two resistors allow setting of the frequencies used in a two-pitch warble of the sound produced. An oscillator on the chip produces the required frequency signals to drive the piezoelectric transducer. A mechanical hook switch serves to disconnect all but the ringer from the circuit when idle. The ringer circuit remains connected to the line during busy periods, but does not react to the low-level speech signals or the dc loop current.

The dc loop current is connected into the set through a bridge rectifier, which allows the set to function even if the loop conductors are reversed. A 1N5274 zener diode across the rectifier output acts to suppress overvoltage conditions. The current is then passed through a series current-limiting transistor (2N5401) to a shunt 12-V zener regulator to provide power to the remaining circuits. The output voltage contains both the dc voltage produced from the line current and the superimposed ac voice signals at levels below 0.5 V.

A Motorola MC34014 speech network chip replaces the electromagnetic transformer of the original telephone. It is powered from the 12 V produced from the line current. The voice signals are separated from the rectifier output by an *RC* coupling network (C_6, R_1, C_2); an internal amplifier and coupling network (R_3, C_1, R_2) add sidetone signal from the microphone, and an amplifier drives the electrostatic receiver transducer. An electret microphone powered from the chip is connected to an amplifier (through R_8, R_9, C_8) to drive the transmission line. A dc equalization amplifier senses the line current and adjusts the telephone terminal impedance to compensate for various loop lengths, so no change in signal levels occurs because of the different loop resistances, and ballast resistors are no longer needed in the central office.

A Motorola MC145412 repertory dialer chip replaces the rotary dialer on the original set. This chip takes switch inputs from a standard MFTD keyboard. It can be mode switched to produce either serial pulse dialing, as in the rotary dial, or it can produce the two-tone coded signals for *multi-frequency tone dialing* (MFTD) (see Table 17.1.2). A 3.58-MHz quartz crystal drives a frequency synthesizer circuit to produce the required tones, and the tone signal is passed to the microphone amplifier circuit for modulation on the loop current. In the pulse mode (used only with older switching machines), each digit is produced as a series of pulses that switches the series pass transistor of the power regulator to interrupt the line current. A logic signal is also produced to mute the receiver amplifier during dialing to eliminate annoying clicks or high-level tones. This chip also contains a memory system that can remember nine 18-digit numbers and a last number called for redialing. The memory is controlled by the keyboard.

Electronic Subscriber Line Interface Circuit

The subscriber line interface circuit described in Fig. 17.1.1 used transformers and electromagnetic relays to accomplish the BORST functions and was heavy and bulky. They were replaced in electronic systems by single-chip LSI circuits (large-scale integrated circuits), one for each line being serviced. One such chip is the Motorola MC3419. Some external components are required with the chip, and these are assembled onto a single line interface board, with the appropriate connections as shown in Fig. 17.1.3. For larger systems, several interface circuits are assembled on a common board to improve space economy.

The terminals on the right show connections to the backplane of the line frame. Inputs include battery positive (ground), battery negative (-48 V), four-wire analog ground, receive and transmit lines, ringing generator bus, ring enable input from the switching machine, and hook switch status out. The two-wire loop pair is connected at tip and ring.

One side of a pair of current mirror circuits in the chip sources positive battery current to a PNP Darlington pair (MJE271) to feed positive battery current to the tip line. The returning ring line is sunk through the MJE270 NPN Darlington driven from the other side of the current mirror pair. Feedback within the chip regulates the magnitude of the loop current. An ac voice signal on the four-wire receive input is amplified to produce a difference component between the tip and ring currents to modulate the loop current. A balance network prevents the received four-wire signal from being retransmitted on the four-wire output line.

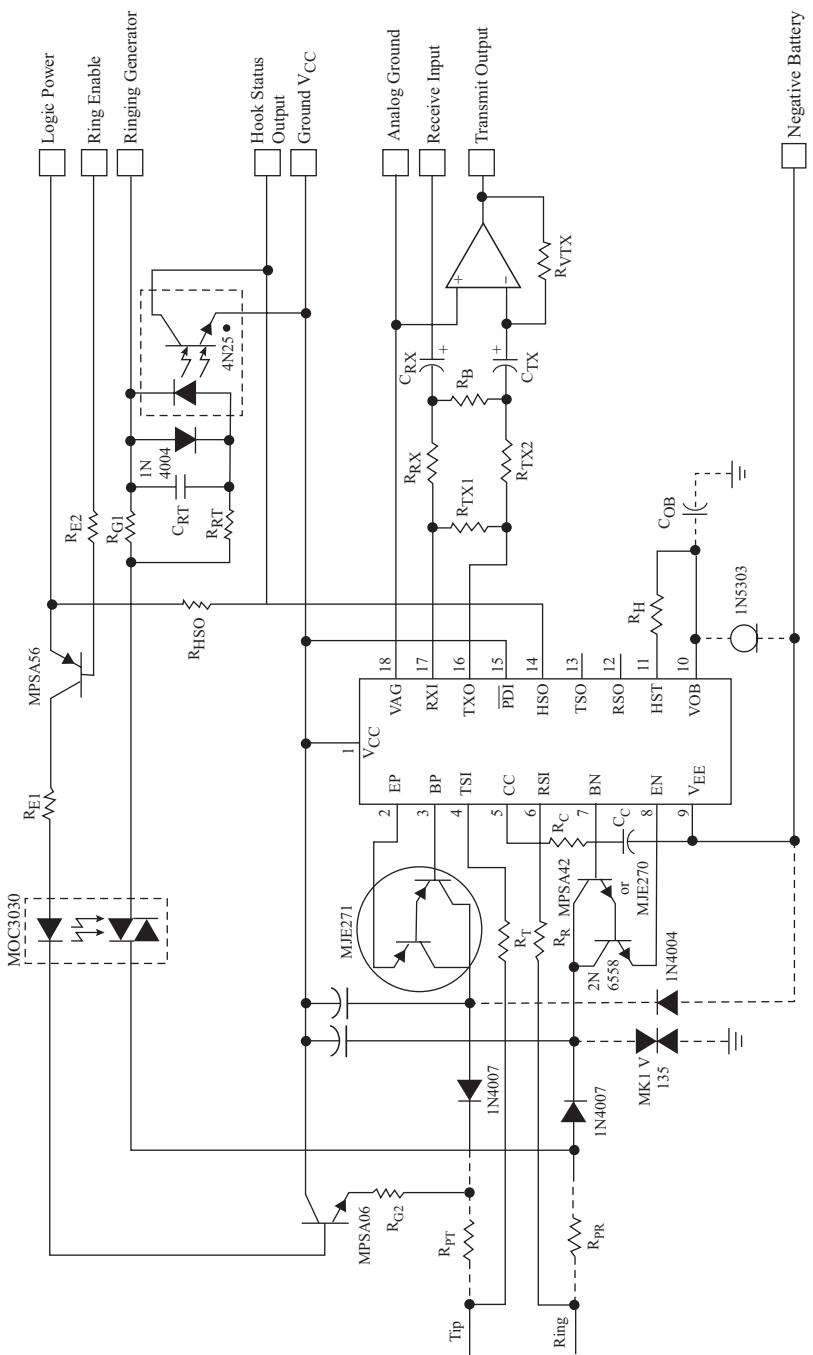


Figure 17.1.3 Single line analog SLIC board using the Motorola MC3419 SLIC chip. (Courtesy Motorola, Inc.)

A voice signal modulated on the line current is sensed through the current mirrors and presented at the four-wire output terminal. An op-amp circuit amplifies this signal to the desired output level before being applied to the four-wire transmit line.

A detector circuit within the chip monitors the loop current, and if it falls below a threshold value, it produces a logic signal at the HSO output of the chip (that is, the hook switch output signal).

A logic signal from the switching machine energizes a light-coupled relay (MOC3030) to connect the ac ringing bus voltage directly to the ring side of the loop (returned through ground or the tip side of the loop). A second light-coupled relay (4N25) senses the presence of ringing current and shorts down the hook switch output signal if it rises above a threshold value. When the telephone hook switch is engaged, the terminal impedance of the telephone is lowered considerably, allowing ringing current to rise above the threshold. This releases the hook switch output to signal the switching machine of a response. When a dial pulse or a hangup occurs, the current mirrors sense the drop in loop current and produce a logic output on the HSO terminal to signal the switching machine. Dial pulses are sensed by another circuit connected to the HSO line.

Ovvoltage protection is afforded by an MK1V135 varistor transient suppressor connected between the ring line and ground and a diode clipper from the tip line to the negative 48-V battery terminal.

Many exchange switching machines are being replaced by fully digital machines, which means that the digital PCM encoding point is being moved closer to the subscriber's terminal. In this case the SLIC for each subscriber loop circuit is expanded to include the PCM codec and also the time slot assignment circuit (TSAC) for interleaved time multiplexing. Figure 17.1.4 shows such an SLIC board. It uses an MC3419 connected in much the same manner as in Fig. 17.1.3, with the exception that the ring bus is connected by an electromechanical relay instead of light-operated relays.

The four-wire connections from the MC3419 are connected directly to the analog receive and transmit connections of the MC14403 PCM codec chip. The digital data receive and transmit lines are connected to the backplane for distribution to the switching machine. The codec produces 8-bit MU-255 companded PCM code for transmission, and the necessary filters are included on the chip as well. (Refer to Chapter 11 for details of the PCM process.)

The MC14418 TSAC chip performs time slot assignment for one unit of a 128-unit channel bank of time slots. The RxE signal gates both the digital data input and output lines of the MC14403 codec during the assigned time slot, allowing simultaneous reception and transmission during that time slot. A local 5-bit address bus on the MC14418 chip is preset to its unique identity code either through a set of switches on the card or through the card wiring. Five logic lines provide connection to the switching machine for control and the transmission and reception of address information. When the received address matches the local identity address, the gating signal is sent to the codec chip. Time slot selection is controlled from the switching machine. It should be noted that the MC14418 TSAC can provide two different time slots, one for transmitting and one for receiving, which makes time domain switching more flexible.

Subscriber Loop Lines

Loop lengths are limited primarily by the dc resistance of the loop pair. If the dc loop resistance is greater than about $1200\ \Omega$, the dc signaling becomes unreliable. Depending on the type of cable used for the loop pair, this value of resistance limits distance to about 5 km using No. 26 AWG copper to perhaps 45 km using No. 16 AWG copper. Open wire lines of heavy gauge copper clad steel conductors were once used for long rural loops.

Further limitation is placed on loop length by the attenuation factor of the cable for voice signals. Loops with losses in excess of 10 dB total are not acceptable, and repeater amplifiers must be included in the loop

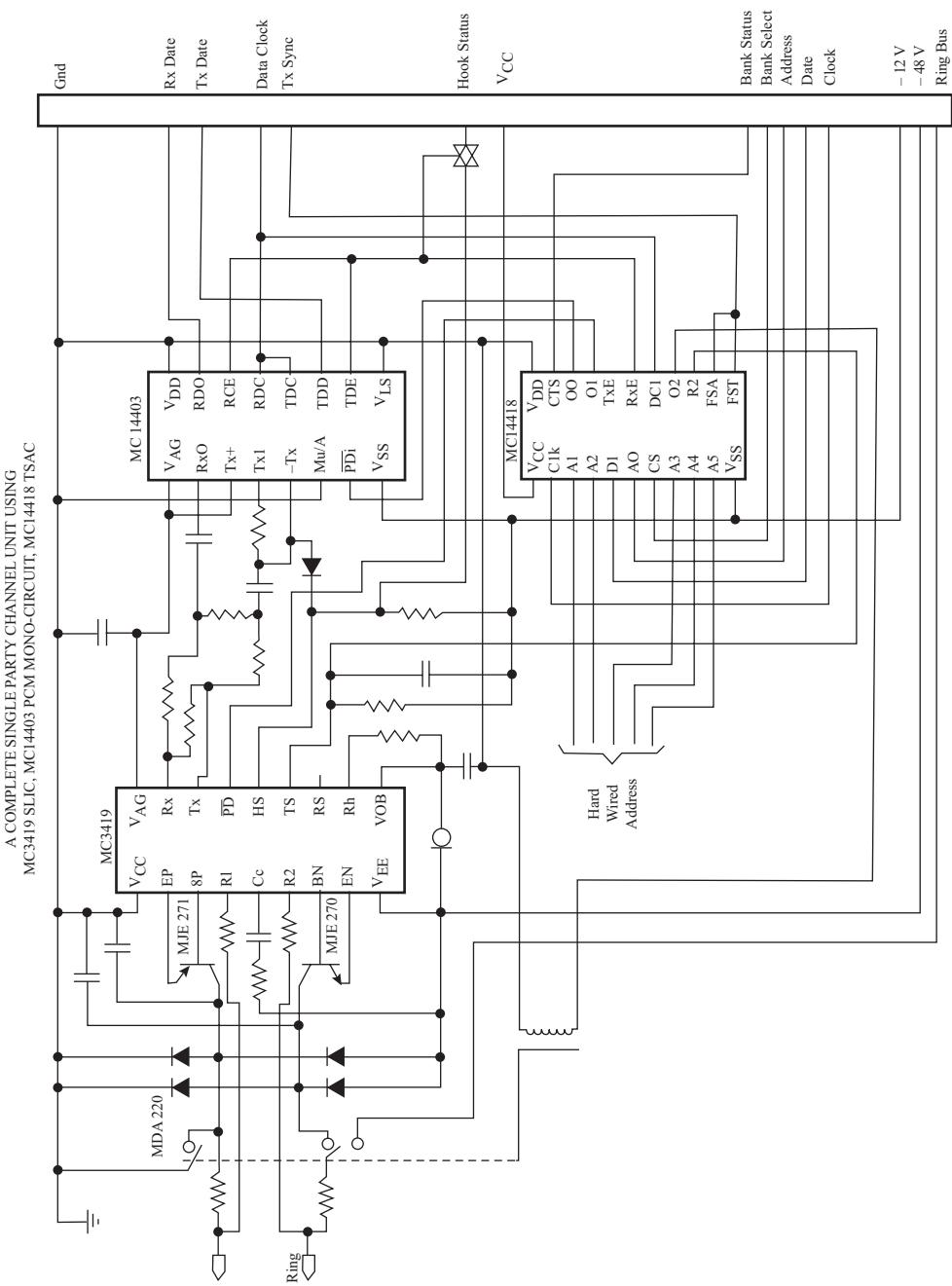


Figure 17.14 Complete single line SLIC circuit providing full BORST functions. (Courtesy Motorola, Inc.)

TABLE 17.1.2 Exchange Loop Cable Characteristics

Cable AWG Gauge No.	Round Trip Loop Resistance (ohms per loop mile)	Attenuation (dB/mile, unloaded)	Factor (M88 ^a loaded)	1200- Ω		
	Loop Length (miles)	1200- Ω Loop (dB, unloaded)	Loss (loaded)			
26	431	2.67	—	2.8	7.5	—
24	271	2.15	1.31	4.5	9.7	5.9
22	170	1.80	0.92	7.1	12.8	6.5
19	85	1.12	0.44	14.1	15.8	6.2
16	42.4	0.76	0.24	28	21.3	6.7

Courtesy Howard W. Sams & Co.

^aM88 loading: 88-mH loading coils placed at 9000-ft intervals (1 ft = 0.3048 m and 1 mile = 1.609 km).

to make up for these losses. Table 17.1.2 shows typical exchange cable characteristics. Uncompensated or unloaded cable exhibits an increasing attenuation factor with increasing frequency. It is necessary to compensate for this in longer loops. The correction is known as *loading* and consists of placing series inductances at intervals along the cable. These inductances act with the cable capacitance to create a distributed low-pass filter with a cutoff frequency of about 3.5 kHz and a flat attenuation function over the frequency range from 300 to 3000 Hz. The attenuation factor tends to be lower than that for unloaded cable, as indicated in the table. The total loop attenuation for the unloaded cable cut for 1200 Ω dc resistance rises with the physical length (that is, with lower gauge number or larger wire diameter), while that for the loaded cable remains approximately constant at about 6 dB.

Transmission Bridges

The term *transmission bridge* applies specifically to the circuit used at the central office end of the loop to separate the voice path from the signaling path and battery and pass it on to another loop. In electronic offices it is referred to as a subscriber line interface circuit, or SLIC. This circuit must isolate the voice signals from the dc signaling circuit and from the central battery supply and pass these signals into the switching network for cross connection to another loop.

Figure 17.1.5 shows a variation of the Stoneman bridge in which transformer coupling is used to connect the loop into the switched network. Transformer coupling minimizes the insertion loss of the coupling and provides balanced operation. Moreover, the transformer allows impedance matching between the loop and the switching system, which may be operating at a different characteristic impedance than that of the loop circuit. Typically, the characteristic impedance of loop circuits is about 900 Ω , while that of interoffice trunk circuits is 600 Ω . Ballast resistors are included in the battery leads to compensate for various loop lengths (loop resistances). The signaling relays operate when dc loop current flows.

Other forms of bridging circuits have been used, but all these are being replaced by electronic subscriber line interface circuits like that one discussed.

Two- to Four-wire Conversion

Originally, many switching machines only operated on two-wire circuits, with four-wire circuits only being provided at the long-distance level. The new digital switching systems have moved the point of conversion to four-wire to the subscriber's SLIC level.

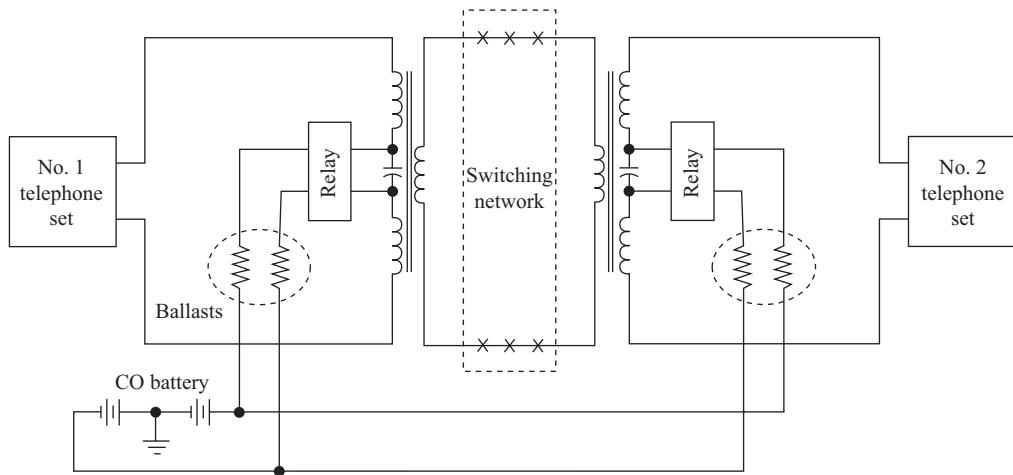


Figure 17.1.5 Transformer-coupled transmission bridge.

In two-wire transmission, the transmitted signal and the received signal must share the same pair of wires or channel. In four-wire transmission, a separate pair of wires is provided for each direction. To prevent an oscillation loop from being set up around the four-wire circuit, the converter circuit at each end must prevent the received signal from one side of the circuit from feeding across and back on the transmitting side. In the past, balanced transformer circuits called *four-wire terminating sets* or *hybrids*, as shown in Fig. 17.1.6(a), were used to provide two- to four-wire connection, as required by loop repeater stations.

This circuit uses two separate transformers, each with four identical windings. The two-wire line is connected to the two A windings, which are connected series aiding. The passive termination is connected to the two B windings, which are connected series opposing. The transmitting pair of the four-wire line is connected to the C, D windings of one transformer, which are connected series aiding, and the receiving pair is connected to the C, D windings of the other transformer, also connected series aiding. The passive termination is built to have an impedance identical to the characteristic impedance of the two-wire line.

Figure 17.1.6(b) shows the equivalent circuit for the currents applied from the two-wire line. Currents in windings A induce identical emf's in windings B in both transformers. These windings are connected series opposing, so that the emf's cancel and no current flows in the passive termination. Currents are induced in the C, D windings of both transformers, however. The currents in the output port are transmitted on one pair of a four-wire line, but those applied to the input port are blocked by the unilateral receiver amplifier connected to the other pair of the four-wire line. Only half of the input power is sent out on the four-wire pair, resulting in a 3-dB insertion loss to transmission.

When the receiver amplifier drives the input port as in Fig. 17.1.6(c), equal currents are induced in the A, B windings. These currents flow from the A windings down the two-wire line and from the B windings to the passive termination, returning through the A, B windings on the second transformer. These A, B windings are connected series opposing for this pair of currents, resulting in cancellation and no current in the output port. In practice, a small amount of current will result in the output port because of imbalance (typically less than -50 dB of the current applied at the input port). This is called the *feedthrough loss*. Again the input power is split two ways, causing a 3-dB insertion loss.

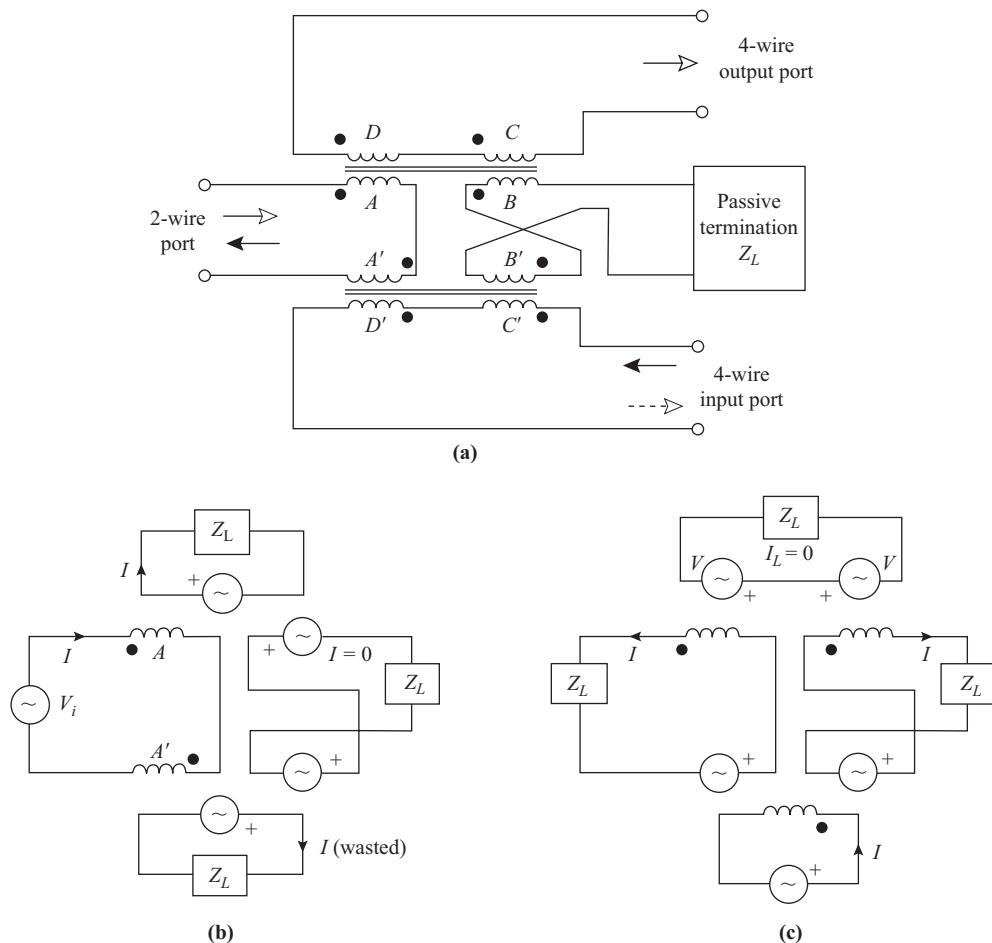


Figure 17.1.6 (a) Four-wire terminating set. (b) Currents and emf's due to a two-wire signal. (c) Currents and emf's due to a received four-wire signal.

Two-wire Repeaters

Repeaters are amplifiers that are inserted in transmission lines at intervals to amplify the signal and compensate for transmission loss on the line. On two-wire transmission lines, transmission occurs in both directions, so any amplifiers used must be bilateral, amplifying equally well in both directions. *Negative impedance amplifiers* employing a principle similar to the negative resistance in oscillators are used. Negative impedance amplifiers are two-terminal bilateral devices. Before these became available, only unilateral amplifiers could be used, and it was necessary to split the two-way transmission into two separate paths (four-wire) before the amplifiers could be inserted. Hybrid transformers like those of the previous section (or their electronic equivalents) are required to divide the two-wire path into a four-wire path and return it to two-wire.

Figure 17.1.7 shows a typical arrangement of a two-wire repeater. Two unilateral amplifiers provide the necessary gain to overcome the insertion losses of the hybrid sets and the transmission losses over a section

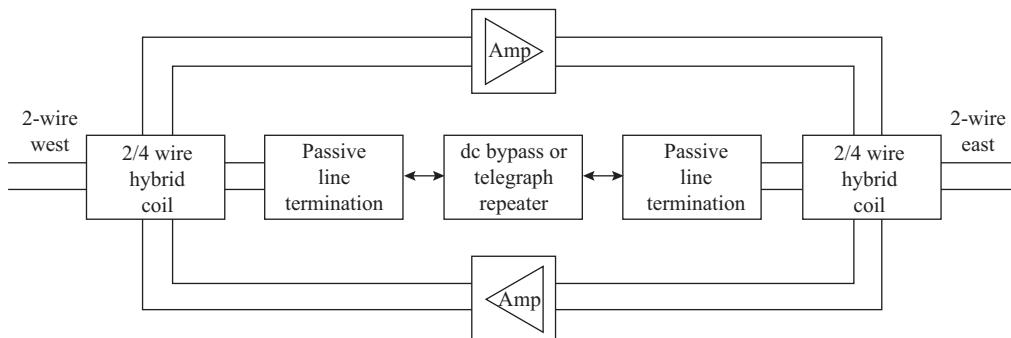


Figure 17.1.7 Two-wire repeater station using unilateral amplifiers.

of transmission line. The two terminating sets split the two-wire circuit into two oneway circuits containing the amplifiers, connected back to back, without introducing a local feedback path causing oscillations. A bypass circuit provides a path for ringing and supervisory dc signals around the repeater.

The conditions for oscillation in a circuit are that a continuous feedback loop must be formed within the circuit with a loop gain greater than unity (0 dB), with a loop phase shift that is a multiple of 360° at the frequency of oscillation (Barkhausen criteria). In the repeater circuit shown, feedback loops can be established in several ways. The longest is for a signal from one end being reflected from the far end and returning to the near end. The shortest is around the loop containing the two amplifiers and the two hybrids. Proper matching will minimize intermediate reflection points.

The amplifiers are required to raise the one-way gain to near 0 dB for good transmission, while not pushing the long loop gain over 0 dB, requiring a terminal reflection loss of at least several decibels. The hybrids must be balanced to provide as good cancellation of return signals into the four-wire circuit as possible (that is, have a high return loss), so that a high gain can be realized from the amplifiers without pushing the short loop gain over 0 dB.

The maximum allowed amplifier gain is limited by the degree of matching on the lines necessary to reduce any feedthrough that can cause oscillations. For the repeater to be stable, it is necessary for the net repeater loop gain on the short loop including the two amplifiers and hybrid feedthrough losses to be less than unity. A loop gain margin included as a protection against oscillation (typically up to 10 dB) provides an upper limit on the practical gains. The term *return loss* is defined as the ratio of power level sent into a circuit to the power level returned to the sending point by a reflection or a feedback path, expressed in decibels. The feedthrough loss of a hybrid is one example of a return loss. The reflection loss of an improperly terminated line is another, in which case signals sent out on the line travel to the far end, are partially reflected, and return to the sending end.

EXAMPLE 17.1.1

- If the repeater of Fig. 17.1.7 uses hybrid transformers with a feedthrough return loss of 50 dB, and the lines are matched so that reflections are negligible, calculate the maximum gain in each direction if the gains are to be equal and if a gain margin against oscillation of 12 dB is provided.
- If the line reflection return loss measured on the west line is 20 dB and that on the east line is 40 dB, find the maximum gain in each direction to give a gain margin of 12 dB.
- If the terminations at each end have a reflection return loss of 15 dB, find the amplifier gain setting required to give a net transmission loss of -6 dB (that is, a margin of 6 dB).

SOLUTION (a) The net loss NFL around the local feedthrough loop is twice the feedthrough loss FTL of the hybrid transformers used.

$$\text{NFL} = 2\text{FTL} = 2 \cdot 50 = 100 \text{ dB}$$

The maximum net feedthrough loop gain NFLG allowing the margin M for stability is

$$\text{NFLG} = \text{NFL} - M = 100 - 12 = 88 \text{ dB}$$

The maximum amplifier gain is one-half of the NFLG, or **44 dB**.

(b) The net loop loss for overall transmission NL is four times the hybrid transformer insertion loss IL (twice going and twice returning), plus the two line reflection losses RLW and RLE.

$$\text{NL} = 4\text{IL} + \text{RLW} + \text{RLE} = (4 \cdot 3) + 20 + 40 = 72 \text{ dB}$$

The allowed net loop gain NLG is the net loop loss NL less the margin M.

$$\text{NLG} = \text{NL} - M = 72 - 12 = 60 \text{ dB}$$

giving a maximum individual amplifier gain of **30 dB**. This is less than the 44 dB maximum on the short loop.

(c) The one-way loss on the west line OLW is the line reflection return loss RLW less the terminal return loss LT,

$$\text{OLW} = \frac{\text{RLW} - \text{LT}}{2} = \frac{20 - 15}{2} = 2.5 \text{ dB}$$

The one-way loss on the west line is

$$\text{OLE} = \frac{\text{RLE} - \text{LT}}{2} = \frac{40 - 15}{2} = 12.5 \text{ dB}$$

The amplifier gain is the sum of the one-way margin OM plus the one-way losses plus the two hybrid insertion losses, or

$$A = OM + OLW + OLE + 2IL = 6 + 2.5 + 12.5 + (2 \times 3) = 27 \text{ dB}$$

Negative resistance amplifiers may also be used for loop gain. The use of such repeaters is dependent on how well the lines can be matched to prevent reflections from occurring, because the amplifier is connected in parallel with the two-wire line and there is no local feedback loop. The only way oscillations can occur is if the total inserted gain of the negative resistance amplifier (that is, its gain) is greater than the sum of the two reflection losses on the lines. The better the lines are matched, the higher these reflection losses are and the more gain allowed. Transformer coupling can be used to place the repeater negative resistance in series with the line. The Bell System E6 exchange loop repeater is a good example of this. It is a transistor amplifier that operates directly from the dc talk battery current in the line and allows the reduction of losses on long loops. Separate signaling repeaters may also be required to regenerate the dc loop signals on very long loops.

Four-wire Transmission

Two-wire circuits are difficult to balance accurately, and the maximum obtainable gain is limited. For this reason, long-distance and interoffice trunk circuits are nearly always done on a four-wire basis, with two wires providing transmission in one direction and the other two the reverse direction. When carrier circuits

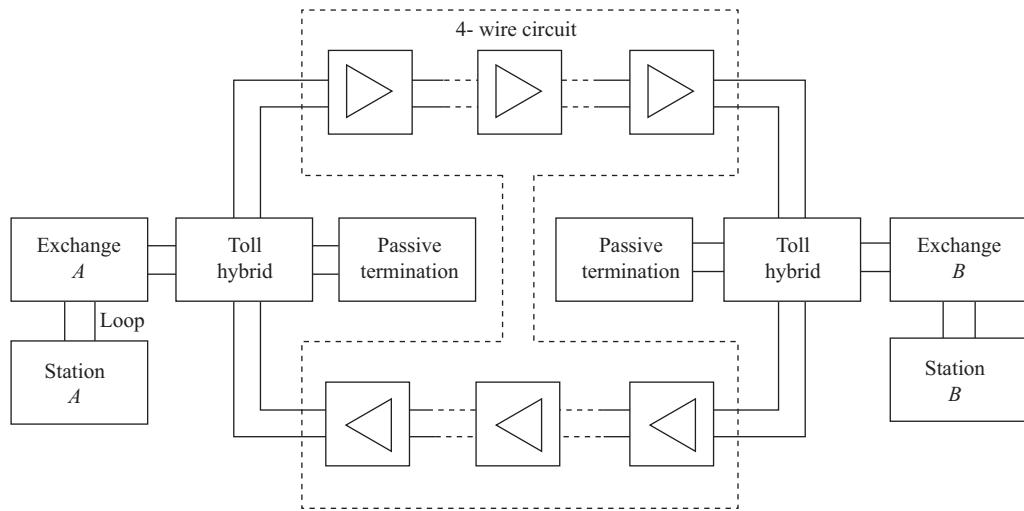


Figure 17.1.8 Four-wire toll trunk circuit arrangement for long distance.

are used, four-wire (two-path) transmission must be used because the carrier amplifiers are unilateral and only allow one-way transmission. Figure 17.1.8 shows a typical arrangement of equipment along the path of a four-wire long-distance circuit. The customer loop at each end is switched on a two-wire basis into a toll (long-distance) switching center. A four-wire terminating set at the toll center converts the circuit to the two-path mode. The long-distance facilities between this toll center and the far end toll center are all four-wire and may include cable circuits, repeaters, carrier circuits, and radio circuits (including satellites). At the far toll center the transmission is converted back to two-wire before connecting to the far customer's loop.

17.2 Public Telephone Network

The System

The basic purpose of the public telephone system is to provide two-way voice communication between any pair of subscribers within the system. A wide variety of subsidiary services has been added in recent years, such as communication with subscribers in other systems, telegraph, teletypewriter and facsimile service, computer data communications, conference calls, and private network service. The basic components of such a system are a network of cable and radio channels linking many centers, switching machines at each center to allow interconnection, and local subscriber stations.

The central component of a system is the exchange switching machine or *office*. This is a switching system that allows direct interconnection among up to 10,000 subscriber stations located mostly within about 10 km of the office, each of which is assigned a separate four-digit number. The office also provides facilities for connection through exchange interoffice trunk circuits to several other nearby exchanges or into toll connecting trunks to a nearby toll or longdistance switching center. Early telephone switching machines were manually operated by operators who physically made the connections necessary to complete a call by means of plug leads on a switchboard. Manual switchboards are still used to serve some private company systems with a few hundred lines, but nearly all public system switching is now done automatically.

In the 1950s the Bell System introduced the concept of continent-wide automatic toll switching, or *direct distance dialing* (DDD). Before this time, all long-distance connections were done by means of manual switching by long-distance operators. Each exchange could automatically switch connections between a subscriber and the nearest toll office, but then the toll operator had to manually complete the call to the far end. This often involved switching by several operators at intermediate centers along the route of the call. With DDD, the connections at the toll center between the toll connecting trunks and the intertoll long-distance trunk circuits are made automatically, and billing information is automatically recorded by a computer. Backup manual facilities are maintained on a limited basis in case of difficulty or for the establishment of special service calls, such as person-to-person or conference calls.

Antimonopoly legislation in the United States in the early 1980s resulted in the partition of the original telephone network into many independent local telephone companies providing regional exchange service, with long-distance interconnection services to be provided by several independent carrier companies operating essentially parallel networks.

For the purpose of making long-distance calling more efficient with respect to the use of toll trunk circuit facilities, a hierarchy of toll offices was established. Fig. 17.2.1 illustrates the interconnections possible within a DDD network. A specific location was chosen within each major population center for the establishment of the local toll connecting offices. These toll connecting offices might serve anywhere from 10 to 100 exchange offices, depending on the local population density. Toll connecting trunk circuits link each exchange office to the toll connecting office. Within each exchange, calls are made on a flat-rate basis so that call connection recording is not required (*A* and *B* in Fig. 17.2.1). Exchange interoffice trunk circuits allow flat-rate calls to be made to other nearby exchange locations within a designated flat-rate calling area (*C* and *D* in the figure). Again, no record is kept of the calls.

Any call that requires a specific charge must be routed through the nearest toll connecting office, where a record of the call is made. At the toll connecting level, four levels of long-distance calls are possible. First, the call may be to a second exchange that is served by the same toll center, in which case a direct connection between two toll connecting trunks is made. Second, the call may be made to an exchange served by a nearby toll center, in which case the call may be routed over a direct toll trunk between the two centers. Third, the call may be between centers in two different regions, separated by a considerable distance. Each region is served by a regional toll switching center, and any calls between regions are routed through these centers. The facilities linking these regional centers are mostly heavy-route microwave relay systems or satellite relay systems or fiber-optic systems. Fourth, in the new regime, the subscriber may elect to connect to "the other carrier" and use a completely different network to complete the call. This requires each independent exchange company to provide interconnections to each of the independent carrier companies and their parallel networks.

A limited number of trunk circuits is provided between any two nearby switching machines, be they exchange or toll. The number provided is based on long-range forecasts of area population growth and future demand for service, done on a statistical basis to provide a certain maximum probability of being able to complete a call during maximum busy hours. Calls between each pair of exchanges within the system are assigned first-, second-, and third-choice routes as available, and the switching machine automatically attempts to select these in sequence before indicating a "trunk busy" condition. An example of this is provided by *C* to *D* and *E* to *F* in Fig. 17.2.1. *C* was able to establish a direct trunk connection to *D* through offices A_x and B_x , a first-choice selection. When *E* tried to call *F* between the same two offices, all the A_x to B_x direct trunks were busy, and the alternate route through C_x was chosen. In a similar manner, toll calls within a region may be made on first choice over direct trunks, and on second choice through one or more intermediate toll offices, or third choice through the regional toll office.

The automatic switching machines for each exchange are designated by a three-digit address code. Any local call within an area must be established by means of seven address digits, three prefix digits for the office address, and four for the particular subscriber's station being called. The digit 1 is reserved as an access

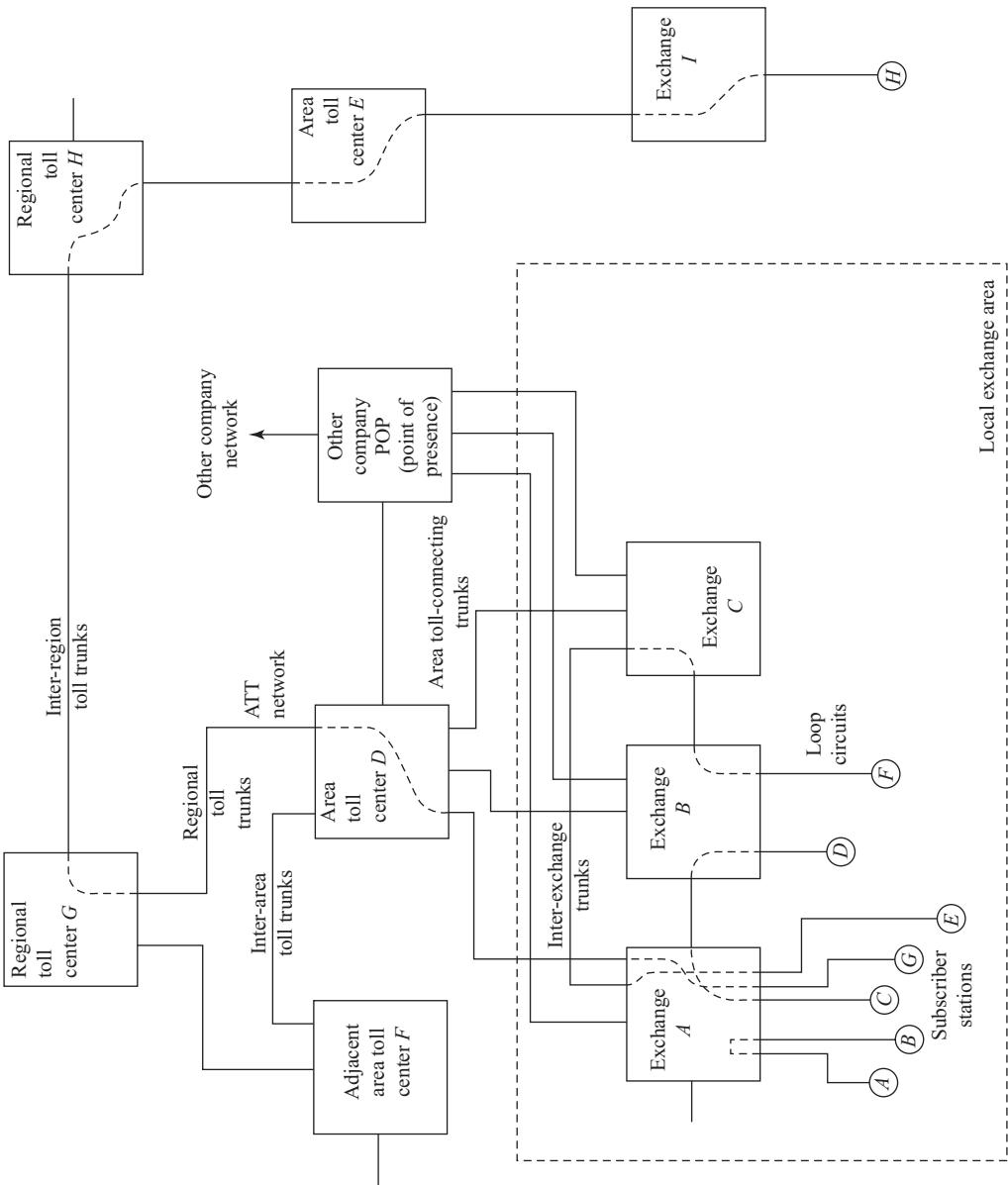
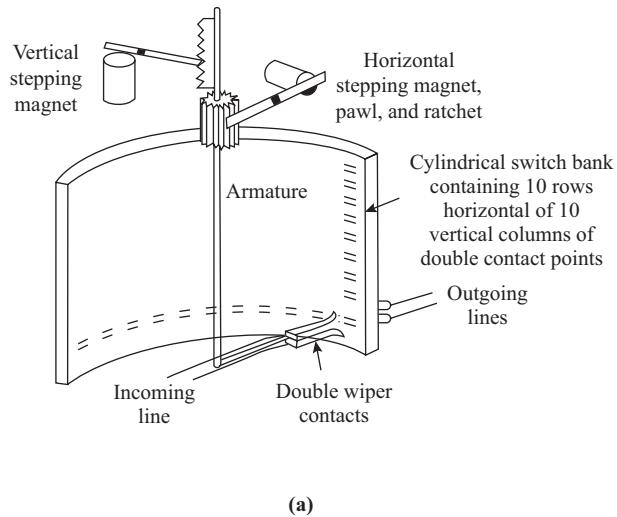


Figure 17.2.1 Direct distance dialing telephone network.

code within each exchange to make connection to a toll connecting trunk circuit. A second three-digit code immediately after the 1 code addresses the toll area in which the destination subscriber is located. This is followed by the seven-digit number, which completes the call in the destination area. Calls outside the country require special routing codes, prefixed with a two- or three-digit country code. Access to a private carrier requires dialing a nine-digit access code plus the additional routing information.

Step-by-step Switching

The heart of the automatic telephone system is the switching machine. The earliest of these systems used step-by-step switching by means of the *Strowger stepping switch*, whose basic mechanism is illustrated in Fig. 17.2.2(a). Each digit dial pulsed into the machine steps one of the switches up by the corresponding number in turn to complete the call.



(a)

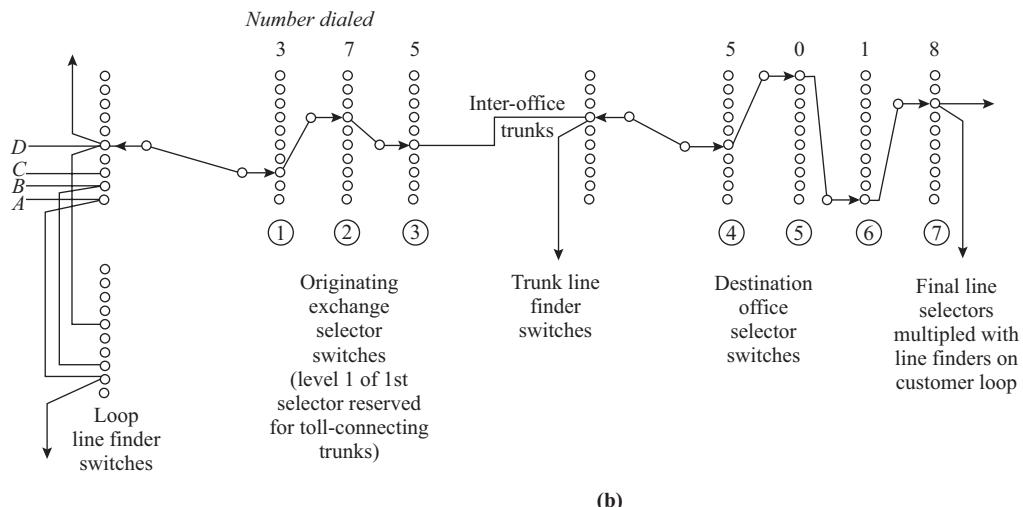


Figure 17.2.2 (a) Stepping switch. (b) Connections through a step-by-step switching machine.

The step-by-step office contains many hundreds of mechanical switches, each with moving parts and wiping contacts. As a result, maintenance on this type of system is costly and difficult. The switching of currents in magnetic circuits also generates a high level of impulse-type noise within the system, which, while tolerable in voice communications, is completely unacceptable in data communications. As a result, step-by-step offices are being retired as rapidly as possible and replaced by modern electronic switching machines, usually of the digital type.

Crossbar Switching

The *crossbar switching system* derives its name from the switching device used, the crossbar switch. Figure 17.2.3(a) shows the mechanical layout of a 200-point crossbar switch unit. This unit is laid out in 10 horizontal rows and 20 vertical columns, with a set of four contacts located at each of the 200 cross-points. Each cross-point contact set can be actuated by energizing one horizontal magnet and one vertical magnet.

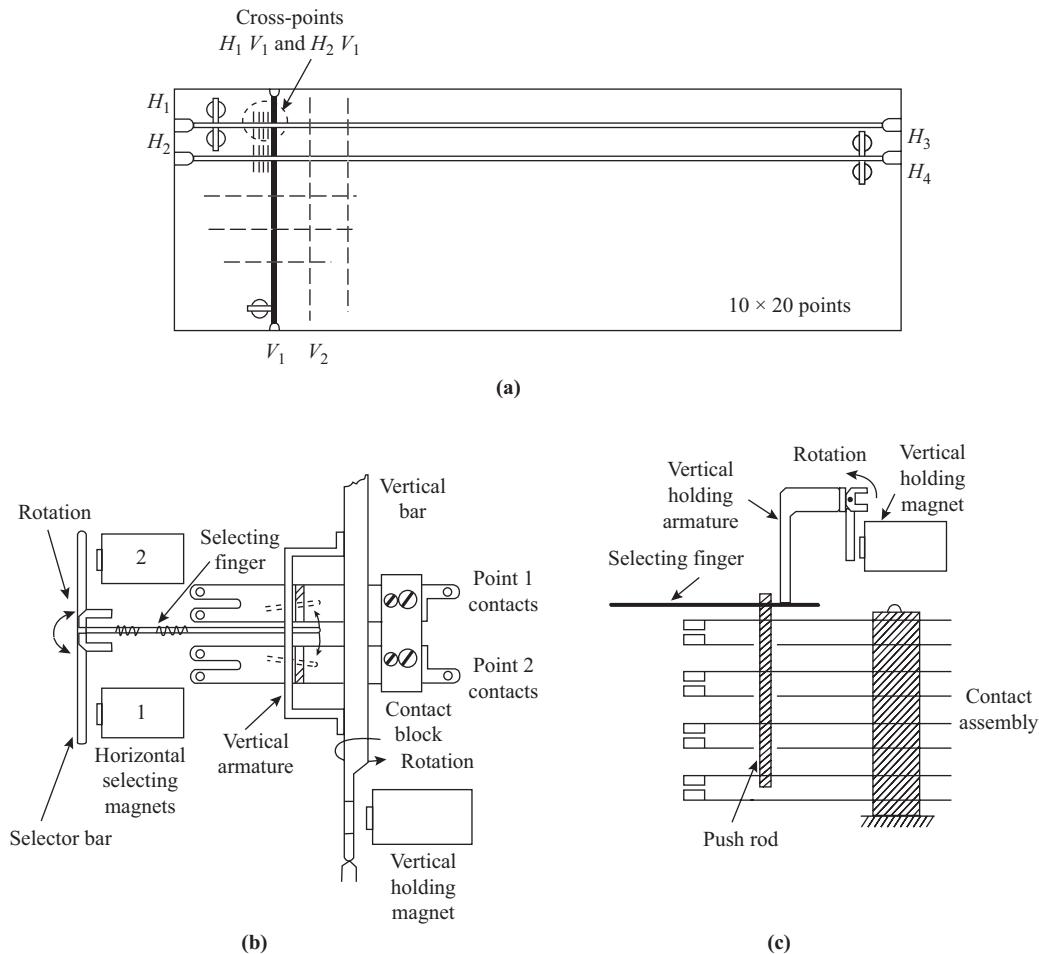


Figure 17.2.3 (a) Structure of a crossbar switch. (b) Two contact assemblies showing the selection wire action. (c) Contact assembly showing the holding armature action.

magnet, first causing a horizontal bar to rotate and then a vertical holding bar to rotate. There are only five horizontal bars, each of which can rotate in two directions, so each bar can control cross-point switches in one of two rows at any one time. There are 20 vertical holding bars that rotate in only one direction, so only 20 simultaneous independent paths can be established through the switch.

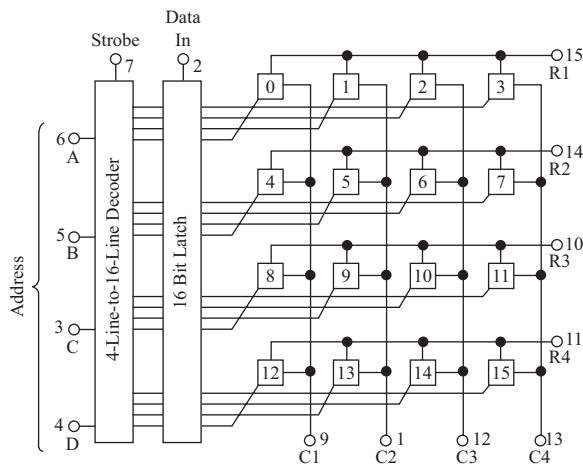
Figure 17.2.3(b) and (c) shows one of the cross-points in more detail. Each horizontal bar carries 20 selecting fingers or wires, which in the idle state lie in the gap between two sets of cross-point contacts. When one of the horizontal magnets is energized, the horizontal selecting bar rotates, causing all 20 selecting fingers to move over on top of the contact sets on one side. When one of the vertical holding magnets is energized, its vertical holding bar pushes down on the selecting spring at the desired cross-point and closes the four contact sets under it through the push rod. When the horizontal selecting bar is released, all springs except the held spring move back into the idle position. The operated spring remains trapped under the vertical holding armature to hold the contacts down until the vertical holding magnet is released. Each of the 200-point crossbar switch units is capable of making and holding a total of 20 independent contact closures at once. Switch units are used in pairs, and several pairs are multiplied within a frame to provide the required fanout of the connections in the switching system.

Since about 1975, Silicon MOSFETs in very large scale integrated circuits (VLSI) have made possible the realization of cross-point switching systems without moving parts. Silicon-controlled rectifiers (SCR) have been used for small, slow-acting switching systems, but they are not efficient in their use of power because of their required holding current. Bipolar transistors can be used as analog switches, but their asymmetry results in complex circuit configurations. However, complementary MOSFETs make good analog switches. Two transistors, one *n*-channel and one *p*-channel, are used back to back to eliminate any asymmetry in characteristics. When they are both switched on, a switch resistance of about $200\ \Omega$ is presented to low-level ac signals. When switched off, they provide a blocking resistance of hundreds of megohms and a very small bridging capacitance.

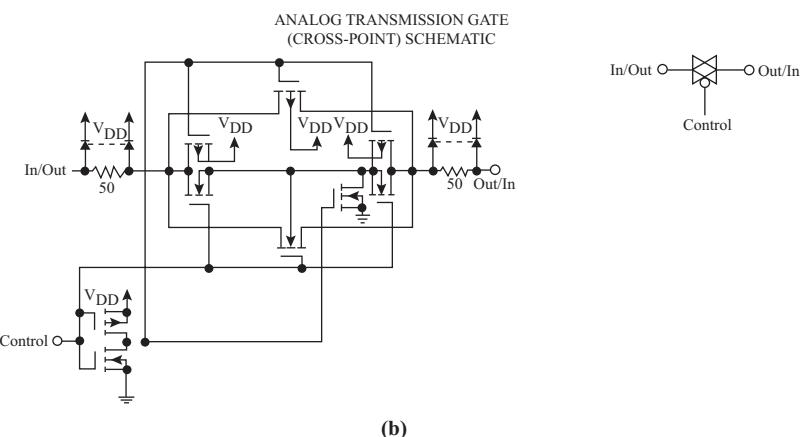
Figure 17.2.4(a) shows the functional block diagram of a Motorola MC142100 cross-point switch realized using CMOS technology. It provides 16 analog transmission gates of the type shown in the schematic of Fig. 17.2.4(b), connecting four input lines to four output lines. The switches are fully bilateral, so signals will transmit in both directions, eliminating the need for a second switch for the return path. When a control address for a cross-point is latched into the on-chip 16-bit memory, the output from the latch applies a high level to the gates of the three series *n*-channel FETs to switch them on. The control signal is also inverted to place a low level on the gates of the three series *p*-channel FETs to switch them on. This inverted signal also turns off the shunt *n*-channel FET to isolate the cross-point from ground. The six parallel FETs provide a bilateral low-resistance path between the input and output lines corresponding to the cross-point. Clamping diodes on the input and output lines act to prevent damage from electrostatic discharge (ESD) on the lines. Each cross-point can be independently switched, so it is possible to cross-connect any or all inputs and outputs, making conference connections possible.

The switch described, only accommodates a 4×4 independent path switching. While it is possible to integrate many more cross-points onto a chip, the number is limited to about 16×16 as a maximum because of the number of input and output pins needed on the chip. A board containing 256 of these 16×16 chips would form a completely nonblocking cross-point matrix for 256×256 circuits. Since the CMOS circuits will operate satisfactorily at speeds in excess of 10 MHz, they can be used in a time-shared space switch to good effect.

Cross-point switches are used in matrixes to form space mode switch units. If the matrix connects N input lines to N output lines, the matrix is square and is completely nonblocking. That is, N independent paths interconnecting the N inputs to the N outputs in any order can be established. In fact, with the square matrix, there are two cross-points for every path, resulting in many more switches than are necessary. It is only necessary to provide for each input line to cross with each output line at only one cross-point.



(a)



(b)

Figure 17.2.4 (a) CMOS VLSI chip cross-point switch. (b) Schematic of one cross-point analog switch. (Courtesy of Motorola, Inc.)

The matrix need not be square. If it has M inputs and N outputs, where N is less than M , then only N connections can be made at one time. The number of outputs provided is based on the statistical maximum demand for connection paths, or the *traffic* that can occur on the M inputs. More elaborate switch units incorporating two or more stages of cross-point switching are set up with enough cross-points to guarantee N simultaneous paths between M inputs and M outputs, where N represents the traffic level expected. If the demand approaches N , then the probability of encountering a blocking condition rises significantly. The number of switches required to make a staged switch of this type is much smaller than the number required for a completely nonlocking $M \times M$ matrix, thus providing a significant reduction of equipment and cost.

Figure 17.2.5 shows the major pieces of equipment in a typical crossbar exchange office and the path of a call established through it. The line link, district link, office link, line connecting link, and incoming trunk link are crossbar switch banks used to establish signal paths through the machine. The sender link is a crossbar switch used to connect the district junctor to a sender during call establishment. The district junctor, sender, marker, and line link controller are each special-purpose computers made of electromechanical relay circuits that control the action of the various switches.

An incoming subscriber's loop is connected to several vertical points within one or more crossbar switches, so it has access to several horizontals. When the calling station loop is energized, a relay network searches for and engages an idle line link controller. This in turn controls the action of the line link crossbar switch. It searches to find a primary cross-point to engage on an incoming line and a secondary cross-point to engage an idle district junctor unit. Once these have been found, the two crossbar switch points are closed, and the holding control is passed to the district junctor unit. At this point the line link controller disconnects. The district junctor unit searches for an idle sender unit and operates the two switches in the sender link to connect the junctor to the sender. The sender in turn connects itself to an idle marker unit. Now the district junctor provides a dial tone, indicating that it is ready to receive dial pulses.

The first three digits of the number being dialed are passed directly to the marker unit, which decodes and stores them as the address of the office in which the called station is located. It then searches for and

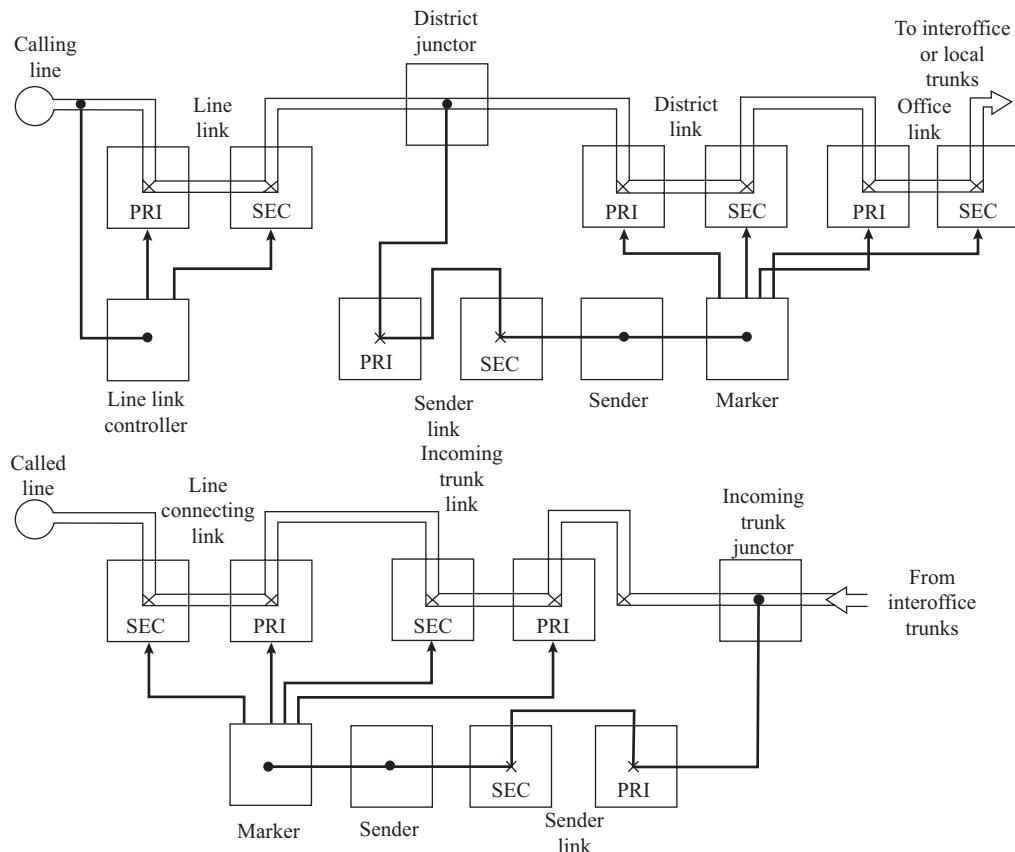


Figure 17.2.5 Exchange crossbar switching system showing the route of a call.

makes cross-point connections in the district link and office link frames necessary to connect the district junctor through an idle trunk circuit out to the destination office. At this point the marker and sender have completed their part of the job, and the marker, sender, and sender link are all released, leaving the district junctor to hold the connections in the district and office link switches. At the far end of the trunk circuit, which may terminate in a remote office or in the originating office itself, an incoming trunk junctor circuit similar to the district junctor takes over control. It engages a new sender and marker unit and waits for the final four digits of the number. These are received, decoded, and stored, and the called line is searched for.

A path between the incoming trunk junctor and the called line is searched for through the incoming link switches and the line link switches, and if the line is not already engaged, the connection is made. At this point the marker and sender are dropped, leaving control to the incoming trunk junctor. If the line is already busy the connection is dropped and a busy tone is transmitted back to the originating station. If it is free, the connections are held and ringing current is sent out to the called station. Operating the hook switch in the called loop closes the dc path and cuts off the ringing current so that the conversation can proceed. Control is maintained by the calling party's loop current, and the connection is not dropped until the calling party hangs up. When it does, the two junctor circuits release all the crossbar switch points in the circuit, which now becomes available to establish other connections.

The number of units of each type provided in an office depends on the statistical probability that a call cannot be completed because all the units are busy. The line link must provide at least one vertical on its primary for each line in the office, and each is multiplied over several verticals. The number of other switch units in the district and office links and the incoming links is considerably smaller and depends on the expected maximum traffic. The sender and marker units are important, but since they do their job in a very short period of time, only a few are needed in a given office. Generally, for a 10,000-line office only six markers are necessary. Only two junctor circuits are necessary for each call in process at a given time. One or two hundred of these are sufficient to operate the office.

Digital Switching

Cross-point switching is a space mode of switching in that it makes a physical connection between two channels on different transmission links or cables. The crossbar switching system is a good example of a space mode switch. However, most of the signals within the telephone system now are in the form of time division multiplexed (TDM) 8-bit PCM signals arranged in a byte-interleaved manner as illustrated in Fig. 17.2.6. In this case, 128 individual telephone signals are sampled at 8 kHz, producing 8-bit PCM. The PCM signals are transmitted 1 byte at a time with a 1-byte slot from each of 128 channels to form a byte-interleaved 128-slot TDM signal. The bit rate of the TDM signal is 8.2 Mbits/s.

Time mode switching is accomplished by passing the TDM signal through a device that interchanges the slot assignments within the signal. This requires receiving and temporarily storing the data bytes in a

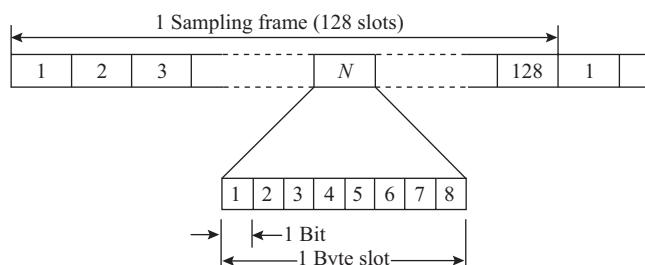


Figure 17.2.6 Time Frame for a 128-channel TDM/PCM signal.

memory and then retransmitting them after a one-frame delay period, but in a different order, as shown in Fig. 17.2.7. For example, suppose a 1-byte data word is received from a line card TSAC in slot 45 during one frame. Slot 45 on the output side is already assigned (busy), so a free slot must be found. The data byte is retransmitted in slot 133 on the outgoing line, which is assigned to a free input on the next switching stage.

The memory must be 8 bits wide (byte oriented) and provide at least 2 bytes for each two-way transmission channel connected to it. It must also be sufficiently fast to accommodate the bit rate.

A time switch does not necessarily have to have the same total number of slots available at both input and output. If it has N input slots and N output slots and N words of memory, then the switch is completely nonblocking. That is, N separate paths can interconnect the N inputs and the N outputs in any order. If, however, it has N inputs and M outputs, with M larger than N , and with N memory words, then only N paths can be connected at any time, and one of the M inputs may be blocked; that is, it will encounter a busy condition. In a slightly different way, if the switch has M inputs and M outputs, but only N memory words (less than M), again only N paths can be established, and the switch can present a blocking condition. The number of output slots and memory words provided in a given unit is based on the statistical probability of the maximum demand for path connections or traffic that the switch must handle.

If a time domain switch is to be used independently, it requires a time division multiplexer on its input and a demultiplexer on its output. However, if a time switch is teamed up with a space switch, then the space switch may also be time shared, resulting in a significantly less complex switching machine in terms of its pieces of hardware. For this reason, most modern digital switching machines use a compound configuration that involves two time switches, with one or more space switches forming an interconnecting matrix between them. The input signal is assigned to a time slot on the input side, which is then passed to an initial time-switching unit. Its output is time-slot-switched to the input of an assigned path through the space switch matrix to the destination slot on the other time switch, which makes the final slot switch into the destination line. The actual path through the space switch on successive frames is not necessarily the same. The time switches search for and select a different free path through the matrix in each time frame, resulting in a time-shared space switch.

The No. 4 ESS electronic digital switching machine was developed by Western Electric for the Bell System as the successor of a series of crossbar switching machines. Although initially designed for switching digitally encoded PCM telephone channels, it can be easily adapted for switching digital data transmission circuits of varying bit rates. The switching architecture for the machine in its largest configuration is shown in Fig. 17.2.8. It is a multistage switch, with a TSSSST configuration (two time switch stages with four intermediate space switch stages).

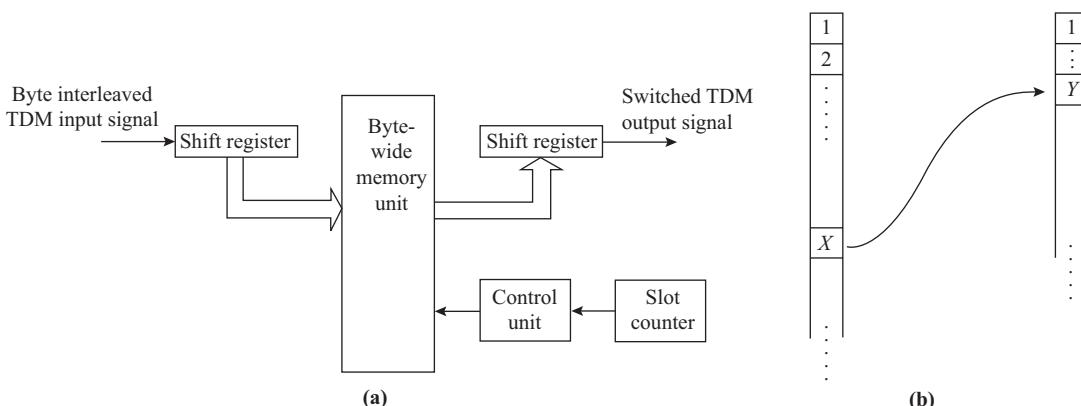


Figure 17.2.7 (a) Time switch unit. (b) Input and output time frames showing slot interchange.

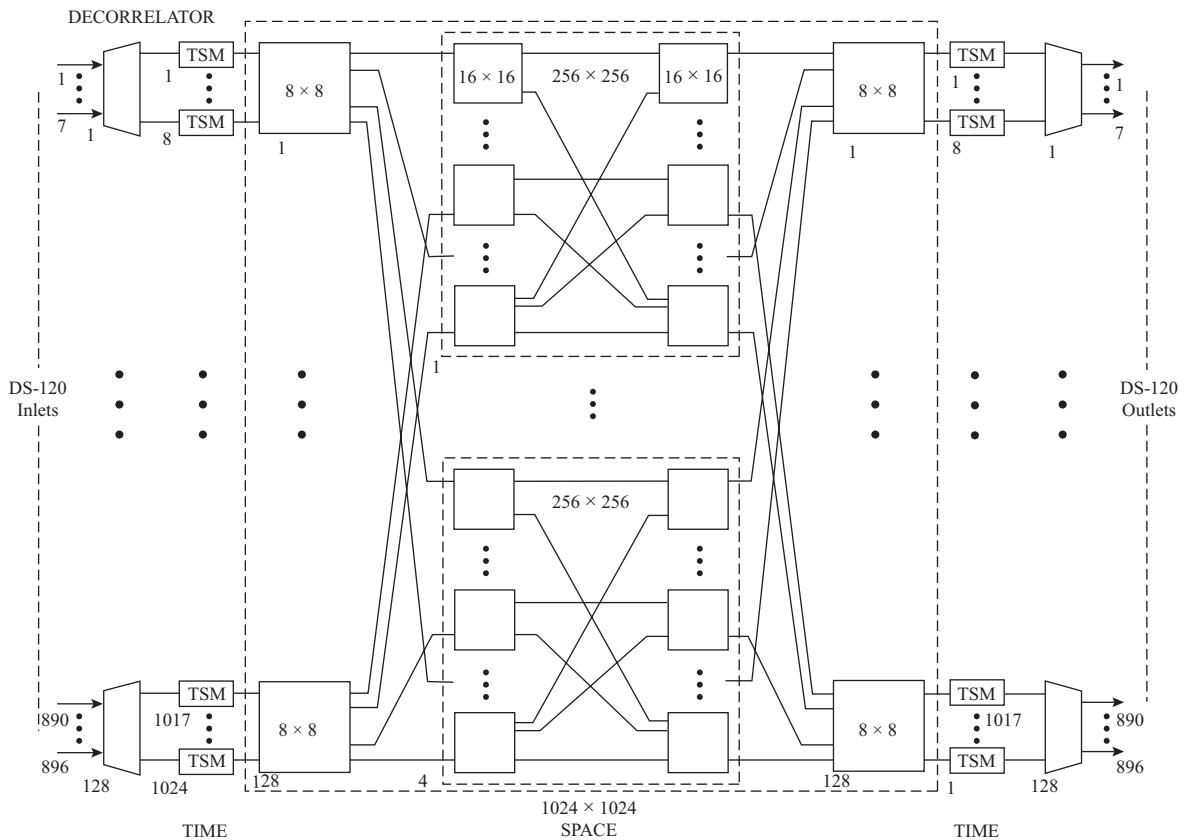


Figure 17.2.8 Matrix architecture for a No. 4 ESS digital switching machine used for toll interconnection. (With permission of John Wiley & Sons, Inc.)

At the exchange level, 24 PCM voice channels are interleaved to form a DS-1 signal with a 1.5 Mbit/s rate. Next, five DS-1 signals are interleaved to form a DS-120 signal operating at 7.7 Mbit/s, providing a fan-in to the switcher. Each physical input line connected to the switcher carries 120 slots per frame providing access for 120 channels. Up to 896 physical input lines are connected into the switcher. They are connected in groups of seven to the inputs of 128 decorrelator units. These decorrelators act in the same way as a time switch, randomly distributing the 120 time slots from each of the seven inputs into 128 slots on each of eight outputs. This has the effect of reducing the chance of a blockage occurring in the space switch stages because of a concentration of traffic on one group of inputs and of effectively minimizing the required size of the space switch. Next each of the 1024 correlator output lines pass through a 128 × 128 slot time switch matrix (TSM) where bit interchanging accomplishes the first time switching step. The 1024 outputs are then connected to the inputs to the first space switch stage.

Four stages of space switches are provided. The first and last each contain 128 8 × 8 cross-point matrixes. The middle two each contain 64 16 × 16 cross-point matrixes, grouped to form four 256 × 256 matrixes. The result is the formation of a 1024 × 1024 space switch matrix. This matrix is time division multiplexed, providing 128 time slots per frame, with the configuration of the space switch being changed for each time slot.

The 1024 output lines from the space switch are connected to the second time switch stage. The 1024 output signals are then passed to 128 slot concentrators, which reverse the process of the decorrelator,

distributing the slots to the appropriate output lines. There are finally 896 output lines, each with 120 slot time division on them (DS-120). This machine will allow interconnection for up to 100,000 telephone channels. Smaller machines can be assembled by omitting one or more groups of switches and correlators from the maximum configuration.

Trunk Circuits

Trunk circuits are transmission circuits interconnecting two different switching centers. These centers may be two adjacent exchanges, or they may be separated by a considerable distance and the connection will involve a chain of several toll (long distance) circuits over different facilities (radio, cable, and the like).

Exchange interoffice circuits are merely extensions of the loop circuits connected at either end. They must provide a two-way voice channel, which can be either a two-way (two-wire) circuit, or two one-way circuits with one in each direction (four-wire), and they must provide a two-way channel for supervisory control signals such as dial information and busy signals. Although two-wire circuits operating in the same manner as loops may be used, four-wire circuits are preferred for interoffice connections. Carrier systems are more commonly used so that fewer cable pairs between offices may be used. The transmission objective on such an interoffice trunk circuit is that the total transmission loss between the points where the two loops are connected should approach 0 dB. Each loop is allowed a 3- to 5-dB loss maximum so that the net connection loss measured at 1000 Hz is no more than 6 to 10 dB.

Toll connecting and intertoll trunks are always designed on a four-wire basis. Ideally, the total transmission loss from end to end of any toll connection should be 0 dB. However, since each end is terminated in a two-wire system, which does reflect energy, it is possible to get an echo returning from the far end. This can result in one of two things happening. If the loss is too low and phase is right, oscillations may occur. More likely, there will be sufficient loss in the loops to prevent oscillation, but there will be an echo. Since the velocity of propagation over long circuits is finite, it is possible to have delays in excess of about 20 ms, with levels that make the echo intolerable to the user. The echo on circuits with shorter delays can also be unacceptable if the return level is too high. Toll trunk circuits are designed on the *via net loss* (VNL) principle, so that each circuit has a minimum amount of return loss built into it to keep the echo within acceptable levels. The VNL required for a toll circuit is calculated from the equation

$$\text{VNL} = \frac{0.2 \cdot L}{V} + 0.4 \text{ dB} \quad (17.2.1)$$

where V = velocity of propagation in km/ms

L = circuit length in kilometers

0.2 = dB loss that must be inserted per ms of delay to obtain acceptable echo levels

(0.2/V) = VNLF (via net loss factor), dB/km

Tables of VNLFs for various transmission facilities are available. (See *Reference Data for Engineers: Radio, Electronics, Computer, and Communications*, 7th ed., 1989, Howard W. Sams & Co., Indianapolis.)

If the total VNL of a given trunk circuit exceeds about 2.5 dB (corresponding to 20 ms of round-trip delay time), devices called *echo suppressors* are inserted in the channel. These are voice-operated relays that sense the high level of a near end talker on the outgoing channel and insert a high loss in the return channel for as long as speech continues. They disconnect about 100 ms after speech ceases so that the far end party can start talking. A second circuit on the return channel does the same for transmission in the other direction. Because the velocity of propagation varies widely depending on the type of transmission facilities, spacing of the echo suppressors may vary from a few hundred kilometers on cable circuits to a thousand kilometers on microwave circuits.

Noise on trunk circuits must also be kept to acceptable levels. Noise on long circuits must be kept below 44 dBrn, and those on short and medium-length circuit below 38 dBrn. The reference level for the noise

measurements is derived from an arbitrary minimum level of output from a white noise source that has first been passed through a weighted filter network with a frequency response similar to a standard telephone circuit. In terms of the Western Electric 3A telephone set (*C* message weighting, or the 500-type telephone receiver), 0 dBrn of noise corresponds to -90 dBm (referenced to 1 mW of 1000-Hz pure tone), or to -88 dBm of white noise in the band pass from 0 to 3000 Hz. Special noise measuring sets are used in practice, which contain a white noise source, a weighting filter, a calibrated attenuator, and a meter for measuring and comparing the channel noise to the reference source noise. The abbreviation dBrn simply means decibels relative to the reference noise level.

Private Telephone Networks

Many companies and institutions use private telephone networks to facilitate communication among their own members. These systems can be anything from two or three handsets on a single loop to an international network involving several centers, each with hundreds of stations. Private networks may be operated on a fully manual basis with all control and switching performed by operators, or they may be fully automatic switching systems similar to the public telephone network, or, as is more probable, they are a combination of both.

The small manually operated system, with several telephone stations near a single switchboard, is called a *private branch exchange* (PBX). Internal switching is performed by the company telephone operator, who often doubles as the office receptionist. One or more dial trunk systems connect the PBX to a nearby public exchange office, where they terminate on loop circuits corresponding to the number assigned to the company. Usually, this is a single number with access to several trunk circuits and a special switch that searches for an idle trunk circuit.

The small automatic system is known as a *private automatic branch exchange* (PABX). This system is usually set up so that calls between stations within the system are done on an automatic switching basis by equipment similar to that used in the exchange offices of the public system. One or more dial trunks connect the switching system with a nearby public exchange and may be arranged for automatic dial-out operation with manual answering or may be completely handled by the operator. An operator is usually provided with a manual switchboard to allow handling of outside calls and also special service internal calls.

Large companies may have several PBXs or PABXs at different locations many kilometers apart. Communications between these PBXs may be done by means of long-distance calls over the public network, but if the volume of traffic is high, the company can get preferred rates by renting the use of private trunk circuits or tie lines between the various PBXs, which are reserved only for their use. Some large companies may even install their own transmission facilities over considerable distances.

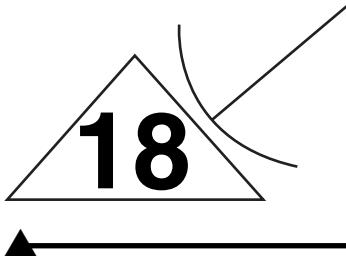
Although a company may purchase, install, and maintain its own private telephone system, it has generally been more convenient to rent such systems from the public telephone companies. All the station and switching equipment then remains the property of the public company, but its use is reserved specifically for one customer. The service provided under these conditions must be at least as good as that provided by the public system and in some instances must meet even higher requirements because of special services such as computer data links.

PROBLEMS

- 17.1. Describe each of the BORSCHT functions of a subscriber line interface circuit.
- 17.2. (a) Describe the nature of signals produced on the subscriber's loop by a pulse dialer. (b) Describe the nature of the signals produced by a Touch Tone dialer.

- 17.3. What frequencies appear on a telephone loop when (a) the 2 key, (b) the 6 key and (c) the * key is depressed on a Touch Tone dialer?
- 17.4. (a) Calculate the dc loop resistance of a telephone subscriber loop comprised of 5.6 km of AWG 22 cable and a further 10 km of 19 AWG cable. (b) If M88 loading is used, find the overall signal attenuation at 1000 Hz.
- 17.5. How many M88 loading coils would be required on a $1200\text{-}\Omega$ loop of No.16 AWG cable?
- 17.6. Explain the purpose of the induction coil in a telephone station set.
- 17.7. Explain the term *full duplex* as applied to telephony.
- 17.8. Explain the principle of operation of the transformer-type transmission bridge in Fig. 17.1.5.
- 17.9. A repeater circuit similar to that of Fig. 17.1.7 is to be included in a two-wire telephone circuit. The maximum reflection return loss on the east circuit is 40 dB and that on the west circuit is 55 dB. The two hybrid transformers each contribute a feedthrough return loss of 50 dB and a transmission loss of 3 dB in either direction. If a 10-dB gain margin to prevent oscillation is allowed, find (a) the maximum gain allowed for each amplifier based on the local loop, and (b) the maximum gain allowed for each based on the transmission circuit loop gain.
- 17.10. Explain with the aid of equivalent circuits the operation of a four-wire terminating set.
- 17.11. (a) What is the main advantage of the crossbar switching system over the step-by-step system? (b) What advantage does the electronic system have over the crossbar system?
- 17.12. Explain the term *nonblocking* as applied to a cross-point switch matrix.
- 17.13. Describe how a time switch unit works. Use a sketch to illustrate.
- 17.14. A DS1 digital signal contains 24 eight-bit PCM channels time-division multiplexed on a cable. One synchronizing bit per frame is supplied, and the sampling rate is 8 kHz. Determine the bit rate on the cable.
- 17.15. Five DS1 signals are interleaved to form a DS120 signal. One extra byte is provided for synchronization per frame. Determine the bit rate of the DS120 signal.
- 17.16. (a) Find the via net loss of a trunk circuit that involves 2000 km of microwave relay circuits with an average velocity of propagation of 299 Mm/s. (b) What is the round-trip delay on the circuit? (c) How long would this circuit have to be to require echo suppressors?
- 17.17. Find the via net loss required for a 5000-km cable circuit on which the average propagation velocity is 80% of that of light. Does the circuit require an echo suppressor?
- 17.18. A telephone channel modulated on a microwave carrier traverses a 41,000-km uplink path and a 42,000-km downlink path. (a) What is the return time delay encountered? (b) Is an echo suppressor required?
- 17.19. Explain the term dBrn as used in connection with telephone circuit noise measurement. Give the definition of 0 dBrn as referenced to a 1000-Hz pure tone.
- 17.20. Calculate the time required to complete the dialing process when the number “9387829029” is dialed using a pulse dialer.
- 17.21. Repeat the above problem when DTMF dialer is used.
- 17.22. Calculate the number of connections (links) required to connect N users in a *completely connected* network.
- 17.23. A DS2 digital signal consists of four interleaved DS1 signals. One extra synchronization byte is inserted every frame. Determine the bit rate of the DS2 system.

- 17.24.** Calculate the dc loop resistance of a telephone subscriber loop comprising of 15.6 km of AWG22 cable and a further 20 km of AWG19 cable. Round trip loop resistances are $170\Omega/\text{mile}$ and $85\Omega/\text{mile}$ for AWG2 and AWG19 cables, respectively.
- 17.25.** Find the VNL of a trunk circuit involving 4000 km of microwave relay circuit with an average velocity of propagation of 299Mm/s.
- 17.26.** What would be the round-trip delay on the above circuit? How long does this circuit need echo suppressors?



Facsimile and Television

18.1 Introduction

In addition to basic signals consisting of speech, music, or telegraph codes, a telecommunications system is often required to transmit signals from a source of a visual nature. *Facsimile* (commonly called fax) means an *exact reproduction*, and in facsimile transmission an exact reproduction of a document or picture is provided at the receiving end. *Television* means *visually at a distance*, and a television system is used to reproduce any scene at the receiving end. It differs from facsimile in that the scene may be “live” (that is, including movement).

Information is transmitted at a much faster rate in television transmission than it is in facsimile transmission. As a result, television transmission requires a much larger channel bandwidth, and special wideband circuits are required. The small bandwidth required for facsimile makes it suitable for transmission over normal telephone channels.

18.2 Facsimile Transmission

Some of the uses of facsimile transmission include the transmission of photographs (for example, for the press), the transmission of documents, weather maps, and so on, and the transmission of language texts for which teleprinters are not suitable (for instance, text in Japanese) so that many of the problems encountered in international operation are avoided.

Facsimile Transmitter

A facsimile transmitter must perform several functions, as follows:

1. It must dissect the source document image into a matrix of picture elements (*pixels* or *pels*) and then sense an image density and/or color for each pixel in a sequential manner a line at a time. This process is called *raster scanning* and is similar to the one used in television cameras.

2. The signal information derived from the scan, along with scan synchronization signals, may be modulated directly onto an analog transmission channel. Alternatively, the information may be encoded for transmission on a digital data channel.
3. The digital information may undergo a process of data compaction to reduce the amount of data that needs to be transmitted and thus reduce the time required for transmission. Synchronizing information is encoded as well.
4. Once the encoding and compaction have occurred, the digital data may be stored in memory for transmission at a later time, or they may be retransmitted without rescanning.

Source Documents. Source documents of a number of types may be transmitted by facsimile. These include cut sheet printed or typed material in a number of standard sizes, handwritten sheets, drawings and maps, and color or black and white photographs.

Scanning. Raster scanning is almost always used in facsimile systems. In a raster scan, as shown in Fig. 18.2.1, the image to be scanned is divided vertically into several horizontal segments called *scan lines*. Each scan line is then divided into a number of equal short segments or picture elements (called pixels or pels). Each pixel may be assigned one of two levels (black or white) in a digital scanning system, or it may be assigned a gray (density) level.

The scanning process examines each line in turn (usually from top to bottom), extracting the value for each pixel on the line in a sequential manner (usually scanning from left to right), until the entire image has

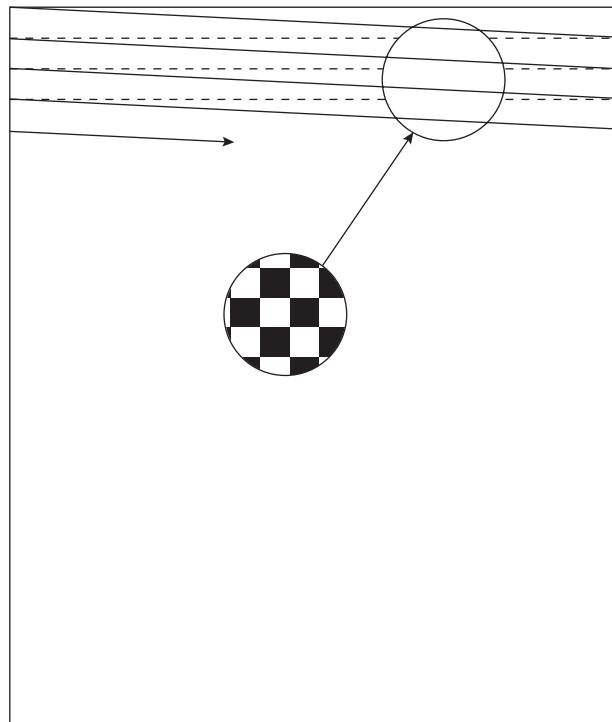


Figure 18.2.1 Raster scanned page showing an alternating black and white pixel pattern.

been scanned. Beginning and end of scan marks are inserted with the picture data to provide synchronization between the transmitter and receiver.

Sensing the level or value associated with each pixel is done optoelectrically. This requires a source of illumination to produce a standard level of light on the source pixel and a photoelectric transducer or sensor to convert the light from the pixel to an electrical signal. Selection of the pixel to be evaluated may involve one of three basic mechanisms, as follows:

1. A finely focused spot of light from the illuminating source is directed at each pixel in turn, and a single broad-angle photocell is used to sense the level of the reflected light from it, as illustrated in Fig. 18.2.2(a). The illumination spot may be moved over the fixed source document (the flying spot), or, alternatively, the source spot may be fixed and the document moved under it. The latter method is more commonly used.
2. The document is completely floodlit and a single photocell is optically focused on the desired pixel, as illustrated in Fig. 18.2.2(b). Again, the document may be moved and the sensor fixed, or the document may be fixed and the sensor moved, or a combination of these.
3. The document is floodlit and the entire image (or a line at a time) is optically focused on a photosensitive surface or photocell array, as illustrated in Fig. 18.2.2(c). Scanning is then controlled electronically and the document, source, and sensor remain fixed in position. A combination is often used in which the document and illumination are fixed, and a line-sensing array is moved vertically along the document.

The mechanics of providing the scanning motion lead to two basic document-handling modes. The first is *cylindrical scanning*, which was very popular in fax machines until recently. The second is called flat-bed scanning, where the document is kept flat and usually in a fixed position.

Cylindrical Scanning. In this method the document is first fixed around a drum by means of clips. The drum turns on its axis, and as it does so it moves along a threaded track so that each turn of the drum displaces it axially by one line width. Figure 18.2.3 illustrates a cylindrical scanner in which the document wrapped around the drum is floodlit. The photocell is optically focused through a beam-forming aperture on a spot on the drum surface, which is fixed relative to the drum axis and starting position. As the drum rotates, the sensor spot follows a spiral path around the drum from one end to the other. The clips can be used to create an end-of-line signal for synchronization.

In earlier models a rotating disc with opaque segments was rotated in the sensor beam so that it “chopped” the beam. This created an alternating current output from the sensor whose amplitude was modulated with the light intensity information and whose frequency depended on the disc rotation speed. Amplification of the resulting ac signal was much easier than the varying dc level from an unchopped sensor. Alternatively, the mechanical chopper may be omitted and the dc signal amplified to a level to allow modulation on an ac carrier electronically.

Typically, the drum was rotated at about 60 rpm with an axial traverse of about 0.25 mm per turn (0.01 in. per turn), producing four scan lines per millimeter along the document, or about 1100 lines for a 280-mm (11-in.) sheet. A North American television scan uses 525 lines to compose a complete vertical scan, while facsimile systems may use 1100 to 1800 lines in a complete document scan.

Electronic CCD Scanning. Photosensitive charge-controlled devices (CCDs) developed in the 1980s have greatly improved the fax transmitter unit. These devices use MOS technology to integrate an array of photosensors with an electronic scanning system built into a special VLSI chip.

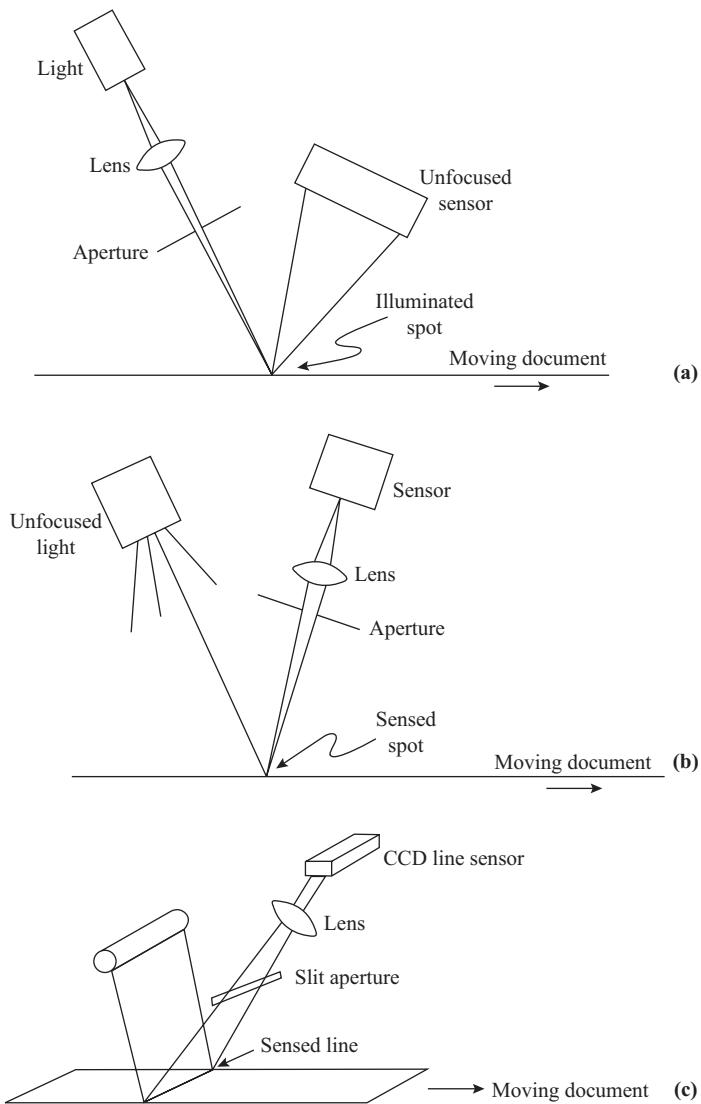


Figure 18.2.2 (a) Flying spot scanning. (b) Flying sensor (floodlit). (c) Electronic scanning (floodlit).

In these devices, each pixel has an associated sensor cell on the VLSI chip that is made up of an isolated charge well connected through a photodiode to a charge supply rail. Initially, the well is discharged. When the diode is exposed to the light from its corresponding pixel, the photocurrent produced in the photodiode allows the well to charge from the supply rail. After a fixed charging time, the charge accumulated in the well is gated through a MOSFET to one cell of a “bucket brigade” shift register. All the sensors are transferred at the same instant and the wells discharged to reference in preparation for the next cycle.

While the sensors are accumulating a new set of charges, the old set is shifted down the shift register and presented, one pixel at a time, to a charge amplifier to create an output voltage proportional to the charge accumulated in the pixel well. Scan synchronization pulses are added to the scan train automatically by the

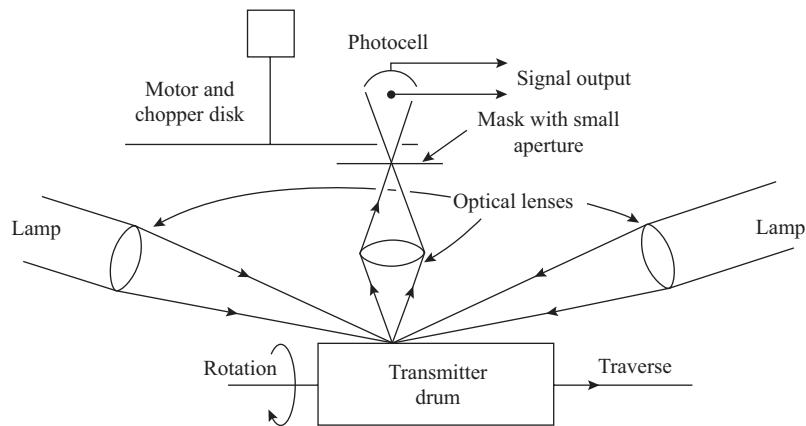


Figure 18.2.3 Drum scanner with focused sensor (floodlit).

shift register. Once all the data have been shifted out to the amplifier, a new set of charges are transferred from the sensor wells into the shift register, beginning a new cycle.

A typical line-scanning CCD chip (such as the Texas Instruments TC103) contains an in-line array of 2048 CCD photocells, each with an area of $12.7 \times 12.7 \mu\text{m}$ so that the chip is more than 2.6 cm long. It is mounted in a 24-pin DIP package with a glass slit window on the top to expose the photocells.

CCD scanners are usually of the flat-bed type, as opposed to the drum scanner. The mechanics of a CCD scanner are illustrated in Fig. 18.2.4. The document is placed face down on a fixed glass plate window

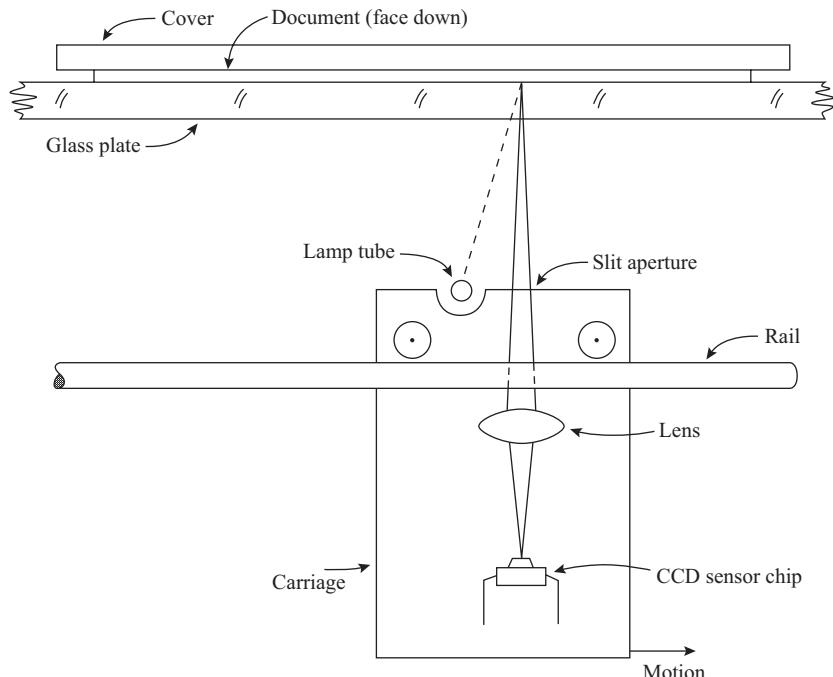


Figure 18.2.4 CCD line-at-a-time flat-bed scanner.

on the top of the unit and a light cover closed over it. The rest of the scanning mechanism is mounted on a carriage on rails so that it can traverse the length of the document. The carriage drive speed is controlled so that it will move one line width (0.25 mm) during the time of one electronic scan cycle in the CCD chip.

A high-intensity tubular lamp mounted on the sensor carriage illuminates the face of the document to be scanned. The image of one scan line across the document is focused through a slit and a series of lenses into the slit-window of the CCD chip onto the sensor array. Once the carriage has completed its traverse of the document, it is driven back to its starting position at high speed.

If the system is to reproduce in color, then the focused line is passed through a three-way beam splitter and three chromatic filters to three CCD sensor chips, one for each of the primary colors, red, green and blue, and the three data streams are transmitted. If the system is to produce color or gray-shaded black and white, then the data from the sensor amplifier for each pixel are quantized and coded, typically as 4 bits for 16 gray levels or 12 bits for 16 levels in three colors.

The quantity of data that must be transmitted for a complete document is very large, and if it is to be transmitted over ordinary telephone channels, then it will take a long time to transmit it. To reduce the amount of time taken, the binary coded systems use special data compaction codes, which in effect only transmit information when a level transition takes place between adjacent pixels. This type of coding is very effective in two-level black-and-white systems, with data compaction ratios of more than 40 : 1 achievable. The amount of compaction possible depends on the type of image being transmitted. Documents with a lot of white space and broad black lines, not too densely spaced, will produce high ratios, while detailed photographs will not produce good compaction at all. A digital system will typically transmit a typewritten 8.5×11 in. page in about 1 min.

The Scanning Spot. The size and shape of the scanning spot on the source document that is presented to the detector is important. Both are determined by the size, shape, and positioning of the aperture used in the spot focusing system.

First, if the scan spot width is less than the center spacing of the scan lines, a strip of scan data between each pair of lines will be omitted. If the receiver recorder has the same size and spacing, a light line will be interposed between the two scan lines.

If the scan lines overlap, then the intensity in the overlap region may be higher than in the nonoverlap region, resulting in a dark stripe interposed between scan lines.

If the scan lines are abutted, but the scan spot is not rectangular or trapezoidal, as shown in Fig. 18.2.5, then the intensity in the overlap region may be above or below the average at the center of a scan line, again giving interposed light or dark lines. In the case of rectangular or trapezoidal spots properly aligned, the intensity does not vary across the overlap region.

Circular scan spots are often used because of the ease of implementation, in which case the line overlap (that is, the spot size) has to be adjusted to minimize the density of the interposed lines created.

Scan spot and alignment problems can occur both in the source scanning and in the recorder scanning at the far end. In the case of the transmitter, the problem can result in intensity error or lost data.

Facsimile Receiver

The facsimile receiver performs the reverse process of the facsimile transmitter. The transmitter scans the source document and extracts data about each pixel in turn, modulates or encodes these data, and sends them on a communications channel. The receiver must get the modulated or encoded signal from the channel, demodulate or decode it to recover the serial pixel information, and then record that on the target document through another scanning process, in the proper order and relation. The scanning process in the transmitter and in the receiver is similar, and very often the mechanism used is identical.

Facsimile systems may operate in one of two basic modes. The first of these, which was universally used until recently, involves scanning, transmission and recording in a real-time, synchronous manner so that the

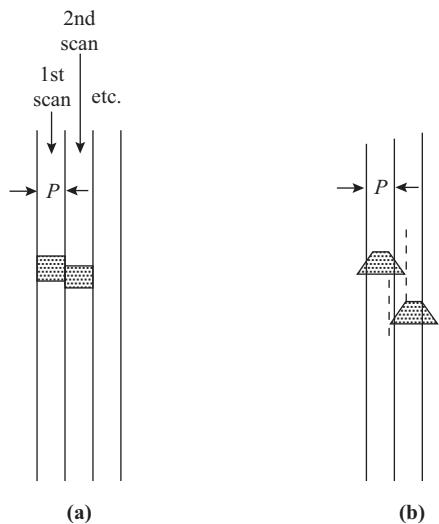


Figure 18.2.5 Scan spot shapes and line overlap.

receiver mechanism scans at the same speed and in step with the transmitter mechanism. The information may be digitally encoded for transmission, but is transmitted at a synchronous rate.

The second method involves the encoding and storage of the data in binary form for transmission at a later time and very possibly at a different rate. This method results in asynchronous operation, where the receiver scanner does not have to operate at the same speed as the transmitter scanner.

In any case, if the received image is to have the correct relationship to the transmitted image, the receiver must be synchronized to the transmitter, phased correctly, and have the same height-to-width ratio as the transmitter.

Synchronization. In a synchronous system it is necessary for the receiver scanner and the transmitter scanner to run at exactly the same speed. If the receiver runs at a slightly faster speed than the transmitter, then each scan line will overlap into the following scan line and delay its start, causing the image to skew diagonally to the right, as shown in Fig. 18.2.6. If the receiver is slower, then the image will skew diagonally to the left.

Some early systems simply used synchronous 60 rpm ac motors running on the common power grid to obtain synchronization. This was satisfactory for some line drawings and handwritten material, but it was not sufficient for documents and photographs.

Some more recent systems used separate but highly precise local oscillators at both transmitter and receiver to maintain synchronization. In this case the generated signals had to be maintained within 10 ppm of the standard frequency in order to maintain this synchronism. The synchronous drive motors are driven directly from the local oscillator signals.

Most systems now use a standard carrier frequency (usually 1020 Hz) generated synchronously by the transmitter, sent over the communications channel. The receiver locks onto this carrier to generate a local oscillator signal whose frequency is forced to be synchronous with the transmitter. This synchronous local oscillator signal is used to drive the receiver scanner at the synchronous speed. Standards require that the receiver speed be maintained within 10 ppm of the transmitter speed, corresponding to ± 0.0006 rpm of 60 rpm.

In the asynchronous digital systems, the horizontal and vertical scanning mechanisms may be run at any convenient speed and do not have to run at exactly the same speed as the transmitter. It is only necessary

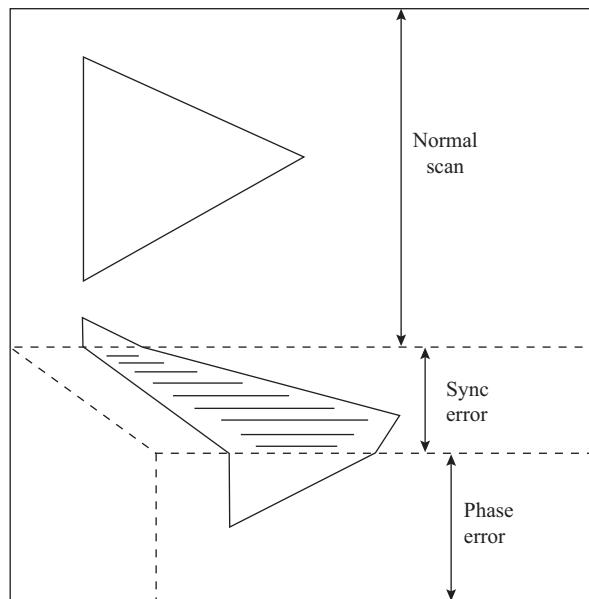


Figure 18.2.6 Facsimile image distortion due to synchronization and phasing errors.

to ensure that the data for each line are placed on that line and are spaced properly within that line. The information for reconstructing the pixel sequence of each line is included in the binary coded data stream, and its extraction is part of the decoding process. Timing of the release of these data to the recording system is controlled by the recorder itself. Synchronization of data on the transmission system is another matter.

Phasing. In synchronous systems, if the receiver and transmitter are in synchronism, but the receiver starts a new line either too early or too late, the image will wrap around. In the drum scanning systems, a 1-s phasing pulse is transmitted at the beginning of a document scan to stall the receiver until its phase matches that of the transmitter. The phasing pulse is ended at the beginning of scan on the first line, and both receiver and transmitter begin the first line at the same time. This results in the end-of-line gap coinciding with the end of the line, at the edge of the recorded document instead of in the middle.

In asynchronous digital systems, phasing is ensured because no line data are released to the recorder until the recorder indicates that it is beginning a new line. The decoding process ensures that only data for that line are sent to the recorder during that line scan period.

Index of Cooperation. The width-height ratio of the document reproduced at the receiver must be the same as that of the originating document if distortion is to be avoided. The index of cooperation (IOC) is a number derived from the width-height ratio, and for proper reproduction the transmitter and receiver must have the same IOC as the transmitter. Figure 18.2.7 shows what happens to the reproduction if the IOC at the receiver is different from that of the transmitter.

The IEEE defines the IOC as the product of scan density and stroke length. Referring to Fig. 18.2.8 and given that

$$\text{IOC} = \text{index of cooperation}$$

$$S = \text{scan density or resolution (lines/mm)}$$

$$W = \text{width of source document or scan stroke length (mm)}$$

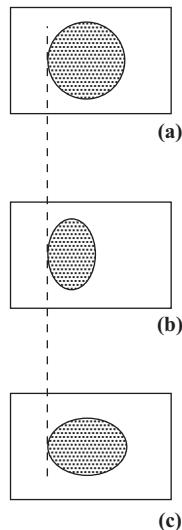


Figure 18.2.7 Image distortion due to index of cooperation incompatibility. (a) Receiver IOC equal to transmitter IOC. (b) Receiver IOC greater than transmitter IOC. (c) Receiver IOC less than transmitter IOC.

L = document length (mm)

n = total number of lines in document height

D = drum diameter (mm)

P = scanning pitch (mm/line)

then, by definition,

$$\text{IOC(IEEE)} = S \cdot W \quad (18.2.1)$$

The scan density is related to the total document length by

$$S = \frac{n}{L} \quad (18.2.2)$$

Thus, in terms of the width-length ratio,

$$\text{IOC(IEEE)} = n \frac{W}{L} \quad (18.2.3)$$

Note that if the received and transmitted documents have the same number of lines and the receiver and transmitter IOCs are the same, then the condition for cooperation is met, and the received document will have the same width-to-length ratio as the transmitted document. This definition of IOC is suitable for both drum and flat-bed scanning systems.

The CCITT (Comité Consultatif International Télégraphie et Téléphonique) developed a slightly different definition of IOC, which is directly applicable to drum scanners. By definition, the CCITT version of the IOC is the ratio of diameter to pitch,

$$\text{IOC(CCITT)} = \frac{D}{P} \quad (18.2.4)$$

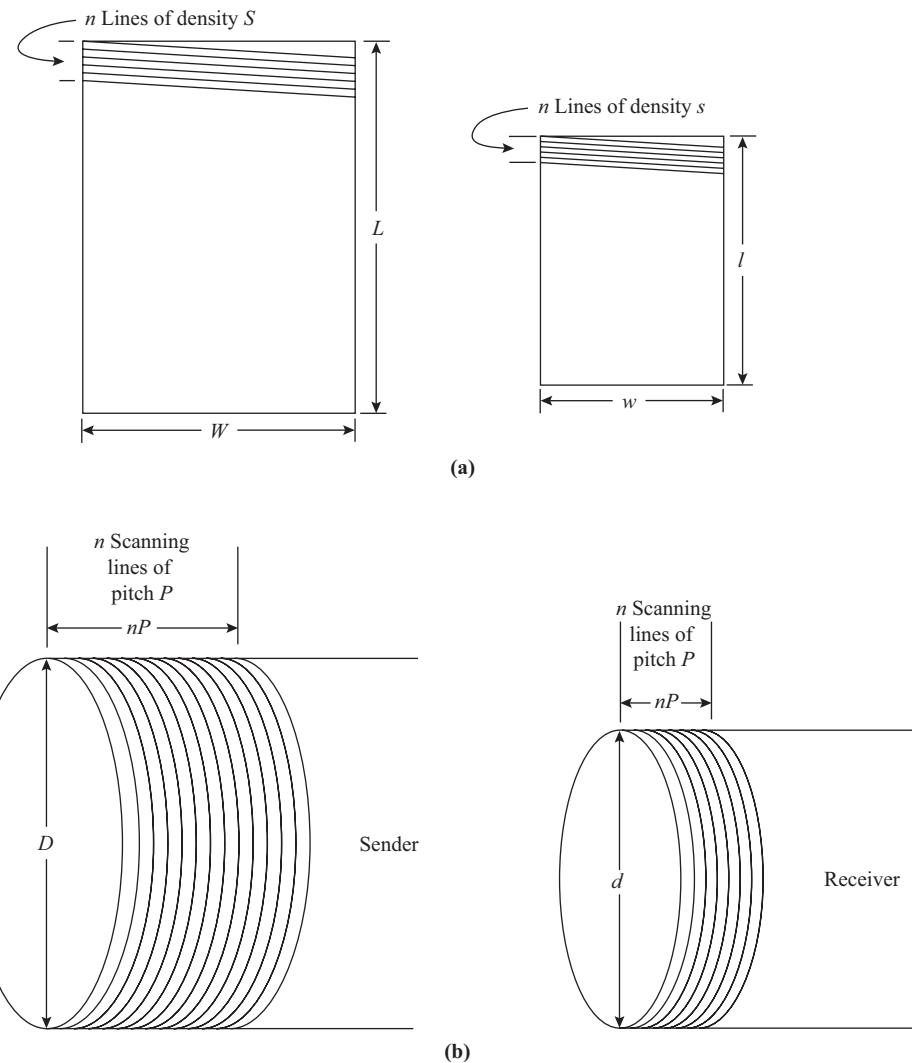


Figure 18.2.8 (a) Page scanning parameters. (b) Drum scanning parameters.

But the document width is exactly the drum circumference,

$$W = \pi D \quad (18.2.5)$$

and the pitch is the inverse of the scan density, or

$$S = \frac{1}{P} \quad (18.2.6)$$

Substituting from Eqs. (18.2.1), (18.2.5), and (18.2.6) into Eq. (18.2.4) gives

$$\text{IOC(CCITT)} = \text{IOC} \frac{(\text{IEEE})}{\pi} \quad (18.2.7)$$

Again, if the IOCs of transmitter and receiver are identical and document lengths are the same, the reproduction will be undistorted.

EXAMPLE 18.2.1

The drum diameter of a facsimile machine is 70.4 mm and the scanning pitch is 0.2 mm per scan. Find the index of cooperation.

SOLUTION Where $P = 0.2 \text{ mm/scan}$ and $D = 70.4 \text{ mm}$, by Eq. (18.2.4),

$$\text{IOC(CCITT)} = \frac{D}{P} = \frac{70.4}{0.2} = 352$$

and, by Eq. (18.2.7),

$$\text{IOC(IEEE)} = \text{IOC(CCITT)} \cdot \pi = 352 \times \pi = 1106$$

Photographic Recording. Once the image pixel information has been recovered from the modulation or encoding at the receiver and placed in the proper sequence, it must be recorded to create the image facsimile. Photographic recording is popular for the transmission of news photographs and other material requiring gray-level coloring, although it may also be used for printed black-and-white documents.

Transmission of photographic facsimile requires either an analog-modulated signal or a digital encoding system that includes gray-level information, typically to 16 levels. In the recording process, photographic print paper or transparency film (either positive or negative) may be used. A focused light spot is required, which can be made to scan the photograph sequentially (as in the flying spot transmitter). The intensity of the light spot must be capable of being modulated with the pixel signal intensity. Photographic recording requires handling of the unexposed paper or film in a darkroom environment, and also requires postprocessing to develop and fix the exposed images, either automatically in the receiver or in a separate darkroom process.

An early form of the photographic receiver used a *Duddell mirror oscilloscope* (a small mirror attached to a galvanometer moving coil) as a modulator. The galvanometer positioned the light beam so that more or less of it passed through an aperture onto the photo paper on a drum scanner, thus modulating the intensity of the light.

A more recent version used a *crater lamp* as the light source for a drum scanner. The crater lamp is a special gas discharge tube whose light intensity is approximately proportional to the applied voltage, making modulation a very simple procedure. The crater lamp has also been used in a flat-bed scanning system, where the document is moved lengthwise, and a rotating multifaceted mirror is used to sweep the spot from the crater lamp along each horizontal line in turn, similar to the laser system described next.

Lasers have provided the means for modulating and focusing a light beam with a minimum of mechanical moving parts. Figure 18.2.9 shows the arrangement of a flat-bed laser-scanned photographic recorder in simplified form. In this system, a semiconductor laser produces a narrow beam of light (usually in the red range). The beam first passes through an electroacoustic modulator where its intensity is varied according to the pixel intensity data received. It is then passed through a lens to focus it to a point at the photographic paper surface. Next it is reflected from a multifaceted rotating mirror, which causes it to sweep along each scan line at one line per facet, and finally reaches the photographic paper, which is carried on a moving flat-bed carriage (or through a series of rollers) to produce the vertical scanning motion. The exposed photo paper is then either passed through a developer system included in the receiver or is taken to a separate system for development under safe-light conditions.

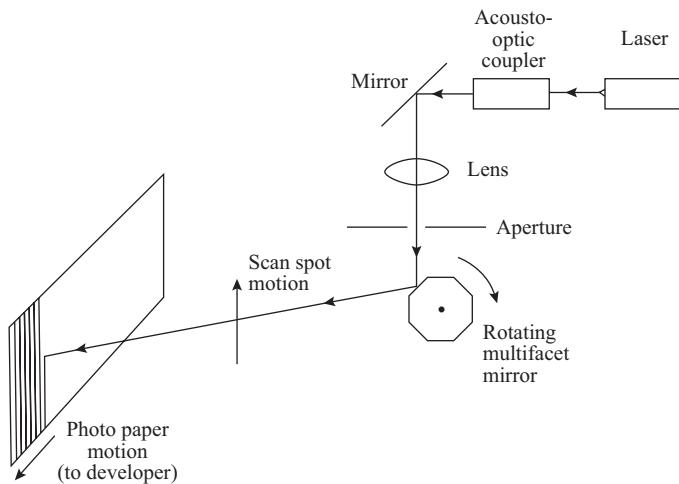


Figure 18.2.9 Photographic laser scan recording.

The 3M Corporation has developed a *dry silver* paper, which is used in Harris Corporation's Laserfax and Laserphoto lines of receivers. This eliminates the wet chemical postprocessing, since the exposed paper need only be passed through a set of heated rollers to develop it.

Direct Recording. Direct recording usually means that the image is transferred directly to the copy paper, and no postprocessing is required to develop the image. One form of this type uses an *electrolytic paper* that is chemically treated with a damp electrolyte. Signal voltage is applied to the paper directly through a scanning stylus to cause the electrolyte to dissociate to produce a metallic salt, which in turn reacts with a coloring agent to mark the paper. Steel styli are preferred since they produce an intense black mark. The paper is damp and must be kept in sealed containers. It is not expensive, but has a short lifetime and produces a poor tonal range and definition.

Another system is the *electroresistance* recording system. This uses a special paper called *Teledeltos* paper, which has a metallized backing, covered with a layer of coloring agent similar to carbon black, and that in turn covered with a thin opaque insulating layer. When a metal stylus passes over the paper, current produced by the modulating signal variously burns the opaque layer, exposing the carbon layer as a mark. Tonal range and definition are not exceptional, the paper is expensive, and the burning process produces an acrid odor that may be offensive.

Electrothermal recording is becoming popular with an improved process. This also uses a special paper treated with a dry chemical that darkens locally when exposed to heat over a threshold temperature value. The heat is applied through a special fast-responding resistive heating stylus whose temperature can follow the variations of the data signal. The temperature must be carefully controlled so that it varies about the threshold value. It has moderately good definition and tonal characteristics, and the paper is not too expensive. Canon markets a dial-up fax system that uses this process.

Transfer xerography means literally "transfer dry pictures," as coined by the Xerox Corporation for their photocopies. The process is also used for facsimile recording and is popular because it uses an ordinary, fine-finish untreated bond paper that is inexpensive and produces acceptable reproductions, especially of black-and-white printed material. The process is outlined in Fig. 18.2.10.

The heart of the system is a rotating drum that is covered with a thin layer of selenium. Selenium is a semiconductor that under dark conditions is a very good insulator, but that becomes a moderately good

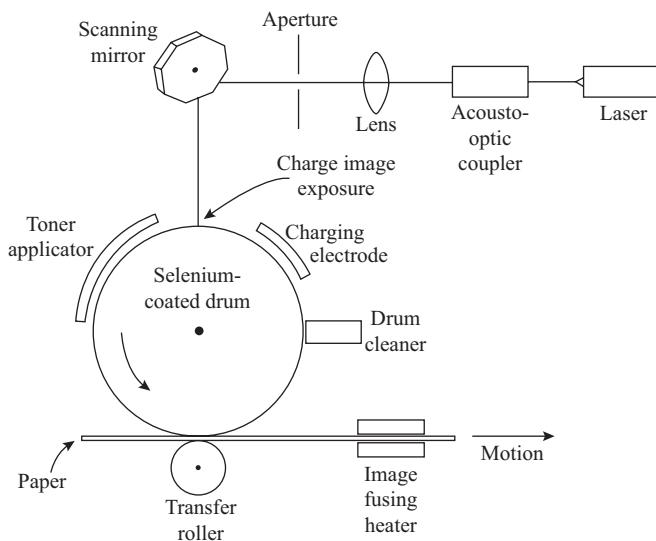


Figure 18.2.10 Transfer xerographic recording.

conductor when it is illuminated because of the creation of hole-electron pairs by the incident light. The layer of selenium is first exposed to a high electrostatic field, which places a uniform layer of trapped charges on the surface. As the drum rotates, it is exposed to the modulated scanning light source, which is either a laser/rotating mirror system as above, or is a one-line cathode ray tube that is focused on the surface. The light areas correspond to white, and where the selenium layer is illuminated, the photocurrent drains off the surface charge, leaving a charge image on the drum.

The turning drum is next dusted with a dry powder black toner that adheres where charge remains. The drum turns then against a pressure roller that transfers the toner image from the drum surface to the copy paper. The paper finally passes through a heater that fuses the toner image to the paper surface to complete the printing process. Meanwhile, the turning drum is cleaned of any remaining toner and recharged, ready for the next page.

Since the image undergoes a mirror reversal during transfer, it is necessary for the data to be scanned in the reverse direction during exposure. Vertical scanning continues from top to bottom.

Transmission of Facsimile Telegraph Signals

Analog Fax Transmission. The basic signal from the scanner of a facsimile machine cannot be transmitted directly on communications channels because these channels will not pass the dc and low-frequency components of the signal. As a result, some form of modulation is always used.

Transmission rates are limited by the capacity of the telephone circuits used. In some cases, special broadband circuits are established for private networks, and on these the data rates are much higher.

Transmission time (or the length of time required to transmit a document) is closely related to transmission rate. Higher transmission rates mean lower transmission times.

Bandwidth requirements for a synchronous fax system can be calculated from the scanning parameters of the transmitter. In the worst case, the image is a pattern of alternate black and white pixels, as shown in Fig. 18.2.1, with each pixel being a square of one line width on a side.

The ideal output waveform will be produced by using a scanning slit that is much narrower than the pixel width and will produce a rectangular waveform at the sensor output, as shown in Fig. 18.2.11. Using a

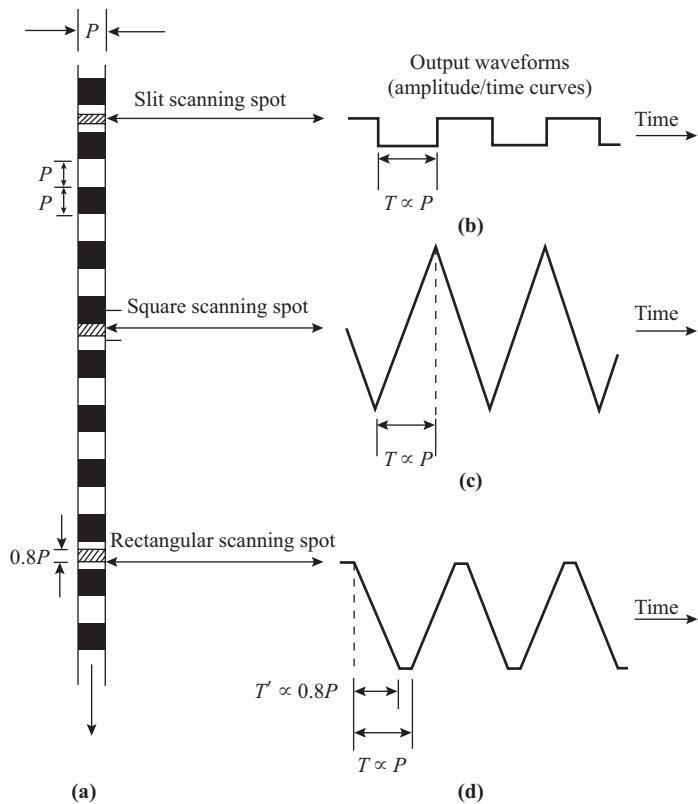


Figure 18.2.11 Effect of scan aperture on output waveform.

square slit of the same size as the pixel will produce a triangular waveform. A good compromise is to use a slit that is 0.8 of a pixel width, which produces a trapezoidal waveform.

In any case, the transmission system must have a bandwidth sufficiently wide to pass the *fundamental frequency* of the sensor waveform, which can be determined as follows. Where

N_{px} = pixel density (pixels/scan line)

R_{px} = pixel rate (pixels/s)

R_s = line scan rate (lines/s)

rpm = speed of drum rotation (revs/min)

then, for a flat bed scanner,

$$N_{px} = WS \quad (\text{pixels/line}) \quad (18.2.8)$$

and, for a drum scanner,

$$N_{px} = \frac{\pi D}{P} \quad (18.2.9)$$

For a flat-bed scanner the line scan rate is R_s , while for a drum scanner one line is scanned for each revolution of the drum, giving

$$R_s = \frac{\text{rpm}}{60} \quad (18.2.10)$$

Now the pixel rate is found to be

$$R_{px} = N_{px}R_s \quad (18.2.11)$$

Every two pixels of the alternating pattern form one cycle of the sensor output, so the output frequency is

$$f = \frac{R_{px}}{2} \quad (18.2.12)$$

Document transmission time is given by the scan rate multiplied by the number of scan lines in the document, or

$$t_d = \frac{n}{R_{px}} \quad (18.2.13)$$

EXAMPLE 18.2.2

The drum scanner in Example 18.2.1 has a pitch of 0.26 mm/line and a diameter of 68.4 mm. The drum rotates at 120 rpm and scans a total of 1075 lines for a standard document page. Find the bandwidth required for the transmission channel and the length of time required to transmit a page. The number of pixels in a scan line is

$$N_{px} = \pi \frac{D}{P} = \pi \frac{68.40}{0.26} \quad 826.5 \text{ pixels/line}$$

The scan rate is

$$R_s = \frac{\text{rpm}}{60} = \frac{120}{60} = 2 \text{ lines/s}$$

The pixel rate is

$$R_{px} = N_{px}R_s = 826.5 \times 2 = 1653 \text{ pixels/sec}$$

The cutoff frequency is

$$f_{\max} = \frac{R_{px}}{2} = \frac{1653}{2} = \mathbf{826.5 \text{ Hz}}$$

Document transmission time is

$$t_d = \frac{n}{60R_s} = \frac{1075}{60 \times 2} = \mathbf{8.96 \text{ min}}$$

The *modulation* method used for fax transmission depends on the nature of the channel to be used and the characteristics of the fax equipment. Telephone circuits are most often used, while a few special services use private broadband radio or cable circuits. Both AM and FM is used extensively, and international agreements have specified carrier frequencies and modulation characteristics for use on telephone-grade circuits.

Typical telephone circuits have a frequency response characteristic that is acceptably flat in the range from 300 to 2600 Hz, as shown in Fig. 18.2.12. The response cuts off sharply below 300 Hz and tails off gradually above 2600 Hz, with usable response up to about 4200 Hz. The fax carrier and its modulation components must fit within this frequency range, which imposes a limit on the speed of operation of the fax equipment.

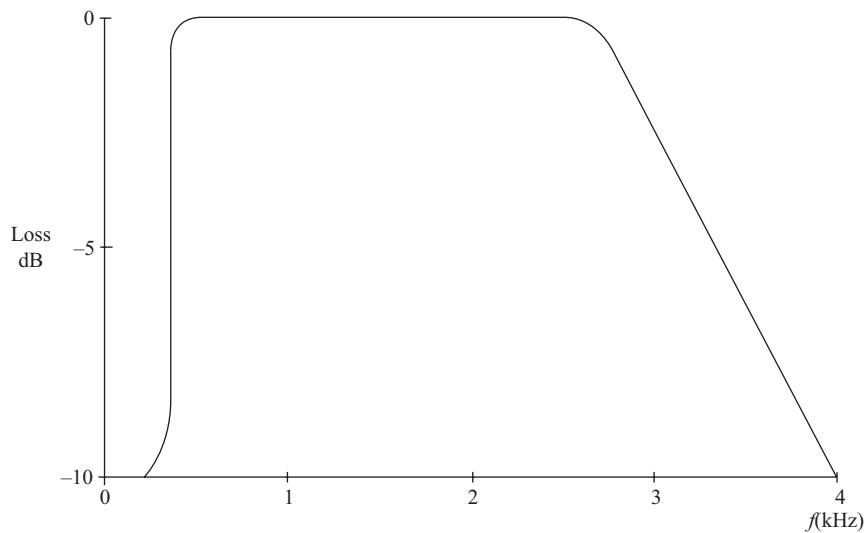


Figure 18.2.12 Frequency response of a typical telephone channel.

Two carrier frequencies have been designated as standard for AM and FM fax carriers on telephone lines, these being 1300 and 1900 Hz. FM modulators are limited to modulation indexes of less than unity, so only one significant pair of side frequencies is produced for each modulating frequency. This is called narrow-band FM.

AM or narrow-band FM with a 1300-Hz carrier is limited in speed to systems scanning at 60 or 90 lines/minute (corresponding to maximum bandwidth requirements of 400 and 600 Hz). Any higher rates would cause the lower modulation sideband to overlap the information baseband and cause interference, as shown in Fig. 18.2.13(a).

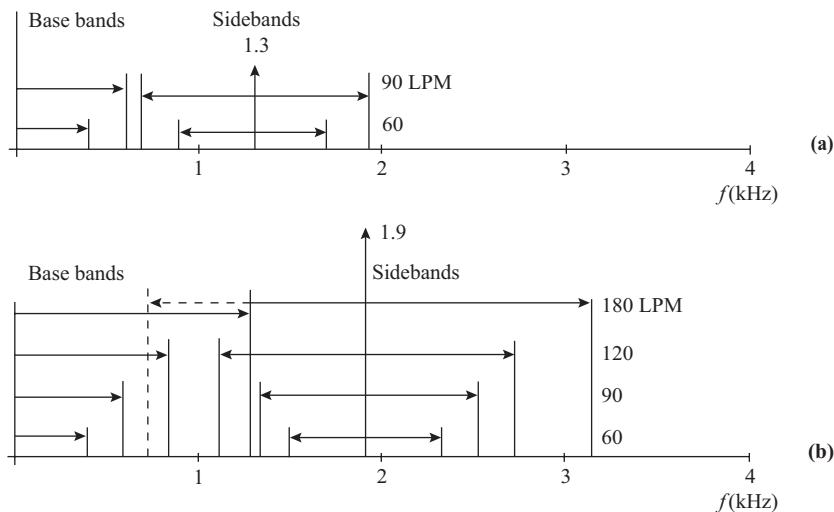


Figure 18.2.13 (a) 1300-Hz fax carrier spectrum. (b) 1900-Hz fax carrier spectrum.

AM or narrow-band FM with a 1900-Hz carrier is speed limited to a maximum of 180 lines/minute (a bandwidth of 2000 Hz). While some overlap at 189 lines/minute occurs, the degradation of quality is minimal. Some AM systems use vestigial sideband modulation, with a portion of the lower sideband suppressed to avoid the overlap. Spectrum capabilities are shown in Fig. 18.2.13(b).

These systems correspond to the CCITT group 1 equipment described in their standards T1/T2.

Noise and distortion cause image distortion and interference and must be kept to a minimum. It has been found that a signal-to-noise ratio of at least 35 dB must be maintained for satisfactory operation. Since AM is more prone to interference, FM is generally preferred. Echo on long-distance circuits is also a problem and must be avoided, especially for equipment used to transmit photographs.

The CCITT also designates in its standard T16 that narrow-band subcarrier FM be used for all HF and LF radio circuits for fax. Radio links used in the telephone network are mostly FM microwave systems anyway, but if HF or LF is used, then the subcarrier FM (SCFM) system must be used to avoid the problems that fading cause. A subcarrier frequency of 1900 Hz is recommended, with deviations limited to ± 400 Hz for HF circuits and ± 150 Hz for LF circuits.

Digital Fax Transmission. The introduction of microcomputer circuitry, digital data handling and transmission, and automatic call-handling equipment has provided the impetus for an explosion in the use of facsimile over the general telephone switched network recently. Also instrumental in this was the implementation of international standards for such equipment, notably the CCITT standards T4 and T30 as revised in 1984. These standards specify the code to be used for data compaction, which allows the transmission of a standard 8.5 x 11 in. page in about 10 s instead of the previous 6 min.

In digital transmission, the output from the scanner is first quantized, usually to two levels for black and white, and then encoded. The encoding does not proceed pixel by pixel, but uses a process of *run length coding* in which each code word specifies the distance in pixel widths from one transition to the next of color, alternating between black and white. A segment of a line is shown in Fig. 18.2.14(a).

To keep the size of the number within reason (a standard line length is specified at 1728 pixels, with larger lengths to 2560 pixels allowed) each run length is specified by two code words. The first codeword gives the integer number of 64 pixel groups in the run length, and the second gives the remainder pixels (0 to 63). End-of-line (EOL) codes are inserted in the string of run length codes at the appropriate places. This form of coding provides a first level of data compaction, since only two code words are needed for each run length group of pixels. In a sparse document with a few broad black strokes on a white background, this results in a considerable saving of code length over the encoding of every pixel. The data word grouping for a document is illustrated in Fig. 18.2.14(b).

The CCITT standard specifies a minimum-redundancy variable-word-length bit-stream code that provides some further compaction over fixed-word-length codes such as ASCII. This code is a *modified Huffman code* (MHC). It provides a unique bit string for each of the 64 black run lengths and each of the 64 white lengths, referred to as *terminating codes*, and 27 black run and 27 white run *makeup codes*, which specify the number of blocks of 64 pixels to add to the terminator to make the run length. A makeup codeword is only included if the run length is greater than 63 pixels long. There are 13 additional makeup code words, which can specify a number of blocks of 64 either black or white, extending from 1728 to 2560 pixels for wider documents. A unique codeword for the end-of-line designator is also provided.

The codewords vary in length from a minimum of 2 bits to a maximum of 12 bits. Each codeword is assigned to its run length number according to the statistical probability of occurrence of that run length in a document, with the shorter codewords being assigned to the run lengths that occur most frequently. Since each codeword is a unique group of bits, there is no need to transmit word delimiter symbols between them (such as the start and stop groups in the ASCII serial code). The bit stream for a segment of coding is shown in Fig. 18.2.14(b).

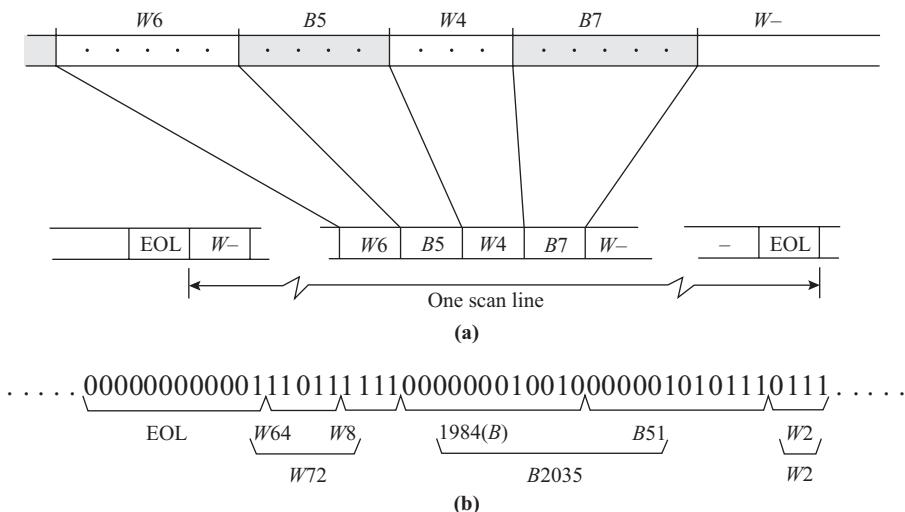


Figure 18.2.14 (a) Run length coding for a portion of a scan line. (b) Run length code string for one scan line.

A further feature of the coding system is that error checking is possible. Since the run length codes in each scan line must always sum to the same number, the presence of an error is easily detected. The simplest form of error correction involves replacing the faulty line with the data from the previous line. This produces less disruption in the finished document than simply carrying on.

This code is extremely efficient and under typical operation provides *data compaction ratios* of up to 40 to 1, with ratios around 30 to 1 being typical. The result is that a document that scanned directly in 6 min can be transmitted at the same data rate in about 12 s.

The encoded bit stream can be transmitted using any data modem pair suitable for the channel to be used. The CCITT standard V27 specifies a data modem set that can transmit at either 4800 or 2400 bps (corresponding to 1600 baud or 1200 baud) over the general telephone switched network. It uses eight-phase differential mode phase modulation of an 1800-Hz carrier for the higher speed and four-phase modulation for the lower speed.

The complete digital fax machine includes CCD line-by-line scanning, run-length and MHC coding, memory capable of storing the data for several pages of documents, the data modem, an electrothermal printer, and facilities for automatic dialing and call handling on the telephone switched network. Either automatic, semiautomatic, or manual operation may be used, and special features such as document broadcast to several receiver stations, network polling, retransmission, and others are frequently included. This is the type of fax machine that has become extremely popular with small businesses and individuals alike.

18.3 Television

Watching television is North America's favorite pastime, and it is rapidly becoming so in the rest of the world as well. As a result, the manufacture of television receiver sets forms a large portion of today's electronics industry. The systems used for television sets manufactured to date have been thoroughly standardized, so sets vary little from one manufacturer to another. The NTSC (National Television System Committee of the United States) system presently used in North America is common to all North American countries, Japan, Korea, and a few others. The PAL (Phase Alternate Line) system and the SECAM (SEquential Coding and

Memory) systems are used in most European countries and differ from the NTSC system mainly in the use of different scanning rates.

Frequency assignments for broadcasting television channels from landbased stations have been set by international agreements and are shown in Table 18.3.1. Channels 2 to 13 are in the VHF band and are the most heavily used ones. Channels 14 to 83 occupy the UHF band and are being increasingly used as the number of broadcast stations in urban areas increases. Channel width is standardized at 6 MHz.

CATV (cable television) distribution systems presently transmit signals over coaxial cable and are limited to a maximum 300-MHz bandwidth. The standard broadcast channel frequencies are used in the same order for channels 2 to 13 so that a standard receiver may be used directly on the cable without a converter. However, nine additional channels are provided (A to I or 14 to 22) in the range from 120 to 174 MHz, and five channels (J to N) in the range from 216 to 246 MHz are usually provided by CATV instead of the UHF channels. Many recent receivers have direct tuning of the CATV channels built in, but earlier models have to be used with an external converter unit to convert each CATV channel into the channel 3 band of frequencies.

Black and white television receivers became readily available around 1950, and color television sets followed around 1960. At present the color transmission system is designed so that all black and white sets can use the same color signals. Most sets being sold presently are color.

Experimentation with HDTV (high-definition television) systems began in the early 1980s and continues. Several recent and expected international standards are in preparation, and when these become accepted, it is expected that HDTV will take an increasing share of the market as people see the great improvement in picture quality that it provides.

TABLE 18.3.1 Television Channel Frequency Assignments

Broadcast Channels	Frequency (MHz)	CATV Channels	Broadcast Channels	Frequency (MHz)	CATV Channels
Other	5–54	Special	7	174–180	7
			8	180–186	8
2	54–60	2	9	186–192	9
3	60–66	3	10	192–198	10
4	66–72	4	11	198–204	11
5	76–82	5	12	204–210	12
6	82–88	6	13	210–216	13
FM Radio	88–108	FM Radio	Other Services	216–222	J (23)
Other	108–120	Other		222–228	K (24)
				228–234	L (25)
Other Services	120–126	A (14)		234–240	M (26)
	126–132	B (15)		240–246	N (27)
	132–138	C (16)		246–252	O (28)
	138–144	D (17)		252–258	P (29)
	144–150	E (18)		258–264	Q (30)
	150–156	F (19)		264–270	R (31)
	156–162	G (20)		270–300	Extra 5
	162–168	H (21)			
	168–174	I (22)	14	470–476	
			UHF	
			83	884–890	

Television Camera

The television camera contains a device that converts the optical image of the program material into an electrical signal for transmission over cable or by radio to a receiver, where the original image can be reconstructed from the received signals. Vacuum tube devices such as the Image Orthicon and the Vidicon (both developed by RCA Corporation during the 1940s) were for many years the standard, and some would argue that they are still the best.

However, advances in very large scale integrated circuit (VLSI) technology after about 1980 allowed the development of image sensor array chips using charge controlled devices (CCD). These chips can be manufactured relatively cheaply, and their availability has given rise to the popular and inexpensive camera-recorder combinations used for home video production.

Charge carriers can be confined to an area within a silicon chip by two means. First, they may be confined in a region of low potential that is surrounded by a region of high potential (for example, holes may be trapped on an island of n -type material that is completely surrounded by p -type material, forming the two plates of a capacitor whose insulator is the depletion zone of the junction separating the two regions). The second method is to form a metal-semiconductor capacitor with an oxide insulator between the two electrodes. In the case of an n -type semiconductor, application of a negative voltage to the metal gate will induce a region or *well* of negative potential below the electrode, which can hold holes in an inversion layer as in a p -channel MOSFET. The depth of the well depends on the magnitude of the voltage applied to the gate electrode.

If a deep well (more negative) is formed next to a shallower well (less negative) that contains a packet of holes (positive charge carriers), the holes will flow into the more negative region of the deeper well much as water would flow from a high tank into a lower tank. Making the gate positive raises the potential floor of the well above the potential of the adjacent area, so the holes then flow out of the well area. This is equivalent to raising the low tank above the level of the high one, thus allowing the water to flow in the reverse direction.

A CCD shift register is illustrated in Fig. 18.3.1. This device uses a two-step well, which means that only one clocking phase is required to accomplish shifting. The shifting channel is isolated from the surrounding chip by a pair of p^+ strips alongside the channel. The $n^+ - n$ region between the capacitor plates forms a stepped holding well as shown by the first profile. The region under the plates has an n^+ well separated from the holding well to the right by a p -substrate barrier. With positive potential applied on all plates, the first potential profile applies as in t_0 . Different-sized packets of holes are shown in the deeper n^+ portions of two consecutive holding wells. The high potential barrier formed by the p region keeps the holes from flowing to the right.

When a negative pulse is applied to all gates simultaneously, the profile under the gates is depressed as in t_1 until the top of the p pedestal of potential is below the floor of the holding well to the right. The packet of electrons now flows from right to left, out of the holding well, through the p region and into the deeper transfer well. The n -region pedestal prevents holes from flowing over the top to the well on the right.

When the negative shift pulse ends, the plates all return to positive voltage, and the profile returns to its original shape as in t_2 . The n^+ transfer well floor rises above the n pedestal to the left, and the packet of electrons spills to the left into the next holding well. The p barrier prevents the holes from flowing back to the right.

Application of a train of alternating positive and negative pulses shifts the contents of the register (varying-sized packets of holes) to the left and into a sensing amplifier for sequential readout.

In the Sony HAD (hole accumulation device) sensor array (TM) the active sensing takes place in an isolated island of n^+ material over a p isolation well. One pixel cell is illustrated in Figs. 18.3.2 and 18.3.3. When an isolated n^+ region is exposed to light, the light photons create surplus hole-electron pairs within the material. This makes the minority carrier density, in this case holes, proportional to the light intensity. Momentarily opening a gate into an adjacent negative well will allow a packet of holes whose size is proportional to the light intensity to be drawn into the potential well. The gate is formed by the horizontal

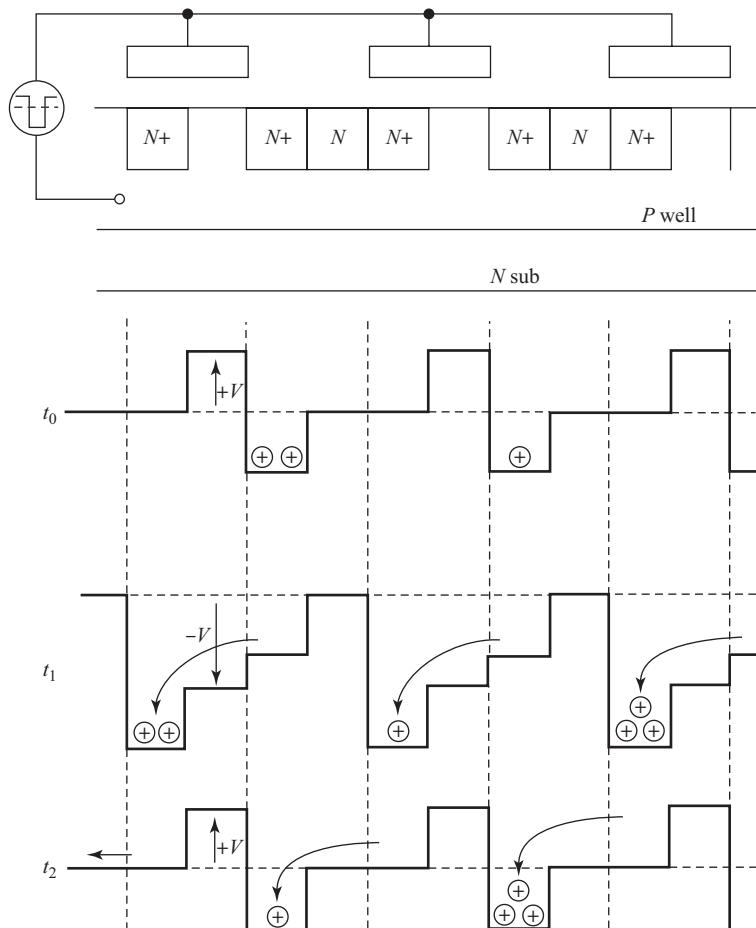


Figure 18.3.1 CCD single-phase shift register. (a) Structure. (b) Potential profiles over one cycle.

aluminum control line that overlies the n^+ holding wells along its length and the isolating p barriers to each pixel cell.

At the beginning of each line scan, the corresponding control line is pulsed negative, and the contents of all the pixels along that line are simultaneously dumped into the shifting line holding wells. Now the vertical signal control lines are pulsed negative to shift the packets to the left at the proper rate to produce the analog video signal at the line sense amplifier. The analog signal can be extracted at this point or the signal may be digitized (very fast PCM).

As shown in the floor plan of Fig. 18.3.3, the active area for each pixel is about $10 \mu\text{m} \times 7 \mu\text{m}$ in area, while the total pixel cell occupies an area of $14 \mu\text{m} \times 11 \mu\text{m}$, giving a chip active area of about $8.8 \text{ mm} \times 6.6 \text{ mm}$. A bubble lens is formed in the window over each of the pixel active areas (Fig. 18.3.2), which concentrates the light into the active region to increase its conversion efficiency.

A typical black and white camera head will contain a single CCD sensor mounted in the focal plane of a standard camera lens system, along with a minimum of interface circuitry. Scan control circuitry,

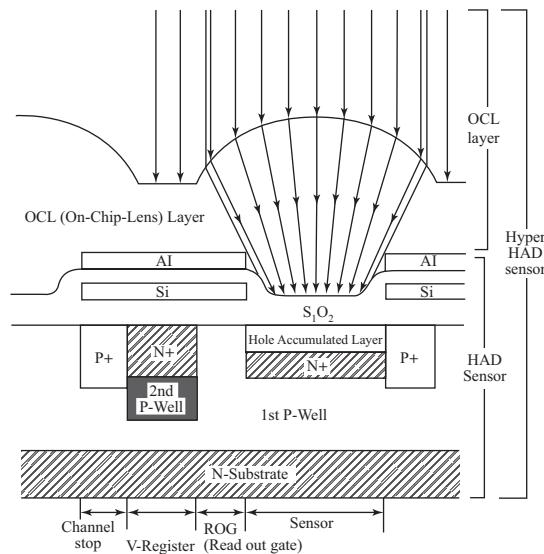


Figure 18.3.2 Section of one pixel of a Sony TV camera sensor chip. (Courtesy of Sony Corporation, Japan.)

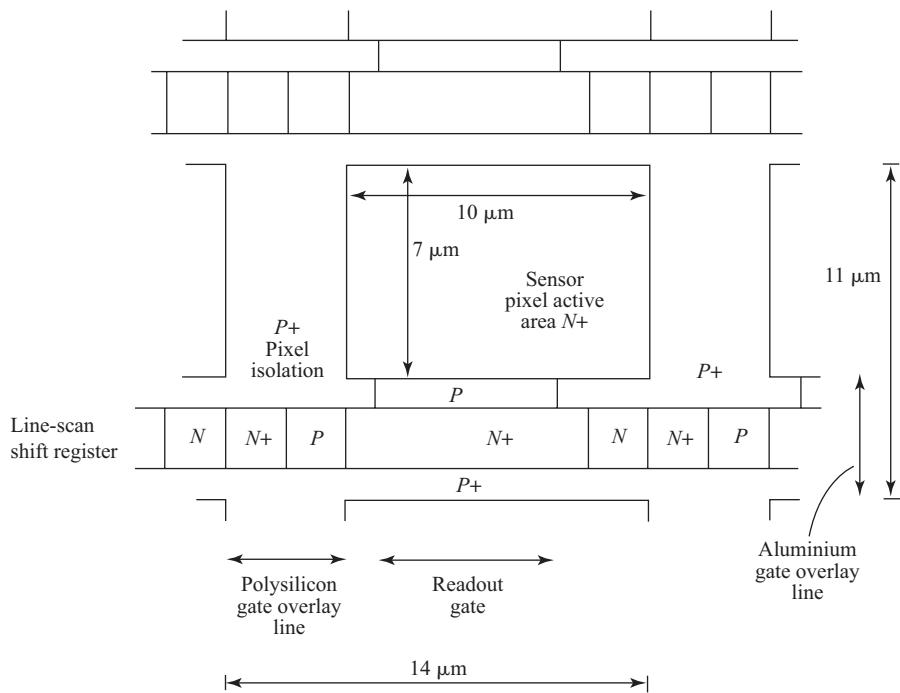


Figure 18.3.3 Floor plan of one pixel of a CCD image sensor array.

synchronizing signal generators, and video and audio modulators are all included in an adjacent camera control unit, which also contains batteries, and in some cases a cassette recorder. A miniature solid-state display produces a local monitor image, which is presented to the operator through an eye piece.

Color cameras are more complex. One popular system uses three separate CCD array sensors. The image from the lens is passed through a color splitter system comprising a prism and special filtering mirrors to separate the image into three, each containing one of the primary color components of the original image. These three images are aligned onto the three sensor arrays, which are scanned synchronously to produce the three color signals. These signals are combined to produce the luminance (I or black and white) signal and the two chrominance components (Q and Y) for transmission.

A special version of the CCD array has been developed for color production in domestic units, in which each pixel is divided into three discrete cells, one for each color. Color separation is obtained by placing an appropriate filtering pigment in the glass lens directly over each pixel segment. Because each pixel requires more area than does a single scanner, the resolution and sensitivity of the resulting device are lower than that of the three-chip camera system. However, the resulting camera is very inexpensive to build since the special color-splitting optics are not required.

The original vacuum tube cameras were large and heavy and used a lot of power. They were only suitable for studio use. The new CCD units do not require large amounts of power for heaters and do not require high voltages for their operation. The result is small, lightweight units that are very portable and produce better results than their vacuum tube equivalents, and at a much lower price.

Television Displays

The CRT (cathode ray tube) has been the traditional means of displaying television pictures. Usually, the picture is directly viewed on the face of the CRT itself. In some systems, the image produced on the face of the CRT is focused through a lens system and projected onto either a back-lit or front-lit screen to produce a larger image than is possible with direct-viewing CRTs. In the last few years much effort has gone into the development of flat, solid-state display panels. While flat panels have been successfully demonstrated, they have not yet been developed much past the laboratory stage, so projection CRTs are likely to be the choice for the new large-screen HDTV systems for some time.

Cathode Ray Tube. Cathode ray tubes were originally developed in the 1930s for displaying radar information and were adapted for displaying television pictures shortly after. In the television receiver the CRT performs the reverse function of the camera tube in the transmitter. It converts the electrical signals containing the picture information into a visible image.

Fig. 18.3.4 shows the basic function of the CRT, with the deflection and focusing mechanism omitted for the moment. An electron gun comprised of a cathode and a series of accelerating and focusing anodes fires a concentrated beam of electrons at the tube screen. The electrical signal representing the picture is applied to the *control electrode* or control grid to modulate the electron beam intensity. The signal causes the potential of the control electrode to vary about a bias level, so that a more negative signal reduces the beam intensity and the brightness of the spot that beam induces in the tube phosphors.

The inside of the CRT screen is coated with a fluorescent material called a *phosphor*, which emits light when it is excited with an electron beam. A thin translucent coating of aluminum behind the phosphor and on the sides of the tube acts as an electron collector or anode. It is maintained at a very high positive dc potential to attract the electron beam. Removal of the electrons from the phosphor increases the brightness produced and prevents dark spots due to charge accumulation. A negative charge accumulation on the tube face would repel the electron beam and reduce the image intensity. The thin aluminum coating also serves to reflect any phosphor light that radiates toward the back of the tube, thus further enhancing the tube brightness.

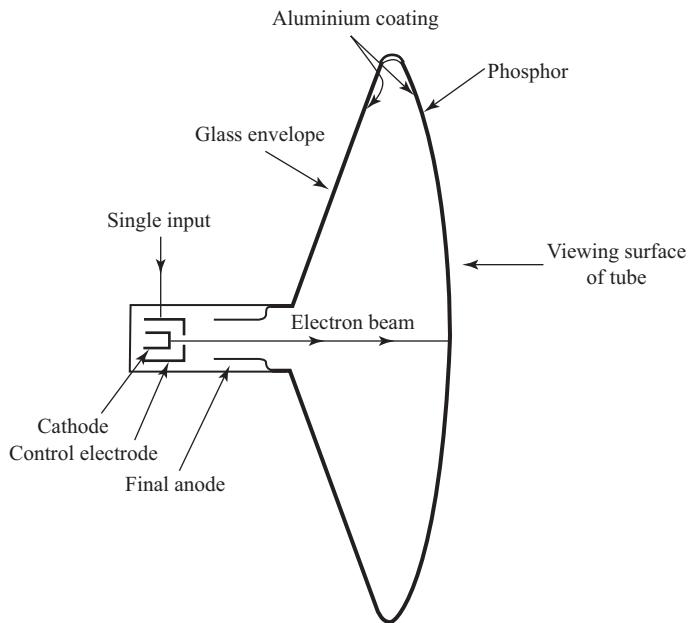


Figure 18.3.4 Cathode ray picture tube structure.

The electron beam must be made to *scan* the face of the tube. Raster scanning similar to that used for facsimile is used. That is, the beam spot must be made to move in a regular repeating pattern over the face of the tube so that every spot on the face of the tube is covered several times a second to give the impression of a continuous image. The phosphor continues to fluoresce for a short time after the electron beam turns off or moves over, so it continues to emit light between scannings. This helps to eliminate flicker and enhance the continuous image impression. However, if the retention continues for more than one or two scan periods, a moving image may blur.

Interlaced raster scanning is usually used in television systems. In this case, each complete picture frame is comprised to two fields, positioned so that the scan lines of one field fall between the scan lines of the other field. Each field is scanned from left to right, from top to bottom, in lines, with a blanked-out retrace between lines. The interlaced scanning reduces flicker by enhancing the image intensity at twice the frame rate. Raster scanning will be discussed further in the next section.

Moving or deflection of the electron beam to create the scanning motion may be achieved either electrostatically or electromagnetically, but electromagnetic deflection permits a shorter tube to be used. Electrostatic deflection is more commonly used in CRTs used for oscilloscope and computer displays.

When the moving electrons in an electron beam pass through a magnetic field, the field exerts a force on them in the same way as in an electric motor. The beam is deflected in a direction that is normal (at right angles) to both the magnetic field direction and to the direction of electron motion, as illustrated in Fig. 18.3.5(a). To obtain the scanning deflection in the tube, an electromagnet is placed around the neck of the tube, as in Fig. 18.3.5(b) or (c), so its field is at right angles to the axis to move the spot horizontally across the face of the tube. A positive magnetic field deflects the beam in one direction, and a negative field deflects it in the other direction. A triangular current waveform applied to the field coil causes the field to start with a large negative value and build to a large positive value, which sweeps the spot across the tube face at the constant fine rate. A second electromagnet at right angles (vertical) sweeps the spot vertically

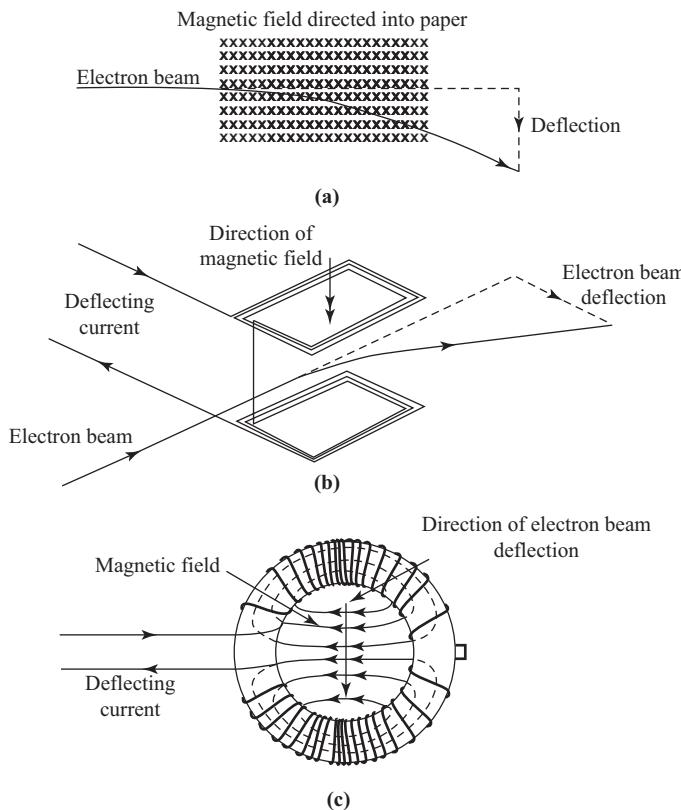


Figure 18.3.5 (a) Electron beam deflected by a magnetic field. (b) Flat horizontal deflection coils. (c) Toroidal vertical deflection coils.

down the face of the tube at the field rate. For practical reasons the horizontal deflection coils are saddle shaped and fitted closely to the neck of the tube, while the vertical deflection coils are of the toroidal shape shown in Fig. 18.3.5(c).

As well as being deflected for scanning, the electron beam must be *focused* to form a sharp spot on the screen. Again, focusing may be accomplished either electrostatically or electromagnetically. Figure 18.3.6(a) shows how a long axially oriented coil can be used for focusing in a camera tube. The coil produces a magnetic field oriented axially along the electron beam before the deflection point. Electrons that are diverging from the axis are caused to spiral back toward the axis and converge onto a focus point some distance down the axis. Adjustment of the field intensity positions the focus point on the target screen. The process is analogous to that of the focusing lens in an optical projector. A short axial cylindrical fixed magnet may also be used to obtain focusing as in Fig. 18.3.6(b), as commonly used for picture tubes. In this case, the magnet is adjusted axially along the neck of the tube until beam convergence is obtained on the CRT face.

Electrostatic focusing is also used in CRTs, especially those used for oscilloscopes. In this case, a double-ring anode is built into the neck of the tube, with a ring cathode placed between them. The field orientation is shown in Fig. 18.3.7. An electron moving along the field axis will not experience any deflection. An electron that is diverging, however, will enter the ring fields and be deflected back toward the axis. The position of the convergence point along the axis is controlled by the potential applied to the ring anode structure and is adjusted so that convergence occurs at the screen surface.

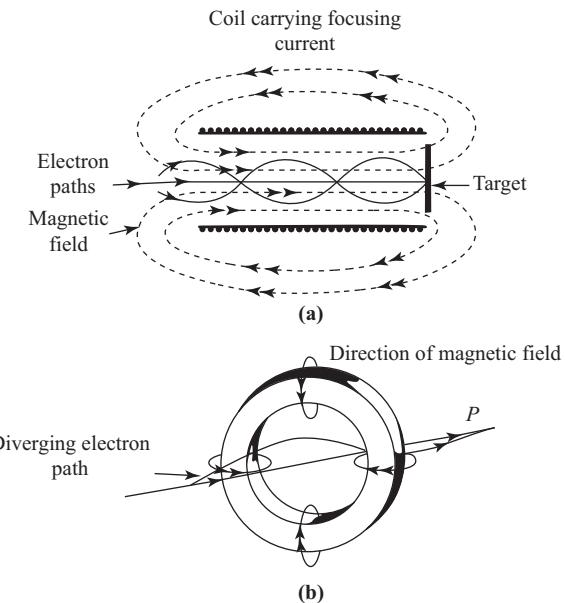


Figure 18.3.6 (a) Solenoid coil focusing. (b) Toroidal coil focusing.

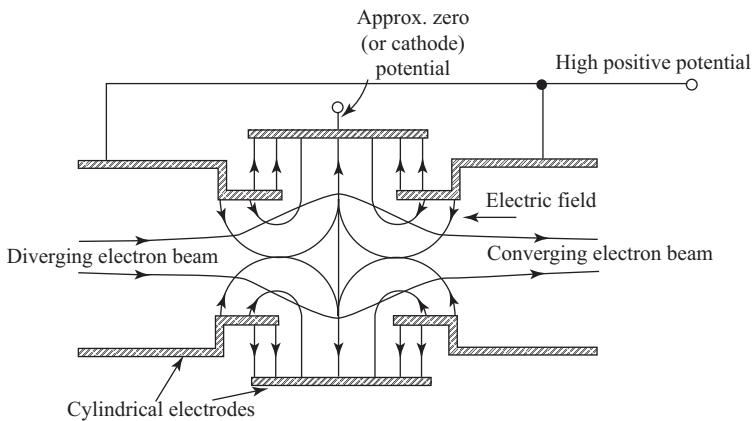


Figure 18.3.7 Electrostatic focusing.

Color Tubes. The first television tubes were monochrome black and white tubes. Color TV was accomplished by using three monochrome cameras, each with a color band-pass filter for one of the three additive primary colors, red, green, and blue. The three signals are independently transmitted to the receiver, where each color signal may be applied to a separate monochrome CRT, which produces the appropriate color. For direct viewing, the three receiving CRTs are combined into one, which has three separate electron beam guns, three separate signal control electrodes, separate focusing, but common deflection circuits. The phosphor target on the back of the CRT face is also a composite, laid out with alternating dots of the three primary color-producing phosphors. By careful focusing and convergence adjustment, each series of color

dots is only scanned by its own electron beam, allowing the three superimposed color pictures to be produced simultaneously and merge on the screen face. These tubes are quite complex and for the larger sizes require very high dc voltage sources, considerably increasing their cost over black and white systems.

Picture tubes are made in a variety of sizes. Size is designated by the *diagonal* measurement across the tube face, with popular sizes being 25, 43, 48, 53, 61, and 74 cm (10, 17, 19, 21, 24, and 26 in.). Recently, the Japanese have developed color picture tubes for HDTV with diagonals in excess of 100 cm (40 in.). Tubes larger than this are unlikely to be practical.

CRT Projection Displays. The physical size of direct viewing CRT displays is limited to about a 100-cm (about 40-in.) diagonal. For situations that require large screen viewing, such as in auditoriums, it is necessary to use a projector system. In this case, three separate CRTs produce three images in white light at high intensity, one for each of the primary colors. Optical color band-pass filters pass only the wanted component of light from each, and the three color images are then focused on a large screen. The screen may be illuminated from the front (usually preferred because it gives greater brightness) or from behind (to give a more compact system). Obviously, these systems are complex and expensive, so they are not commonly found in homes. However, they are likely to be favored for the wide-screen HDTV applications that will occur soon and should become much more popular and less expensive as demand grows.

Flat Panel Displays. Flat panel television displays have always been considered desirable to reduce the bulk and weight of TV receivers. However, to date, only small experimental units have been developed. The first attempt to produce a flat display used a matrix of Light-emitting diodes (LEDs), but this proved to be impractical because of power dissipation and poor resolution.

Two technologies presently are being pursued. The first is plasma discharge tubes, and the Japanese recently planned to demonstrate a 127-cm (50-in.) panel. The second technology is liquid crystal displays, but it may be 10 years before a practical LCD display can be developed. In the meantime, the projection CRT system will remain the favorite for HDTV and the direct-view CRT for NTSC home units.

Interlacing and Vertical Synchronization Frequency

One complete scan of the target area allows reproduction of one complete picture at the receiver of a television system. A large amount of information must be transmitted during this period, and if the picture is repeated at a high rate, then the bandwidth required for transmission becomes excessive. In television, if the picture rate is made too low, moving scenes have a stop-and-go jerky movement in the same way that slow-motion movies do. Furthermore, the tube phosphors have a relatively low persistence, so the picture fades out between scans, producing a flicker at the picture rate. The picture rate must be sufficiently high so that the normal persistence of the viewer's eyes overrides the flicker and merges the picture series into smooth motion. This minimum picture rate has been found to be about 35 to 50 pictures per second.

It has also been found that in television systems, if the scanning rate is near but not exactly equal to the supply frequency, voltage pickup from the ac power circuits modulates the scanning circuit amplifiers and causes distortion and jitter in the picture. This interference can be minimized by making the picture scanning rate a multiple of the supply frequency. In the American (NTSC) system the *frame* or picture rate is 30 Hz, one-half of the 60-Hz supply frequency. The European system uses a 25-Hz frame rate with 50-Hz ac power mains.

The 30-Hz frame rate is too low to eliminate flicker. For this reason each picture frame is divided two separate *fields*, each with half the total number of scan lines. The scan lines of the two fields are made to *interlace* so that the picture area is scanned twice during each picture frame period, as shown in Fig. 18.3.8. In the U.S. system, each frame is scanned with 525 *lines*. During the first field, 262.5 odd-numbered lines are scanned from top to bottom of the picture. At the middle of the 263rd line scan, a retrace moves the scan

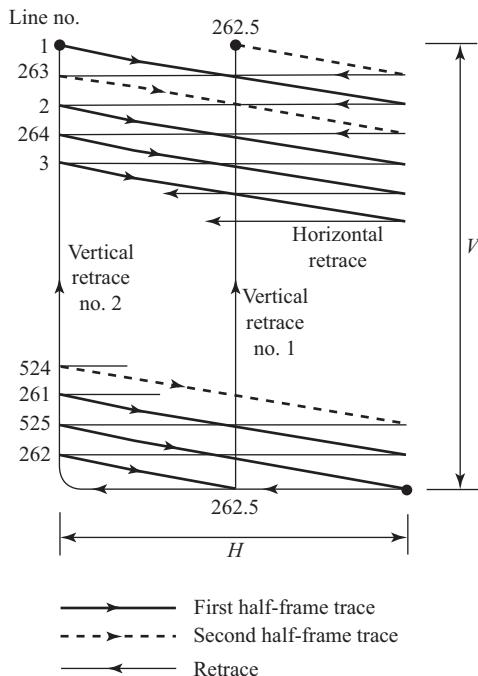


Figure 18.3.8 NTSC television frame 2 : 1 scan interlacing pattern.

spot vertically up the middle of the screen to the top center, and 262.5 even-numbered lines are scanned during the second field period, again from top to bottom. The second field lines fall midway between the first field lines to complete the picture. At the end of the 525th lines, vertical retrace takes the spot from the bottom right to the top left of the screen, ready to begin the next frame. Since each picture frame is being scanned twice during each frame period, the field repetition rate is twice the frame repetition rate. Thus in the U.S. system the frame repetition rate is 60 Hz, which is well above the flicker threshold.

Two vertical retraces take place during each complete picture frame period. For each retrace a vertical synchronizing pulse is produced in the master camera control unit, which is used to initiate the vertical retrace in the camera and in all the receivers at the same time. This ensures that the receivers stay in synchronization with the camera.

Where

$$f_v = \text{vertical synchronization frequency (Hz)} \\ (\text{same as the field repetition rate})$$

$$P = \text{frame (picture) repetition rate (Hz)}$$

$$f_v = 2P \quad (18.3.1)$$

The vertical synchronization pulse is superimposed on top of a longer *blanking* pulse, which is used to turn off the electron beam of the CRT during retrace so the retrace lines cannot be seen. The structure of the combined vertical synchronization and blanking pulses is shown in Fig. 18.4.1(d), which shows that about 20 lines of video information are lost during each vertical retrace, or about 40 lines out of each complete frame.

Picture Definition

Picture definition is determined by the size of the smallest element in a picture, called a *pixel* or *pel* in the same way as for facsimile. Referring to Fig. 18.3.9, let

N = total number of scan line periods per frame period

N_s = number of scan lines suppressed during retrace

N_v = number of active lines

V = vertical dimension of the viewing area of the CRT

w = width of each scan line (or line separation)

Then

$$N_v = N - N_s \quad (18.3.2)$$

and

$$w = V/N_v \quad (18.3.3)$$

Square pixels are used. That is, each pixel has the same width as its height, that is, w . The active viewing area has an *aspect ratio*, which is defined as the ratio of the viewing width to viewing height. Where

α = aspect ratio

H = horizontal viewing dimension

N_h = number of active pixels in a line

Then

$$N_h = N_v \times \frac{H}{V} \quad (18.3.4)$$

and

$$a = \frac{H}{V} = \frac{N_h}{N_v} \quad (18.3.5)$$

Note that 40 lines out of every 525 are blanked out to leave time for the vertical retraces. Similarly, a portion of each line is blanked out to allow time for the horizontal retrace from the right side to the left side of the picture. Fig. 18.4.1(c) shows that 16.5% of the line is blanked, so all of the active pixels are transmitted during the remaining 83.5% of the line period.

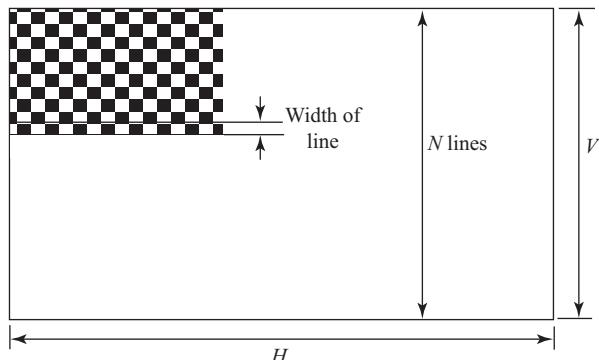


Figure 18.3.9 Checkerboard pattern of alternating black and white pixels and scan parameter definition.

Where

N_L = total number of pixel periods per line period

$$N_L = \frac{N_h}{0.835} \quad (18.3.6)$$

EXAMPLE 18.3.1

In the U.S. NTSC system the aspect ratio is 4/3, the total number of line periods per frame is 525, and the number of suppressed lines is 40 per frame. Find the picture height and width in number of pixels. Also find the number of pixel periods in a line period.

$$N_v = N - N_s = 525 - 40 = \mathbf{485 \text{ lines}}$$

$$N_h = a \times N_v = (4/3) \times 485 = \mathbf{647 \text{ pixels}}$$

$$N_L = \frac{N_h}{0.835} = \frac{647}{0.835} = \mathbf{775 \text{ pixels}}$$

Horizontal Synchronization Frequency

In each picture frame, N lines are scanned, and a horizontal synchronization pulse is produced at the beginning of each line. Frames are scanned at the frame or picture rate P , so the horizontal synchronization pulse frequency is given by

$$f_h = N \times P \quad (18.3.7)$$

and the line scanning time is given by

$$T_h = \frac{1}{f_h} \quad (18.3.8)$$

The synchronization pulse-generating circuits in the camera use a very stable oscillator to generate a primary timing pulse train with a frequency of twice the horizontal synchronizing frequency (31,500 Hz in the U.S. system). This master pulse train is then frequency divided by 2 using digital circuits to produce the horizontal synchronizing pulse train. The master pulse train is also divided by N (the number of lines per frame) to give the vertical synchronizing pulse train.

EXAMPLE 18.3.2

In the U.S. TV system, $N = 525$ lines per frame and $P = 30$ frames per second. Find the horizontal and vertical synchronization frequencies, and the time required to scan one line.

$$f_h = N \times P = 525 \times 30 = \mathbf{15,750 \text{ Hz}}$$

$$f_v = 2 \times P = 2 \times 30 = \mathbf{60 \text{ Hz}}$$

$$T_h = \frac{1}{f_h} = \frac{1}{15,750} = \mathbf{63.5 \mu\text{s}}$$

Video Bandwidth

The bandwidth required for transmitting a video signal is estimated by assuming that adjacent pixels on a scan line are alternating black and white. Under these conditions, two pixels will comprise one cycle of the maximum rate video signal. Scan lines are repeating at the horizontal synchronizing frequency f_h , and each scan line can have a maximum of N_L pixels on it. Based on this, then the highest video frequency that must be passed is given by

$$f = \frac{(f_h \times N_L)}{2} \quad (18.3.9)$$

Experiment has shown that a reduction in picture resolution by a factor of 0.70 (known as the *Kell factor*) can be tolerated, so the highest video frequency that need be transmitted (that is, the video bandwidth) becomes

$$B_v = f_{\max} = 0.70 \times f = 0.35 \times f_h \times N_L \quad (18.3.10)$$

EXAMPLE 18.3.3

Find the video bandwidth required for the U.S. NTSC system.

SOLUTION From the examples above, the horizontal synchronizing frequency $f_h = 15,750$ Hz and the number of pixel periods per line $N_L = 775$. The bandwidth is then

$$B_v = 0.35 \times f_h \times N_L = 0.35 \times 15,750 \times 10^6 \times 775 = \mathbf{4.27 \text{ MHz}}$$

18.4 Television Signal

The NTSC television signal as transmitted is a complex one. Four separate components are included in the signal; (1) sound, (2) picture brightness or luminance, (3) vertical and horizontal synchronization, and (4) color or chrominance information. Each television channel is allotted 6 MHz of bandwidth (see Table 18.3.1) in which the composite video signal is transmitted. Video information is transmitted as an amplitude-modulated vestigial sideband (partially suppressed sideband) carrier located 1.25 MHz from the lower edge of the channel band. All the 4.3-MHz-wide upper sideband and 1 MHz of the lower sideband of the amplitude modulation are transmitted, along with the carrier. The remainder of the lower sideband is filtered out. Figure 18.4.1(a) shows the spectrum distribution of the video signal within the channel band pass.

The audio signal is transmitted as a monaural frequency-modulated carrier located 4.5 MHz above the video carrier, near the upper edge of the channel band pass, and requiring a bandwidth of about 80 kHz about the sound carrier. The sound carrier is located well above the video band-pass and can be removed easily from the video signal by filtering.

The color chrominance information is double-sideband suppressed-carrier modulated on two quadrature-related color subcarriers located at 3.58 MHz above the video carrier and added to the video modulation. Only the sidebands are transmitted, so two additional channels are provided for transmitting the color information. The luminance or Y signal is directly amplitude modulated on the video carrier, while the two color components I and Q are modulated on the two phase-shifted color channels. The Q signal requires a sideband channel width of ± 0.5 MHz, while the I signal uses a lower sideband of 1.5 MHz and a vestigial upper sideband of 0.6 MHz. While these two color signals are included within the same band pass as the luminance signal, interference between them is minimal because they are so closely related.

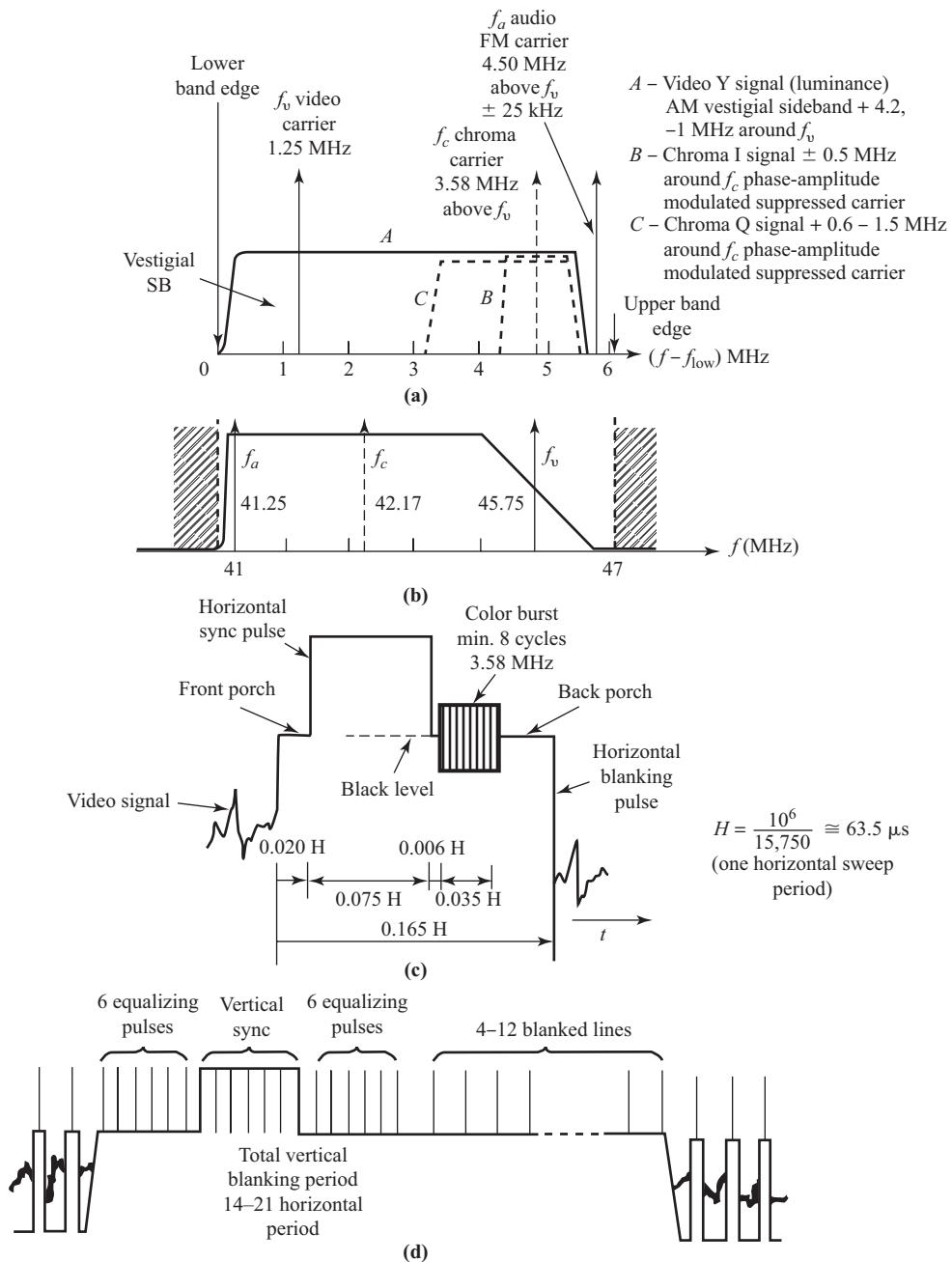


Figure 18.4.1 (a) Signal spectrum within the 6-MHz channel assignment. (b) Video IF band-pass characteristic. (c) Horizontal sync pulse structure. (d) Vertical sync pulse structure.

Synchronization information is carried on the video carrier during retrace periods between lines and between fields. During the horizontal retrace time (16.5% of each line period), the video signal is raised to the black level to *blank* out the picture during retrace. A horizontal synchronizing pulse 7.5% of the horizontal

period wide is added to the blanking level, as shown in Fig. 18.4.1(c), which is used to trigger the horizontal oscillator in the receiver and force it to synchronize to the signal. A burst of 8 cycles of the 3.58-MHz chroma carrier is also transmitted during the horizontal blanking period, which is used to lock the phase of the receiver chroma oscillator to that of the transmitter to allow proper demodulation of the color signals.

Vertical retrace takes longer, and a period of about 20 lines on each field is blanked out. During this period, as shown in Fig. 18.4.1(d), six horizontal sync pulses are transmitted at twice the rate to stabilize the synchronization. During the next three horizontal periods, the signal is raised to the synchronizing level, and inverted horizontal sync pulses are sent at twice the rate (total 6) to form the actual vertical synchronizing pulse. Following this, six more equalizing pulses at twice the rate are sent, and then the normal horizontal synchronizing pulses resume.

18.5 Television Receivers

Black and white television receivers were developed before color receivers, and the color system was designed to be an “add-on” that doesn’t affect the performance of any black and white receivers still in use.

Black and White Receivers

Figure 18.5.1 shows the functional block diagram of a typical black and white television receiver. Starting from the antenna or CATV cable input terminals, the first section is the tuner assembly. The input signal is coupled into an RF amplifier stage (a class A broadband small signal amplifier) whose output is fed to a mixer—oscillator circuit. Usually two separate tuners are built into the same receiver, one for the UHF bands (channels 14 to 83) and one for the VHF bands (channels 2 to 13).

In the older sets, tuning is accomplished for the VHF bands by means of a turret switch that places a different set of coils and capacitors in the circuit for each channel. A small variable capacitor allows fine tuning. In the UHF tuner, variable capacitance tuning is used, with a large variable capacitance to select channels and a smaller one for fine tuning. In later sets, varactor tuning has become commonplace, with a wafer switch to select tuning voltages to apply to the varactor in the oscillator circuit and a potentiometer for fine tuning trimming.

The intermediate frequency (IF) has been standardized to lie in the 41-to 47-MHz band, with the frequency spectrum inverted (using the difference signal from the mixer, with the oscillator frequency above the channel frequency). This spectrum is shown in Fig. 18.4.1(b), with the video carrier at 45.75 MHz and the audio carrier at 41.25 MHz. The IF filters are designed to pass this band, with the response curve sloped between 44.5 and 47 MHz to compensate for the vestigial sideband in the detector. The IF string is typically three to five stages of narrow-band RF amplifiers coupled through bandpass filters to produce the desired band-pass characteristic. One or more of these amplifiers have AGC applied, as does the RF amplifier in the tuner. Video detection is accomplished using a simple envelope detector at the output of the IF string.

The detected video signal is fed to the input of the audio IF amplifier string, where tuned circuits isolate the audio carrier at 4.5 MHz, with a band pass of about ± 40 kHz about the carrier. Audio limiting and detection are done by any of a number of standard FM detector circuits, such as the ratio detector. A volume control and one or more audio amplifier stages drive a speaker.

The video signal is also passed to a video amplifier, which raises the level, filters out the sound carrier, and produces the control bias for the picture tube electron beam gun. This bias modulates the electron beam intensity and thus the light intensity on the face of the CRT.

A sample of the video signal is also passed to the synchronization circuits, where a clamp circuit removes the video portion of the signal to leave only the sync pulses. These sync pulses are differentiated and used to trigger the horizontal sweep oscillator circuit. The linear ramp voltage produced by the oscillator is

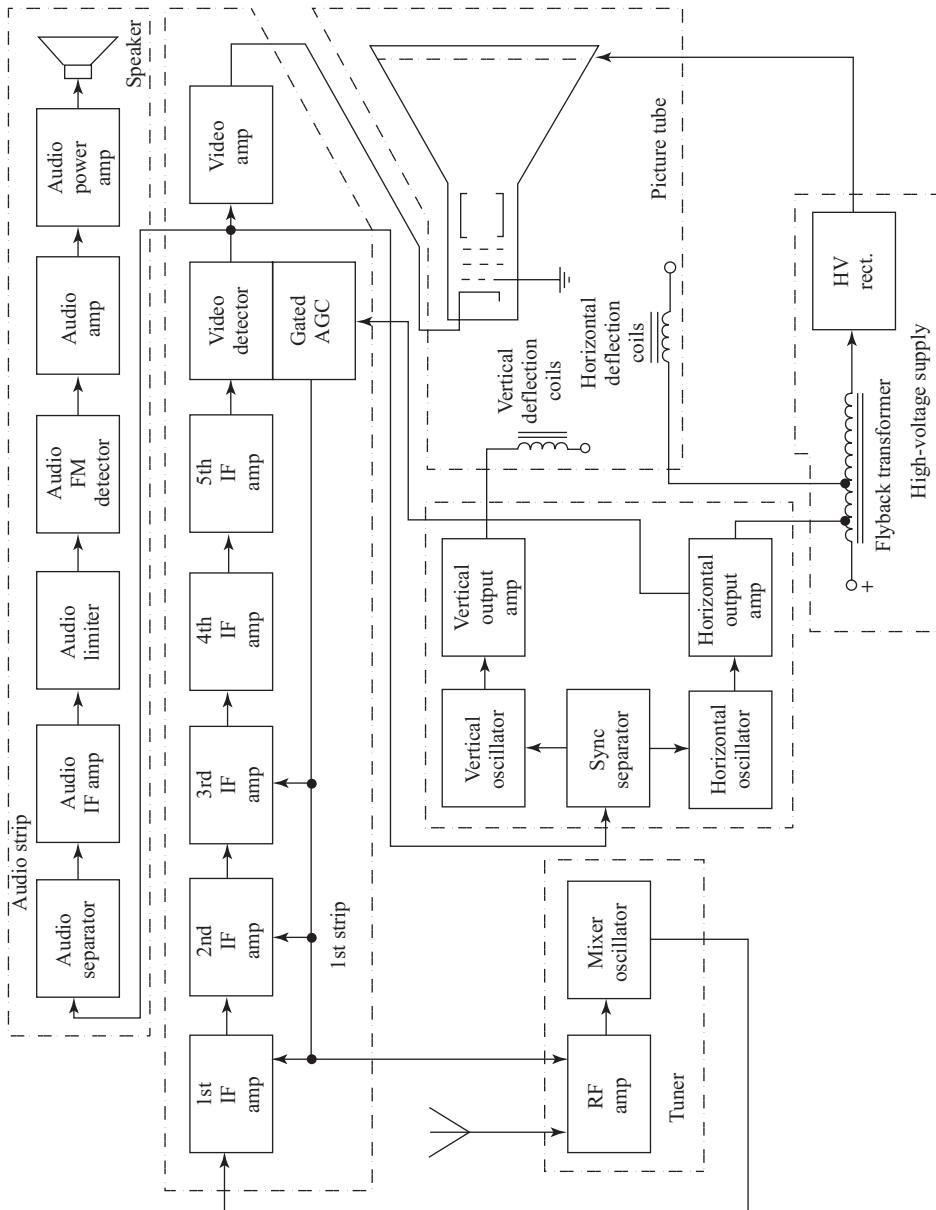


Figure 18.5.1 Black-and-white television receiver block diagram.

amplified to drive the horizontal output transformer and the horizontal deflection coils. The large pulse produced by the flyback transformer during the retrace period is rectified to produce the very high dc voltage (10 to 20 kV) for the picture tube target and also to provide the gating signal to control the AGC circuit. The clamped sync pulses are also integrated to extract the vertical sync pulse and then used to trigger the vertical oscillator to produce the vertical sweep voltage. This is amplified and applied to the vertical deflection coils.

Color Receivers

The color television receiver is more complicated than the black and white receiver because of the necessity to decode the color signals from the video signal before it is applied to the picture tube. However, many of the circuits in both sets serve the same function and have the same specifications. These include the tuners, the IF amplifier strips, the deflection circuits, and the audio circuits. The video amplifier has different characteristics, the picture tube has three separate electron beam guns, and additional circuits are required to decode the color information and control the convergence of the three electron beams in the tube. Figure 18.5.2 shows those circuits in a color set that differ from those in a black and white set.

The detected video signal is passed through a video preamp, which raises the level to compensate for the four-way split that follows. The preamp has a 4.2-MHz bandwidth and a *trap* (notch filter) to remove the sound carrier at 4.5 MHz. The first path is to the input of the synchronization separator circuit and deflection circuits, which function in exactly the same manner as those in a black and white set.

The second path is through a delay line, which compensates for the delays in the chroma demodulator circuits, and then to the luminance (Y) video amplifier. This amplifier has a full 4.2-MHz bandwidth to provide the video signal to the picture tube during black and white reception. This signal also serves to produce most of the detail in color pictures.

The third path for the video signal is to the chroma amplifier, which has a band pass of +0.6/-1.5 MHz about the 3.58-MHz color carrier frequency. The output from the chroma amplifier feeds the I and Q signal inputs to the color demodulators.

The fourth path is to the color burst pickoff circuit. It responds to the 3.58-MHz bursts present on the horizontal synchronizing pulses and during black and white reception turns off the I and Q circuits through the color killer circuit to prevent color blooming. The color burst pickoff also provides the phase synchronization signal for the 3.58-MHz color local oscillator to lock onto. Thus the color demodulators start off locked onto the signal at the beginning of each scan line.

The I and Q demodulators extract the two color signals and present them, along with the Y signal, to the inputs of the color decode matrix. This circuit restores the three color signals, red, green, and blue, by means of weighted sums of the Y, I and Q signals.

$$\begin{aligned} R &= +0.62Q + 0.95I + Y \\ G &= -0.64Q - 0.28I + Y \\ B &= +1.73Q - 1.11I + Y \end{aligned} \quad (18.5.1)$$

The three color signals (R, G, and B) each control a separate electron beam in the picture tube. Special convergence controls steer the three beams so that they only illuminate the desired phosphors on the screen, which are laid out in an alternating dot matrix, to recreate the color image.

The picture tube power supplies are complicated by the need for carefully regulated voltages to keep the three beams properly converged. The high target supply voltage (20 to 24 kV) *must* be carefully regulated. If it becomes too high, either the picture tube or the shunt regulator tube may produce harmful x-rays. This regulator and the high-voltage rectifier tubes are carefully shielded to prevent radiation or flashover in dust accumulations.

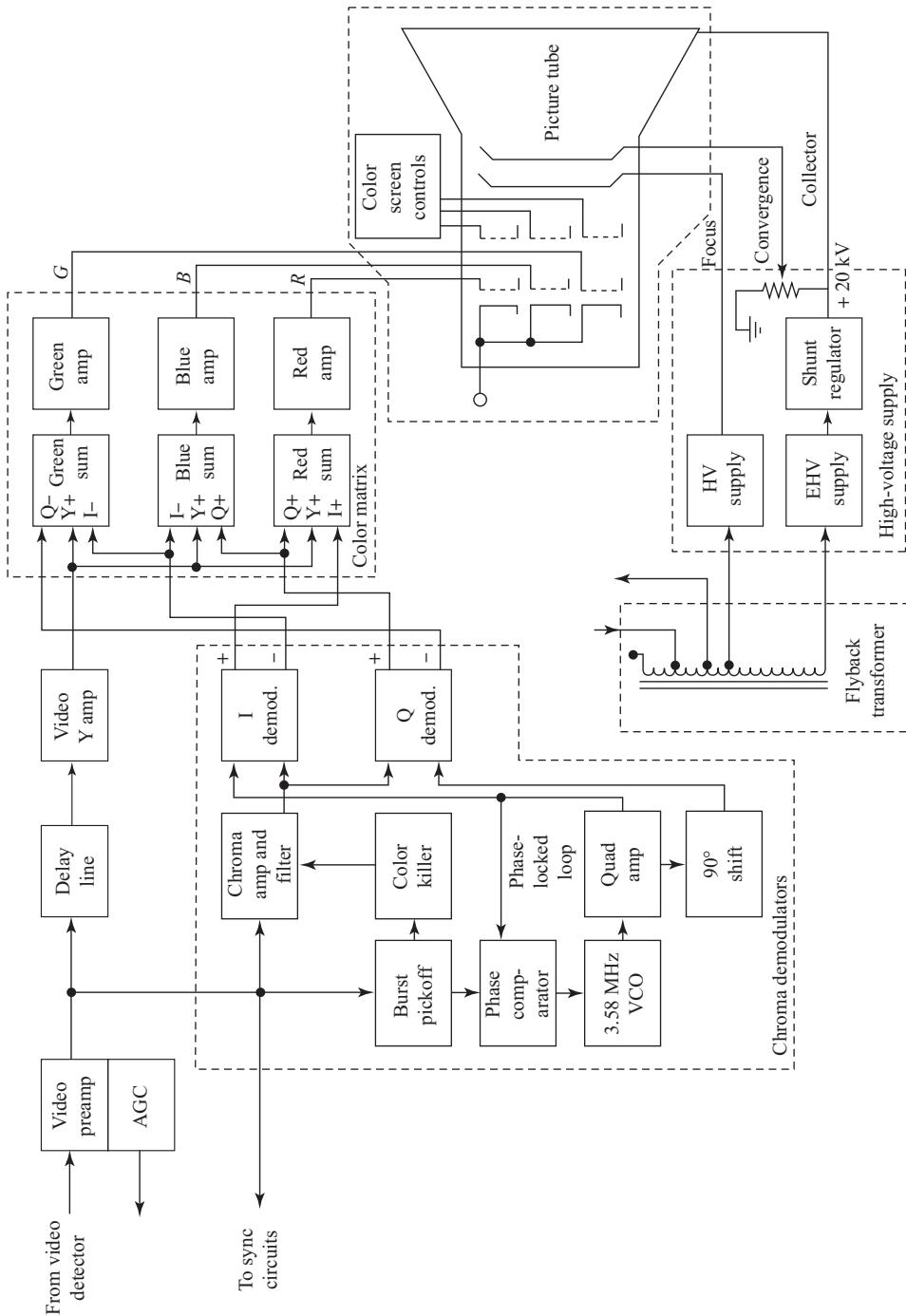


Figure 18.5.2 Color television receiver block diagram, showing those parts of the receiver that differ from those of a black and white set.

18.6 Television Transmitters

It is unlikely that any commercial black and white only transmitters still exist. Color transmission has been commonplace since the mid 1950s and most distributors of television signals find it necessary to be up to date. In some cases, black and white signals may be transmitted, as would be the case when viewing old black and white movies. However, these signals can be generated easily by turning off the color modulators in a color transmission system. Therefore, only a full color transmitter will be discussed here.

Color Transmitters

In this discussion, for simplicity, it is assumed that each of the three primary colors (R, G, and B) is recorded by a separate camera image sensor. In some inexpensive camera systems, a special image sensor that combines the functions of the three color cameras into one is used, although with lower quality than the three-camera system.

Figure 18.6.1 shows the functional block diagram of a simplified system for transmitting an NTSC color TV signal. A common optical system focuses the image and splits it into three beams. These beams pass through color absorption filters to leave only one of the three primary colors (R, G, or B) in each beam. These three color beams are then focused onto three image sensor arrays. When these arrays are simultaneously scanned, they produce three separate video signals, one for each of the primary colors (R, G, or B).

The three color signals (R, G, and B) are applied as inputs to the color combining matrix (which performs the reverse function of that in the receiver). The color signals are combined in weighted sums to produce the luminance (Y) and the two chrominance (I and Q) signals as follows.

$$\begin{aligned} Y &= +0.30R + 0.59G + 0.11B \\ I &= +0.60R - 0.28G - 0.32B \\ Q &= +0.21 - 0.52G + 0.31B \end{aligned} \quad (18.6.1)$$

The luminance or Y signal contains all the information required to reconstruct a black and white picture from the signal. It is passed through a low-pass filter to limit its bandwidth to 4.2 MHz. The I signal is limited to a bandwidth of 1.5 MHz, and the Q signal is limited even further to 0.5 MHz. The bandwidth limiting is used to reduce interference among the three signals.

A 3.58-MHz crystal-controlled oscillator generates the subcarrier for the chroma signals. The system reference phase is specified as 0° and the transmitted carrier burst for synchronization lags the reference phase by 180° . The Q signal carrier phase leads the reference phase by 33° , and the I signal carrier leads the Q phase by 90° , giving a quadrature relationship between the two. These phase relations are illustrated in Fig. 18.6.2.

In the camera shown in Fig. 18.6.1, the color burst is gated directly from the chroma crystal oscillator, which is leading reference phase by 180° . The oscillator signal is also passed through a phase shifting network to create a lag of 147° relative to the oscillator signal to create the Q phase signal. A further phase shifter produces a 90° lag from the Q signal to give the I phase signal. Two balanced modulators modulate the I and Q signals on their respective subcarriers in a double-sideband suppressed carrier mode. The color burst gate allows a short burst (about 8 cycles) of the chroma carrier to be transmitted during each horizontal blanking period to allow phase synchronization of the carrier reinsertion oscillator at each receiver.

The signal from the chroma oscillator is also fed to a binary frequency divider chain, which generates the 15,750-Hz horizontal and 60-Hz vertical synchronizing signals and the horizontal and vertical blanking pulses. These signals control the scanning in the camera system and are also transmitted to control the scanning at the receiver.

Audio signals are produced by one or more microphones used with the camera and are mixed to form either a single monaural or a multiplexed stereo sound channel associated with the picture. A frequency

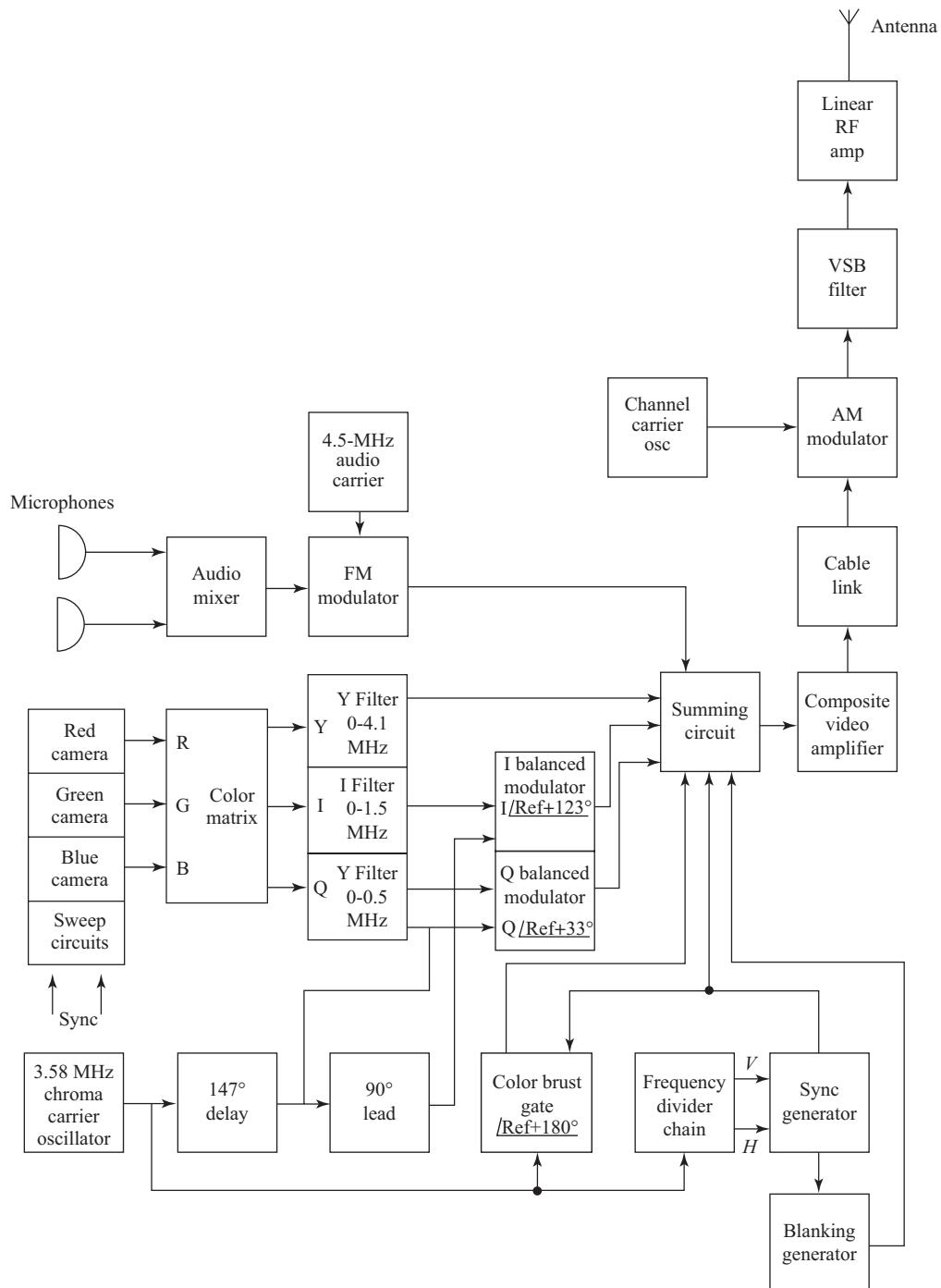


Figure 18.6.1 NTSC color television transmitting system.

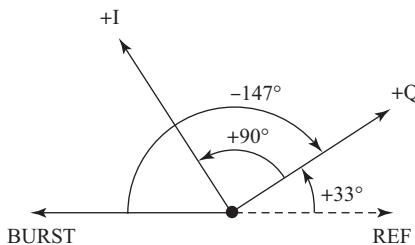


Figure 18.6.2 Phase relationships of the NTSC color carrier signals.

modulator circuit modulates this channel on the output from a 4.5-MHz crystal-controlled oscillator with a maximum modulation index of about 2 and a bandwidth of about 80 kHz. This FM carrier may also have an additional SCSA channel multiplexed onto it for special services that do not interfere with the basic television sound.

A summing circuit combines the components of the video baseband signal. These are the Y signal, the I and Q modulated signals, the sound carrier, the color burst, the blanking signals, and the horizontal and vertical sync pulses. The FM sound carrier is also added in at this point. This video baseband signal contains frequency components from dc to about 4.6 MHz, so transmission in this form for more than a few feet is not practical. For transmission to the broadcast transmitter site (of cable head or satellite link head), the video baseband signal is vestigial sideband modulated on an intermediate frequency (IF) such as the channel 3 video carrier at 61.25 MHz. The output from the modulator circuit is passed through a band-pass filter to remove all but 1 MHz of the lower sideband and pass the 1-MHz vestigial lower sideband, carrier, and full upper sideband. The resultant is amplified and transmitted over a coaxial cable to the transmitter site or is modulated onto a point-to-point microwave link or satellite link connected into a distribution network or a CATV system.

At the transmitter site, the IF signal is picked off the incoming cable, amplified, and applied to the input of a frequency converter. This converter combines the IF with the transmitter oscillator signal and moves it into the desired broadcast channel spectrum slot, with the proper sideband orientation for transmission (lower sideband vestigial). A chain of tuned linear power amplifiers raises the signal level to the transmitter output level (10 to 200 kW) for application to the antenna. Typical transmitting antennas are multiple dipole vertical arrays oriented to provide omnidirectional coverage in a main lobe parallel to the earth's surface. This gives good coverage over an area of perhaps 80-km radius.

18.7 High-definition Television

There has lately been much interest in the possibility of upgrading the existing color TV systems to provide high-definition television (HDTV). Much work has been done on the technical problems of providing HDTV, and international action is presently underway to establish a set of universal standards for HDTV, which will allow development of a universally accepted system.

The Japanese have led the way in this research, and in 1989 they demonstrated a prototype version of their system and planned to launch a direct broadcast satellite in 1991 to provide broadcast facilities for HDTV. European interests in 1989 began a joint research project called Eureka 95, with the aim of having an HDTV system on air in 1995. The Americans have only just started to think about the problem in a serious way, and they have taken the view that any new system must be compatible with existing NTSC receivers, a restriction that artificially limits the possibilities of HDTV.

The International Telecommunications Union is presently trying to establish a series of standards to which HDTV systems can adhere to provide a semiuniversal compatibility. The Japanese system at the moment appears to be the most likely basis for these standards, since they were the first to go on air. Some aspects of HDTV are presented next to provide an insight into the future.

Definition and Aspect Ratio for HDTV

It was shown above how the picture definition in pixel density is related to the signal and scan characteristics. But this only gives definition in terms of the display on the picture tube. However, the size of the picture tube and the distance of the viewer from the screen also affect the apparent resolution seen by the viewer.

To illustrate this, suppose we have a standard NTSC system feeding a receiver with a 48.25-cm (19-in.) diagonal picture tube. If the viewer is seated about 2.3 m (7 ft) away from the screen, he or she will not see any details of the scan lines on the screen. However, if the viewer moves closer, he or she will see some of the line details. If the viewer moves farther from the screen, he or she will have the sensation of improved definition, although the picture will appear to be smaller because it is farther away. Again, if the same viewer sits at 2.3 m (7 ft) from a 76-cm (30-in.) diagonal screen, he or she will again see line details, so for a good picture the viewer must move farther from the screen. Alternatively, if the system were changed to "high definition" so that the pixel density available were higher, then the viewer could move closer or could use a larger screen without degradation of the picture.

It has been shown that the limit of definition for the human eye is such that it can discern an object that is large enough and close enough to subtend a viewing angle of about 1 minute of arc. It has also been shown that the viewer will have a feeling of being "in the picture" if the picture width sub-tends an angle greater than about 30°.

Figure 18.7.1 illustrates how the screen dimensions, viewing distance, and viewing angle are related. H and V are the viewing area horizontal and vertical dimensions as defined above. Where

D = diagonal dimension of the viewing area

X = distance from screen to viewer

θ = angle subtended by the horizontal dimension H

$\alpha = H/V$ = the aspect ratio

by Pythagoras

$$D^2 = H^2 + V^2 = H^2 \left(1 + \frac{1}{\alpha^2} \right) \quad (18.7.1)$$

and

$$\tan\left(\frac{\theta}{2}\right) = \frac{H}{2X} \quad (18.7.2)$$

EXAMPLE 18.7.1

The NTSC system presents 485 active lines with 647 active pixels each in a picture (see Example 18.3.1), with an aspect ratio of 4 : 3. A receiver uses a CRT tube with a viewing area diagonal of 48.26 cm (19 in.). Find the minimum distance a viewer should sit from the tube to eliminate seeing any scan lines and the viewing angle subtended by the screen.

SOLUTION: The horizontal width of the viewing area H is, from Eq. (18.7.1),

$$H = \sqrt{\frac{a^2 \cdot D^2}{1 + a^2}} = \sqrt{\frac{1.333^2 \cdot 48.26^2}{1 + 1.333^2}} = 38.6 \text{ cm}$$

There are 647 active pixels in each line, so the width of one pixel is

$$w = \frac{H}{N_h} = \frac{38.6}{647} = 0.060 \text{ cm}$$

Each pixel must subtend 1 minute of arc, so the total viewing angle must be

$$\theta = N_h \times \frac{1 \text{ minute}}{60} = 647 \times \frac{1}{60} = 10.783^\circ$$

Now the minimum viewing distance X can be found as

$$X = \frac{H}{2 \cdot \tan(\theta/2)} = \frac{0.3861}{2 \cdot \tan[10.783 \cdot (1/2)]} = 2.05 \text{ m (6.7 ft)}$$

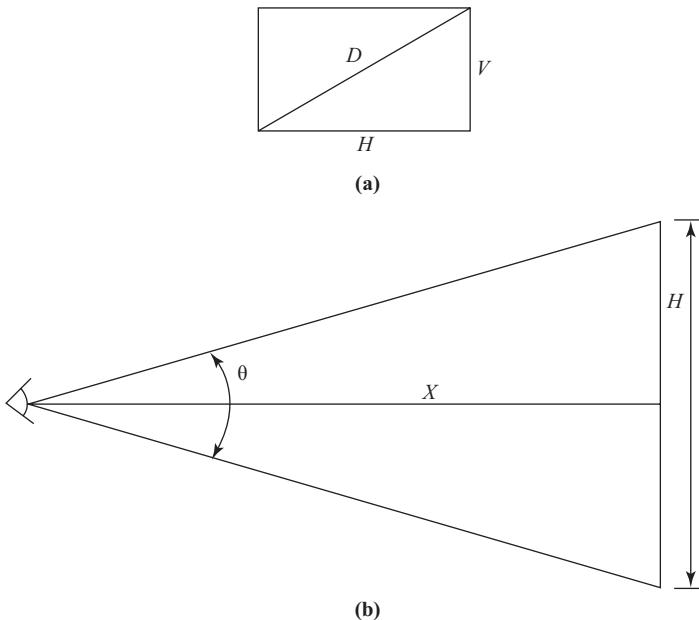


Figure 18.7.1 (a) Television screen viewing area dimensions. (b) Relationship of viewing angle to distance from screen.

The aspect ratio of the existing normal TV systems is 4 : 3. As this example illustrates, this aspect ratio is large enough to allow viewing at a comfortable distance from existing CRT displays, but it is not large enough to provide that “in the picture” feeling. It has been agreed by most people that an aspect ratio of 16 : 9 is most likely to be standard for HDTV. Although development work on flat panel large TV displays is progressing, they are not likely to be available for many years yet. This means that HDTV will in most cases use large screen projection systems for some time.

The larger aspect ratio means that the viewing area is increased by 4 : 3 over the old system, with a corresponding increase in the number of pixels per frame that must be transmitted. A further increase in

EXAMPLE 18.7.2

Repeat Example 18.7.1 for the parameters proposed for the HDTV system given above. Assume that a projector screen with a diagonal dimension of 1.40 m (55 in.) and an aspect ratio of 16 : 9 is used with 1125 vertical lines (90 suppressed).

SOLUTION The horizontal width of the viewing area is

$$H = \sqrt{\frac{a^2 \cdot D^2}{1 + a^2}} = \sqrt{\frac{1.778^2 \cdot 1.40^2}{1 + 1.778^2}} = 1.22 \text{ m}$$

Ninety lines are suppressed during each frame for vertical sync, leaving 1035 active lines per frame. Assuming square pixels, this gives 1840 active pixels per line. Allowing 1 minute of arc for each pixel gives the total active viewing angle as

$$\theta = N_h \times \frac{1 \text{ minute}}{60} = 1840 \times \frac{1}{60} = 30.7^\circ$$

$$X = \frac{H}{2 \cdot \tan(\theta/2)} = \frac{1.22}{2 \cdot \tan[30.7 \cdot (1/2)]} = 2.23 \text{ m (6.8 ft)}$$

Note that the viewing distance is about the same as it was in Example 18.7.1, but now the viewing angle is three times larger and the viewing area is much larger for the same degree of definition.

definition is obtained by increasing the number of vertical scan lines per frame from the present 525 (or 625 in Europe) to most likely 1125. This results in a further increase of about 2.1 : 1 in the number of pixels required, for a total increase of about 2.86 : 1. This increased pixel density means that the viewing distance from the screen can be reduced considerably, as illustrated in Example 18.7.2.

Transmission Bandwidth Requirements for HDTV

The Japanese multiple sub-Nyquist sample encoding (MUSE) system uses 1125 lines per frame, with two interlaced fields per frame, at a line frequency of 33,750 lines per second. The vertical synchronization pulse requires suppression of 90 lines, leaving 1035 active lines. With an aspect ratio of 16 : 9, this gives 1840 active pixels per line. Horizontal synchronization requires suppression of 12.7% of each line, which means that there are 2108 pixel periods in each line period. Allowing a 15% reduction of bandwidth for the Kell factor, the minimum video bandwidth required is given by a modification of Eq. (18.3.10).

$$B_v = f_{\max} = 0.85 \times f = 0.425 \times f_h \times N_L \quad (18.7.3)$$

For the numbers given, this yields a bandwidth requirement of about 30 MHz for the Japanese HDTV system, which is more than five times the bandwidth required for the NTSC system. It is proposed that this bandwidth be used directly for studio use, recording, and studio satellite links transmissions.

Because of the restricted spectrum space available, it is not feasible to use this wide bandwidth channel for direct broadcast of HDTV, and some form of bandwidth compression is required. Obviously, any bandwidth compression scheme results in a compromise that lowers the overall definition of the system.

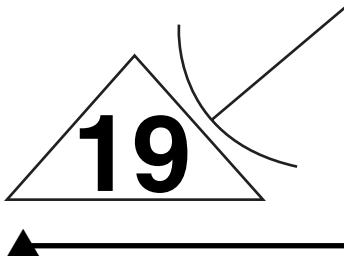
However, it has been observed that the human eye observes maximum definition of an image if that image is not moving and observes a reduced definition if the image is moving, with further reduction as speed of motion increases. The Japanese MUSE system uses this fact as the basis for a bandwidth compression system that reduces the transmission bandwidth to about 7.5 MHz, or about twice that required for an NTSC channel.

Except for the digital sampling and bandwidth compression techniques used, much of the other circuitry used in the HDTV system will be similar to that in the NTSC system.

PROBLEMS

- 18.1. Discuss and compare television and facsimile telegraphy as methods of transmitting documentary information such as photographs.
- 18.2. Discuss one form of facsimile transmitter and explain how the scanning spot is obtained. Explain also how the basic signal may be modulated during the scanning process.
- 18.3. Explain the term *raster scanning* as applied to facsimile.
- 18.4. Explain what is meant by synchronization and why this is necessary in picture transmission. Show, with the aid of a sketch, the form of distortion that occurs in a facsimile link if the receiver speed is greater than the transmitter speed.
- 18.5. Explain what is meant by index of cooperation and why this is important in facsimile telegraphy. Show also how it affects the bandwidth of the signal. If the index of cooperation of a facsimile transmitter is 352, what must be the index of cooperation of the corresponding receiver?
- 18.6. The index of cooperation of a facsimile machine is 352, its speed of rotation is 60 rpm, and the scanning pitch is $\frac{5}{16}$ mm. Find the theoretical bandwidth required for the transmission channel.
- 18.7. A fax transmission with an IOC(IEEE) = 900 is received by a drum scanner that uses 8-in.-wide paper (the scanning surface). Find (a) the drum diameter in millimeters, (b) the scan pitch, and (c) the scan line density.
- 18.8. The drum receiver in Problem 18.7 is 14 in. long and turns at 120 rpm. How long does it take to receive the document page?
- 18.9. A flat bed scanner scans 1200 lines for a 28 cm \times 21.6 cm page. Document transmission takes 140 s. Find (a) the IOC(IEE) and IOC(CCITT), (b) the scan density, (c) the scan rate, and (d) the transmission bandwidth.
- 18.10. Explain what is meant by subcarrier frequency modulation. A facsimile signal is transmitted on an SCFM system whose carrier frequency is 1800 Hz. The SCFM signal is then amplitude modulated on an RF carrier. Assuming only first-order sidebands need be considered in each modulation step, estimate the bandwidth required at the RF carrier frequency if the highest frequency in the basic signal is 550 Hz.
- 18.11. FCC specifications for a facsimile drum scanner are line length = 18.85 in. and scan density = 96 lines/in. Find the IEEE and CCITT indexes of cooperation. For a scan speed of 90 lines per minute, find the bandwidth required.
- 18.12. An electronic scanner produces 1.11 million pixels per document page and uses a run length code with Huffman coding to produce a 35 to 1 average data compression. If the maximum transmission bandwidth is 1 kHz, find the document transmission time.
- 18.13. Explain the purpose of the aluminum coating used on the inside of CRT television tubes. Why is this coating held at the final anode potential?

- 18.14.** Explain how scanning is achieved in a television system. What is the purpose of the synchronization signals? In the U.S. 525-line twin interlace system, how many field synchronizing signals occur during each frame period?
- 18.15.** Describe and compare magnetic and electric methods of focusing an electron beam in a cathode ray tube.
- 18.16.** In the British television system the number of lines is 625, the number of pictures transmitted per second is 25, and the aspect ratio is 4/3. The number of lines suppressed during the field blanking period is 48, and twin interlace is used. The line blanking period is 16% of line duration. Find the video bandwidth required for transmission.
- 18.17.** What are the vertical and horizontal oscillator frequencies for the British TV system in Problem 18.16?
- 18.18.** An experimental TV system uses twin interlace scanning and has Kell factor = 0.8, video bandwidth = 40 MHz, horizontal deflection frequency = 45 kHz, vertical deflection frequency = 50 Hz, vertical blanking of 50 lines, and horizontal blanking of 12% of the period. Assuming square pixels, find the picture aspect ratio.
- 18.19.** An arena uses an NTSC projection system on a large screen, producing an image containing 650×490 pixels. The nearest seat is 30 m from the screen. Find the width, height, and diagonal dimensions of the screen.
- 18.20.** Outline the differences between a black and white television signal and an NTSC color signal. Use a sketch of the waveforms to illustrate, and explain the reason for each component. Also sketch the signal frequency spectra.
- 18.21.** Discuss the major differences between the proposed HDTV system and the NTSC system as they affect screen size and shape, viewing distance, and the impact on the viewer.
- 18.22.** An HDTV system uses dual interlaced scanning with 1125 lines and a horizontal scan rate of 33750 lines/s. The aspect ratio is 16/9, 90 lines are suppressed for vertical synchronization, and 12.7% of each line is suppressed for horizontal synchronization. A Kell factor of 85% applies. Find the transmission bandwidth required for the uncompressed analog video signal.
- 18.23.** Find the total number of pixels required for a frame in the HDTV system of Problem 18.22 and compare that to the number for the NTSC system.
- 18.24.** Compute the time required to send a page as a facsimile if the data circuit has a speed of 19.2 kbps and the number of bytes in the scanned image is 320 kB.
- 18.25.** An image file of size 4 MB is to be transmitted over a digital link with a speed of 19.2 kbps. If a compression mechanism is employed before the image is transmitted, which would reduce the file size to 78%, calculate the time required to complete the transmission.
- 18.26.** An electronic scanner produces 1.12 mega pixels per document page and uses a run length code with Huffman coding to produce a 40 to 1 average data compression. If the maximum transmission bandwidth is 1 kHz, find the document transmission time.
- 18.27.** Derive the *bandwidth* of monochrome video. Deduce that of *color video* from it.
- 18.28.** Derive the transmission bandwidth requirements for HDTV.



Satellite Communications

19.1 Introduction

A communications satellite is a spacecraft that carries aboard communications equipment, enabling a communications link to be established between distant points. An all-embracing definition of a spacecraft would include deep-space probes such as the Voyager series, but in this chapter only those satellites that orbit the earth will be considered. Satellites that orbit the earth do so as a result of the balance between centrifugal and gravitational forces. Johannes Kepler (1571–1630) discovered the laws that govern satellite motion. Although Kepler was investigating the motion of planets and their moons (so-called heavenly bodies), the same laws apply to the artificial satellites launched for communications purposes. Before examining the role these satellites play in telecommunications, a brief introduction to Kepler's laws will be presented as they apply to such satellites. Kepler's laws apply to any two bodies in space that interact through gravitation. The more massive of the bodies is called the *primary* and the other the *secondary* or *satellite*.

19.2 Kepler's First Law

Kepler's first law states that the satellite will follow an *elliptical path* in its orbit around the primary body. An ellipse has two focal points (or *foci*) shown as F_1 and F_2 in Fig. 19.2.1.

The center of mass of the two-body system, termed the *barycenter*, is always centered on one of the foci. In our specific case, because of the enormous difference between the masses of the earth and the satellite, the center of mass always coincides with the center of the earth, which is therefore at one of the foci. This is an important point because the geometric properties of the ellipse are normally made with reference to one of the foci, which can be selected to be the one centered in the earth.

The *semimajor axis* is shown as a and the *semiminor axis* as b in Fig. 19.2.1. The *eccentricity* of the ellipse is defined as

$$e = \frac{\sqrt{a^2 - b^2}}{a} \quad (19.2.1)$$

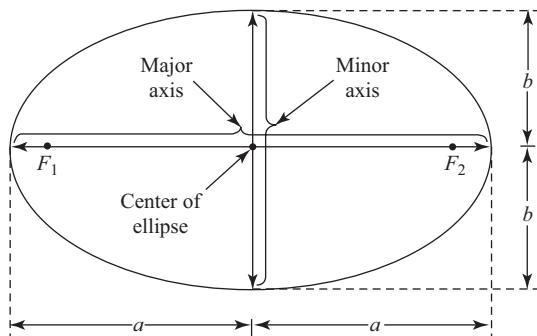


Figure 19.2.1 Foci F_1 and F_2 , the semimajor axis a and the semi-minor axis b of an ellipse.

As will be shown shortly, the semimajor axis a and the eccentricity e are two significant parameters in the study of the orbits followed by communications satellites.

19.3 Kepler's Second Law

Kepler's second law states that for equal time intervals the satellite sweeps out equal areas in the orbital plane, focused at the barycenter. Referring to Fig. 19.3.1, assuming that the satellite travels distances S_1 and S_2 meters in 1 s, the areas A_1 and A_2 will be equal. The average velocities are S_1 and S_2 m/s. Because of the equal area law, it is obvious that distance S_1 is greater than distance S_2 and hence the velocity S_1 is greater than velocity S_2 . Generalizing, it can be said that the velocity will be greatest at the point of closest approach to the earth (termed the *perigee*) and will be least at the farthest point from the earth (termed the *apogee*). This also has fundamental significance in the selection of orbits for communication satellites, as will be shortly shown.

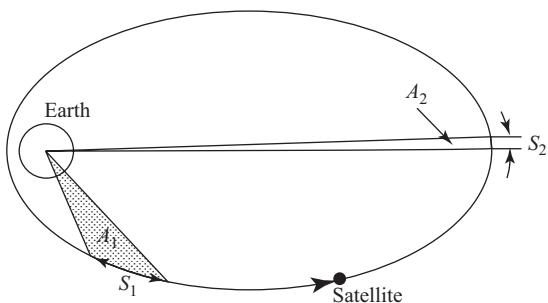


Figure 19.3.1 Kepler's second law.

19.4 Kepler's Third Law

Kepler's third law states that the square of the periodic time of orbit is proportional to the cube of the mean distance between the two bodies. The mean distance as used by Kepler can be shown to be equal to the semimajor axis, and the third law can be stated in mathematical form as

$$a = AP_o^{2/3} \quad (19.4.1)$$

where A is a constant. With a in kilometers and P_o in mean solar days, the constant A for the earth evaluates to

$$A = 42241.0979 \quad (19.4.2)$$

These equations apply for the ideal case of a satellite orbiting a perfectly spherical earth, with no disturbing forces. In reality, the earth's equatorial bulge and external disturbing forces will result in deviations in the satellite motion from the ideal. Fortunately, the major deviations can be calculated and allowed for. Satellites that orbit close to the earth (coming within several hundred kilometers) will be affected by atmospheric drag and by the earth's magnetic field. For the more distant satellites, the main disturbing forces are the gravitational fields of the sun and the moon.

19.5 Orbits

Although an infinite number of orbits are possible, only a very limited number of these are of use for satellite communications. Some of the terms used in describing an orbit are

Apogee. The point farthest from the earth.

Perigee. The point of closest approach to the earth.

Ascending node. The point where the orbit crosses the equatorial plane going from south to north.

Inclination. The angle from the earth's equatorial plane to the orbital plane measured counterclockwise at the ascending node.

Figure 19.5.1 shows three orbits. The *polar orbiting* satellite follows an orbit that is close to the earth and passes over, or very close to, the poles; that is, the inclination is close to 90° . The average height of these orbits is typically 800 to 1000 km above the earth, and they are used mainly for earth observation and surveillance (weather, pollution monitoring, and the like), and for search and rescue work. More recently, trials have been conducted using small satellites for data communications and position determination (ORBICOMM System), which may provide low-cost services in these areas.

The *inclined highly elliptical orbit* is used where communications is desired to regions of high latitude. Kepler's second law shows that the orbital velocity is least at the apogee, and hence by placing the apogee above the high latitude regions the satellite remains visible for a longer period from these regions. The Russian *Molniya* series of satellites use highly inclined orbits. One effect of the earth's equatorial bulge is to cause the orbit to rotate, such that apogee and perigee move around the earth, this being referred to as *rotation of the line of apsides*. However, at one particular value of inclination, $i = 63.4^\circ$, the rotation of the line of apsides is zero, and satellites that are required to have the apogee remain fixed over a particular region are launched into orbits with this value of inclination. These orbits are referred to as being in the 63° slot.

A recently introduced service that uses near-circular, non-geostationary orbits is the *Global Positioning Satellite* (or GPS) service, which is essentially a navigation and position determination service. The GPS system utilizes 6 orbits with 4 satellites in each. The ascending modes of the orbits are separated by 60° and the inclination of each orbit is 55° .

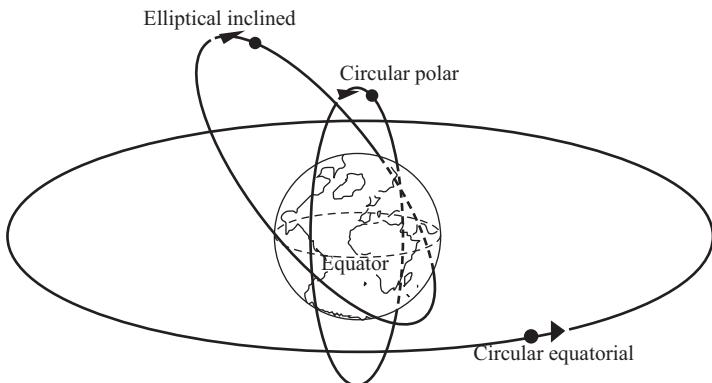


Figure 19.5.1 Three possible orbits. (From *Telecommunications Satellites*, Kenneth W. Gatland, Ed. Englewood Cliffs, NJ: Prentice Hall, 1984.)

19.6 Geostationary Orbit

A *geostationary* satellite is one that appears to be stationary relative to the earth. There is only one geostationary orbit, but this is occupied by a large number of satellites. It is the most widely used orbit by far, for the very practical reason that earth station antennas do not need to track geostationary satellites (except for certain very high gain earth station antennas that require a limited range of tracking, as will be described later).

The first and obvious requirement for a geostationary satellite is that it must have zero inclination. Any other inclination would carry the satellite over some range of latitudes and hence would not be geostationary. Thus the geostationary orbit must be in the earth's equatorial plane. The second obvious requirement is that geostationary satellites should travel eastward at the same rotational velocity as the earth. Since this velocity is constant, then from Kepler's second law it can be deduced that the orbit must be circular, since as previously shown the velocity in an elliptical orbit varies from a maximum at perigee to a minimum at apogee and hence is not constant.

The earth makes one complete rotation, relative to the fixed stars, in approximately 23 h 56 m. Notice that this is slightly less than the time required for one complete rotation about its own axis, which is 24 h. Substituting $P_o = 23\text{ h }56\text{ m}$ in Eq. (19.4.1) for Kepler's third law, along with the value for A given in Eq. (19.4.2), results in

$$a_{\text{gso}} = 42,164 \text{ km} \quad (19.6.1)$$

The subscript gso is included to remind us that this is the value for the geostationary orbit. It will be recalled that because the orbit is circular this is also the radius of the orbit measured from the center of the earth. The earth's equatorial radius is approximately 6378 km, and hence the height of the geostationary orbit above the earth is

$$\begin{aligned} h &= 42,164 - 6378 \\ &= 35,786 \text{ km} \end{aligned} \quad (19.6.2)$$

This value is often rounded up to 36,000 km for use in calculations. It will be seen that there is only the one value of a that satisfies Kepler's third law for the periodic time of 23h 56m, and hence there can only be one geostationary orbit.

19.7 Power Systems

A satellite stays in orbit essentially as a result of natural forces and in the absence of external disturbances would orbit the earth indefinitely without having to carry fuel for propulsion. In practice, disturbance torques and forces exist, as described in the following sections. As a result of these disturbances, satellites must carry fuel on board so that corrective forces can be applied from time to time, usually through thruster jets. The need to carry fuel imposes one of the major limitations on the useful life of a satellite.

In addition, the satellite must receive energy to power the electronic equipment on board. This is invariably supplied by solar cells. With cylindrically shaped satellites, these are arranged around the body of the satellite, as shown in Fig. 19.7.1(a). The advantage of the cylindrical arrangement is that the satellite can be set spinning to maintain its position through the gyroscopic effect, but with this arrangement only about one-third of the satellite body is illuminated by the sun at any given time, and so the power available is limited. As an example, the Intelsat VI satellite employs the cylindrical arrangement that is designed to provide at least 2 kW throughout the expected 10-year life of the satellite.

An alternative arrangement is to employ solar sails, as shown in Fig. 19.7.1(b). With this type of construction, spin stabilization cannot be used and other methods are discussed in the next section. The orientation of the solar sails can be adjusted automatically for maximum solar illumination, so high power outputs can be obtained. For example, the European *Olympus* satellite employs solar sails that are capable of generating 7 kW throughout the 10-year projected lifetime of the satellite.

For a period of about 45 days around each equinox, the satellite is eclipsed by the earth, the eclipse lasting for a maximum period of around 70 min at its peak during each eclipse. Battery backup supplies must be provided during these periods, and long-life batteries have been especially developed for this purpose.

19.8 Attitude Control

By *attitude* is meant the satellite's orientation in space. Attitude control is necessary to keep the directional antennas aboard the satellite pointing to desired regions of the earth. The antennas will also have specific *footprints* to maximize the coverage of certain areas, and, again, attitude control is necessary in order to maintain the proper orientation and positioning of the footprint. A satellite's attitude can be altered along one or more of three axes, termed the *roll*, *pitch*, and *yaw* axes. These are illustrated in Fig. 19.8.1.

Geostationary satellites are stabilized in one of two ways. *Spin stabilization* can be utilized with satellites that are cylindrical. The satellite is set spinning with the spin axis parallel to the N–S axis of the earth, as shown in Fig. 19.8.2. Spin rates are typically in the range from 50 to 100 rpm. Since the antennas are oriented to point to fixed regions on earth, the antenna platform must be “despun” at the same rate as the satellite spins.

In the absence of disturbance torques, the spinning satellite would maintain its correct attitude relative to the earth. Disturbance torques, notably those produced by the gravitational fields of the sun and the moon, can alter the satellite's attitude. Also, movement aboard the satellite, for example, that experienced by redirecting antennas, and bearing friction can decrease the spin rate. Corrections must be applied periodically using impulse thrusters or jets.

Spin stabilization is obviously not possible where solar sails are used. In this case, stabilization is achieved through the use of momentum wheels inside the satellite. A number of arrangements are in use, one of which is shown in Fig. 19.8.3. Here the satellite is stabilized through the gyroscopic effect of the spinning wheels.

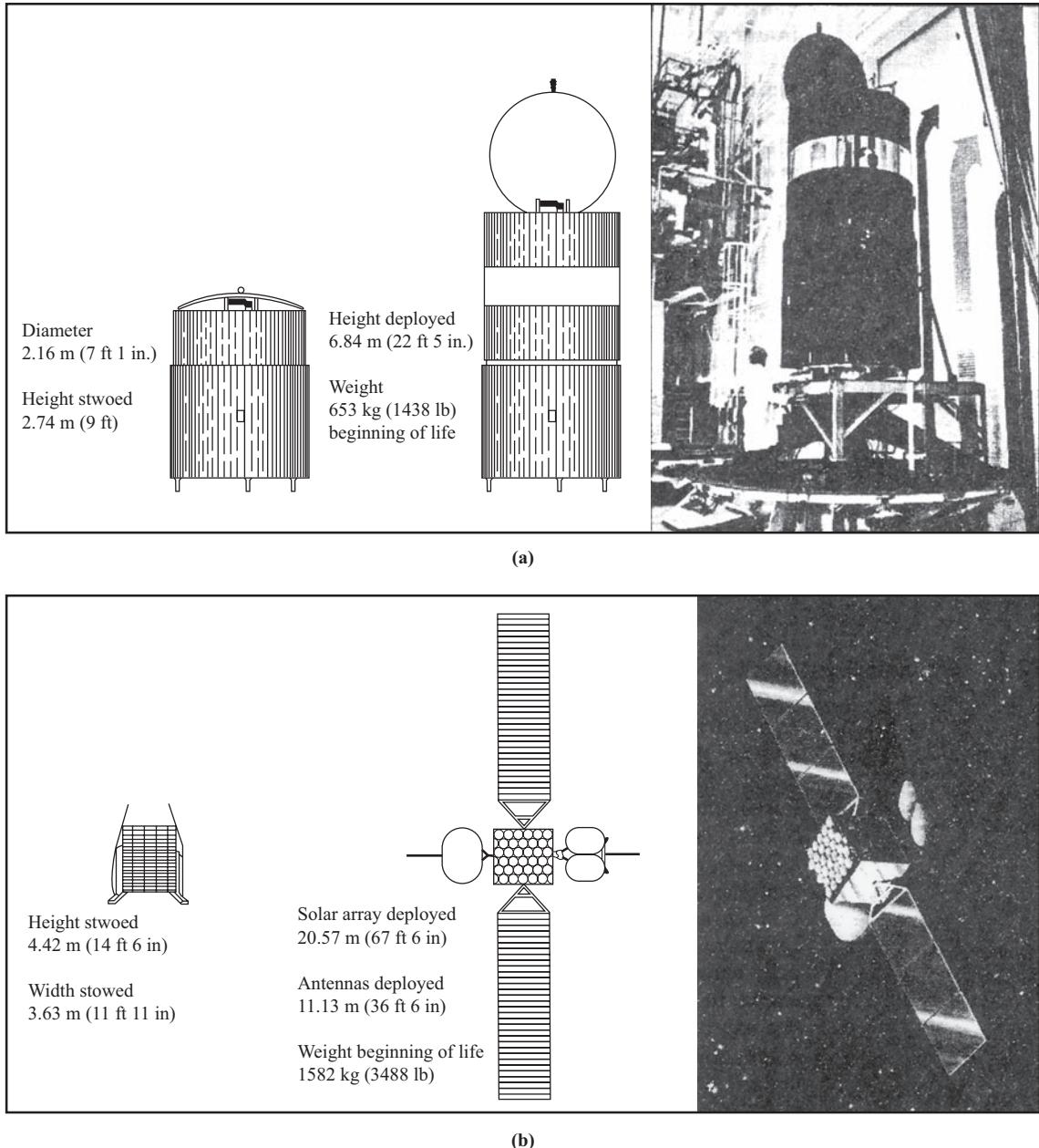


Figure 19.7.1 Solar cell arrangements: (a) cylindrical body and (b) solar sails. (Courtesy Hughes Aircraft Co. Space and Communications Group.)

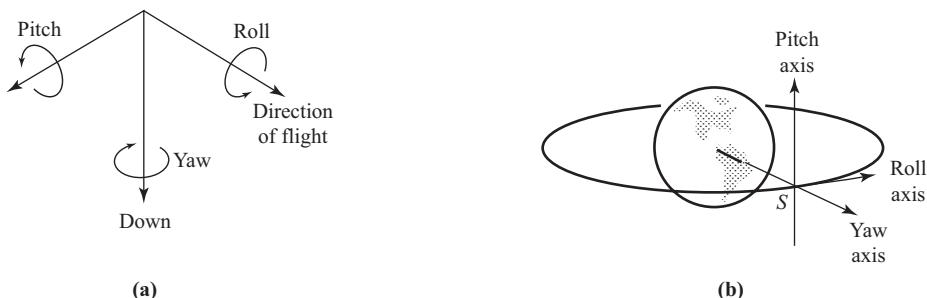


Figure 19.8.1 (a) Roll, pitch, and yaw (RPY) axes. The yaw axis is directed toward the earth's center, the pitch axis is normal to the orbital plane, and the roll axis is perpendicular to the other two. (b) The RPY axes for a geostationary satellite. Here the roll axis is tangential to the orbit and lies along the satellite velocity vector.

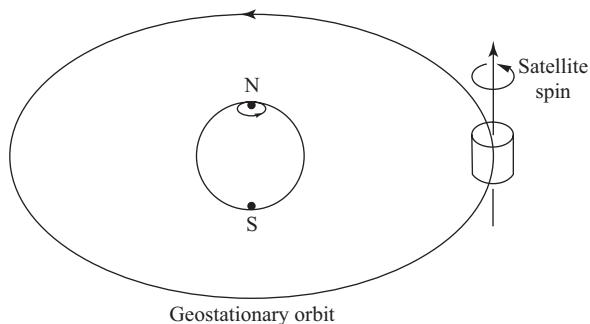


Figure 19.8.2 Spin stabilization in the geostationary orbit.

19.9 Satellite Station Keeping

Left to itself, a geostationary satellite would in fact drift from its initial position as a result of perturbing forces. The gravitational field of the moon, and to a lesser extent that of the sun, causes a drift in the angle of inclination, which amounts to about $0.85^\circ/\text{year}$. The drift is cyclic, the angle of inclination increasing from zero to a maximum of 14.67° in a period of about 26.6 years, thereafter drifting back to zero inclination again in about the same period. For satellites operating in the C band (6/4 GHz), the drift must be kept within $\pm 0.1^\circ$, and for Ku band (14/12 GHz) satellites, within $\pm 0.05^\circ$, so that north-south station keeping maneuvers are required. These are carried out by means of thruster jets once every few weeks. The extra weight added by the fuel needed for the north-south corrections is a major factor in the cost of the launch.

Kepler's laws apply for bodies that are perfectly spherical. The earth departs from a true spherical shape, a flattening occurring at the poles, and the equatorial circumference is not quite circular. Overall, the nonsphericity of the earth results in geostationary satellites drifting eastward toward one of two gravitational nodes separated by 180° . These are located at longitudes 105°W and 75°E and are sometimes referred to as satellite graveyards because satellites that are out of commission tend to drift toward these as their "final resting place." The longitudinal tolerance is also $\pm 0.1^\circ$ for C band and $\pm 0.05^\circ$ for Ku band satellites, so that

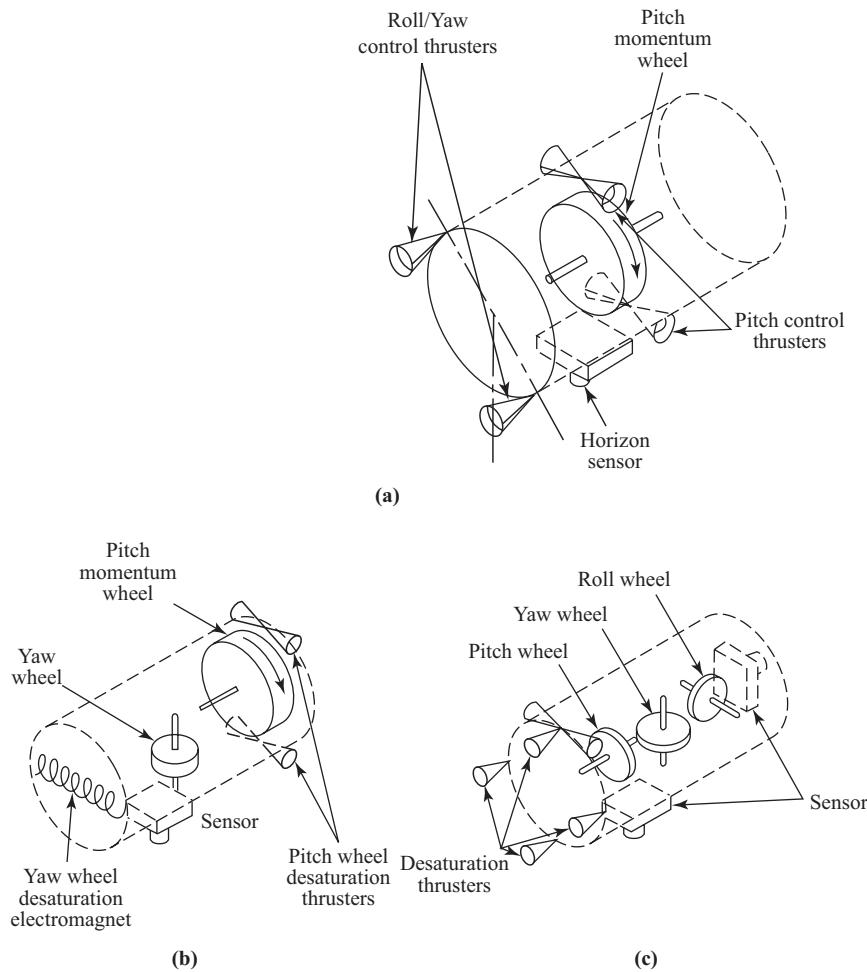


Figure 19.8.3 Momentum wheel stabilization. (Reprinted with permission from *Spacecraft Attitude Determination and Control*. Edited by James A. Wertz. Copyright © 1984 by D. Reidel Publishing Company, Dordrecht, Holland).

east–west station keeping maneuvers are required in addition to the north–south maneuvers. These are also carried out once every few weeks, but require considerably less fuel than the north–south maneuvers. Typical satellite motion is shown in Fig. 19.9.1.

19.10 Antenna Look Angles

To maximize transmission and reception, the direction of maximum gain of the earth station antenna, referred to as the antenna *bore sight*, must point directly at the satellite. To align the antenna in this way, two angles must be known. These are the *azimuth*, or angle measured from the true north, and the *elevation*, or angle measured up from the local horizontal plane, as shown in Fig. 19.10.1.

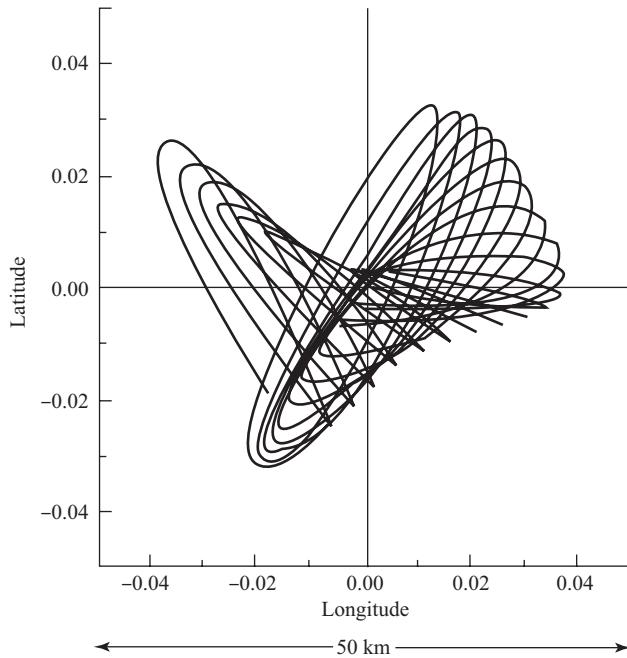


Figure 19.9.1 Typical satellite motion. (From Telesat Canada, 1983, courtesy Telesat Canada.)

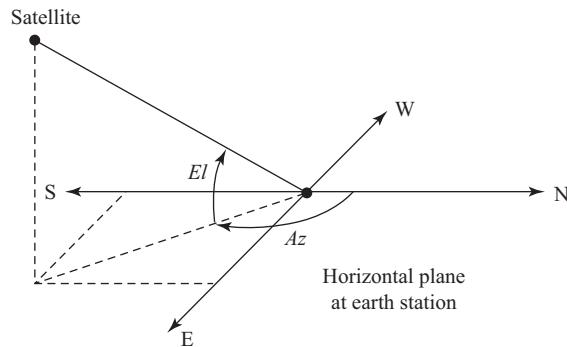


Figure 19.10.1 Angles of azimuth Az , and elevation El , measured with reference to the local horizontal plane and true north.

The azimuth and elevation angles are usually referred to as the *look angles*. In addition to the look angles, it is often necessary to know the range or distance from the earth station to the satellite. The earth's constants needed in these calculations are

$$\text{Mean radius: } R = 6378 \text{ km} \quad (19.10.1)$$

$$\text{Radius of geostationary orbit: } a_{\text{gso}} = 42164 \text{ km} \quad (19.10.2)$$

In addition to these constants, the other pieces of information needed to determine the look angles and the range are

Satellite longitude, ϕ_S

Earth station longitude, ϕ_E

Earth station latitude, λ_E

The conventions used in the calculations are that east longitudes are positive numbers and west longitudes are negative numbers (measured from the Greenwich meridian). Latitudes are positive measured north and negative measured south from the equator. Certain rules known as *Napier's rules*, which apply to spherical trigonometry, must be used in these calculations. Figure 19.10.2(a) shows the situation. SS is the subsatellite point (which must lie on the equator for a geostationary satellite), and ES is the earth station, which for clarity is shown in the southern hemisphere. A property of spherical triangles is that all the dimensions including the sides are in angular measure. Angle a is measured from the north pole to the subsatellite point, and since the subsatellite is on the equator, $a = 90^\circ$. Because one of the sides is a right angle, the spherical triangle is referred to as a *quadantal* spherical triangle.

Angle B is the difference in longitude between the earth station and sub-satellite longitudes. Keeping in mind the sign conventions referred to above, angle B is given by

$$B = \phi_E - \phi_S \quad (19.10.3)$$

Also keeping in mind that southern latitudes are assigned negative values, the angle c is given by

$$c = 90^\circ - \lambda_E \quad (19.10.4)$$

For example, if $\lambda_E = 30^\circ\text{S}$, then $c = 120^\circ$

Knowing angles B and c , angle A can be found by the application of certain of Napier's rules. For the quadrantal triangle these result in A being obtained from

$$\tan A = \frac{-\tan |B|}{\sin \lambda_E} \quad (19.10.5)$$

The azimuth can be determined once angle A is known. Four situations must be considered; these are shown in Fig. 19.10.3. For these situations, the azimuth is given by

$$\text{Figure 19.10.3(a): } \lambda_E < 0 \text{ and } B < 0, \quad \text{Az} = A \quad (19.10.6)$$

$$\text{Figure 19.10.3(b): } \lambda_E < 0 \text{ and } B > 0, \quad \text{Az} = 360^\circ - A \quad (19.10.7)$$

$$\text{Figure 19.10.3(c): } \lambda_E > 0 \text{ and } B < 0, \quad \text{Az} = 180^\circ + A \quad (19.10.8)$$

$$\text{Figure 19.10.3(d): } \lambda_E > 0 \text{ and } B > 0, \quad \text{Az} = 180^\circ - A \quad (19.10.9)$$

These equations do not take into account the special case when the earth station is on the equator, and determining the look angles for this situation is left as an exercise for the reader.

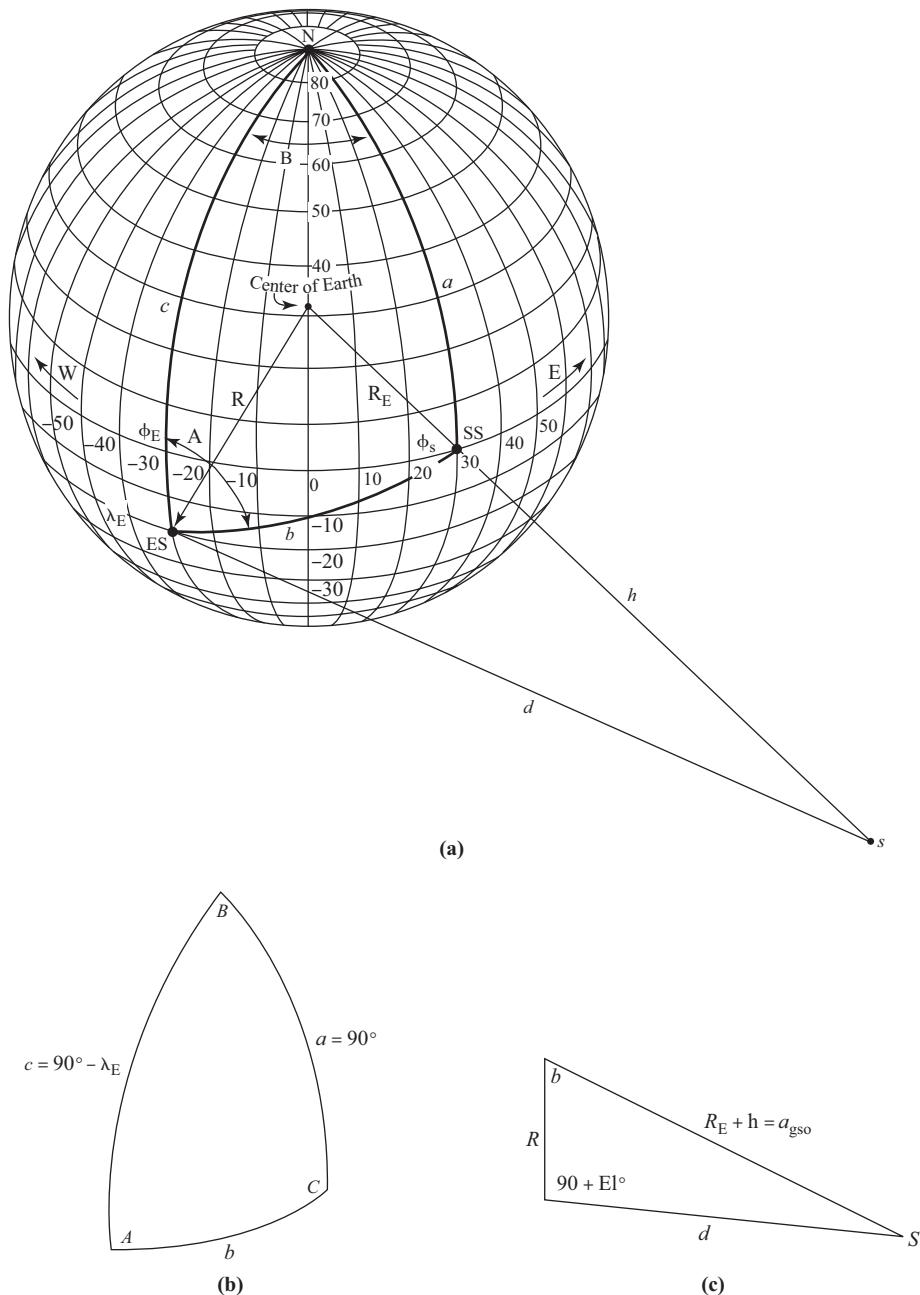


Figure 19.10.2 (a) Geometry used to calculate look angles and range for a geostationary satellite at S . (b) Spherical quadrantal triangle obtained from (a). (c) Plane triangle obtained from (a).

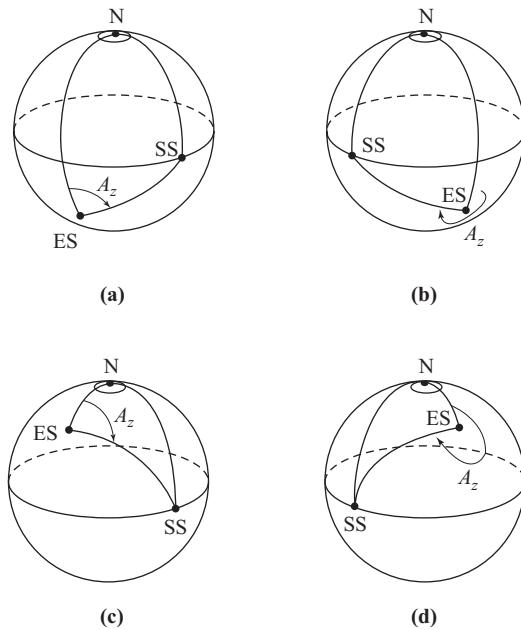


Figure 19.10.3 Azimuth angles for four possible situations: (a) earth station in the southern hemisphere, west of the subsatellite point; (b) earth station in the southern hemisphere, east of the subsatellite point; (c) earth station in the northern hemisphere, west of the subsatellite point; (d) earth station in the northern hemisphere, east of the subsatellite point.

To find the range and elevation, it is first necessary to find the side (angle) b of the quadrantal triangle. Another of Napier's rules can be used to show that

$$\cos b = \cos \lambda_E \cos B \quad (19.10.10)$$

Attention can now be transferred to the *plane triangle*, Fig. 19.10.3(c). This includes the radius of the earth at the given latitude of the earth station. It will be noted that the earth's radius does not come into the calculations for azimuth. The shape of the earth is more closely approximated as an oblate spheroid rather than a perfect sphere, for which the radius is a function of latitude and the surface represents mean sea level. The assumption of a perfectly spherical earth and ignoring earth station altitude introduces about a tenth of a degree error in angle of elevation, and a few km at most in a range of about 40 km. For our purposes the assumption of a spherical earth and ignoring earth station altitude is quite adequate. The mean radius is taken as:

$$R = 637 \text{ km} \quad (19.10.11)$$

Application of the cosine rule to the plane triangle gives the range d as

$$d = \sqrt{R^2 + a_{\text{gso}}^2 - 2Ra_{\text{gso}} \cos b} \quad (19.10.12)$$

The elevation can now be determined from application of the sine rule for plane triangles. This yields

$$\cos El = \frac{a_{\text{gso}}}{d} \sin b \quad (19.10.13)$$

These computations are illustrated in the following example, worked out in Mathcad.

EXAMPLE 19.10.1

An earth station at latitude 20° S and longitude 30° W is working into a geostationary satellite situated at longitude 30° E. Determine the look angles and the range.

Constants: $R := 6371$ km, $a_E := 42,164$ km

Given data: $\phi_S := 30^\circ$, $\phi_E := -30^\circ$, $\lambda_E := -20^\circ$

Computations:

$$B = \phi_E - \phi_S \quad \text{Eq. (19.10.3)}$$

$$A := a \tan \left(\frac{-\tan(|B|)}{\sin(\lambda_E)} \right) \quad \text{Eq. (19.10.5)}$$

$$Aza := \text{if } [(\lambda_E < 0) \cdot (B < 0), A, 0]$$

$$Azb := \text{if } [(\lambda_E < 0) \cdot (B > 0), 2\pi - A, 0]$$

$$Azc := \text{if } [(\lambda_E > 0) \cdot (B < 0), \pi + A, 0]$$

$$Azd := \text{if } [(\lambda_E > 0) \cdot (B > 0), \pi - A, 0]$$

Since only one of these will be other than zero, the azimuth can be found by setting

$$Az := Aza + Azb + Azc + Azd, \quad \mathbf{Az = 78.8^\circ}$$

$$b := a \cos(\cos(\lambda_E) \cos(B)) \quad \text{Eq. (19.10.10)}$$

$$d := \sqrt{R^2 + a_E^2 - 2Ra_E \cos(b)} \quad \text{Eq. (19.10.11),} \quad \mathbf{d = 39,572 \text{ km}}$$

$$El := a \cos \left(\frac{a_E}{d} \sin(b) \right) \quad \text{Eq. (19.10.12),} \quad \mathbf{El = 19.9^\circ}$$

A plot of azimuth and elevation for an earth station located at the authors' home city is shown in Fig. 19.10.4. For accurate pointing, the azimuth and the elevation angles must be adjusted independently, which requires two drive motors. Many domestic (backyard) installations use what is termed a *polar mount*, which employs a single drive motor. This can be installed so that the pointing is accurate for one satellite, but pointing errors occur for satellites on either side of this. Figure 19.10.5 shows the are followed by the antenna boresight compared to the true geostationary arc. Figure 19.10.6 shows how a polar mount antenna is installed.

The antenna is first installed so that its *polar axis* is pointing to true north and elevated so that its boresight is parallel to the earth's equatorial plane. Assuming a spherical earth, the earth station latitude is λ_E

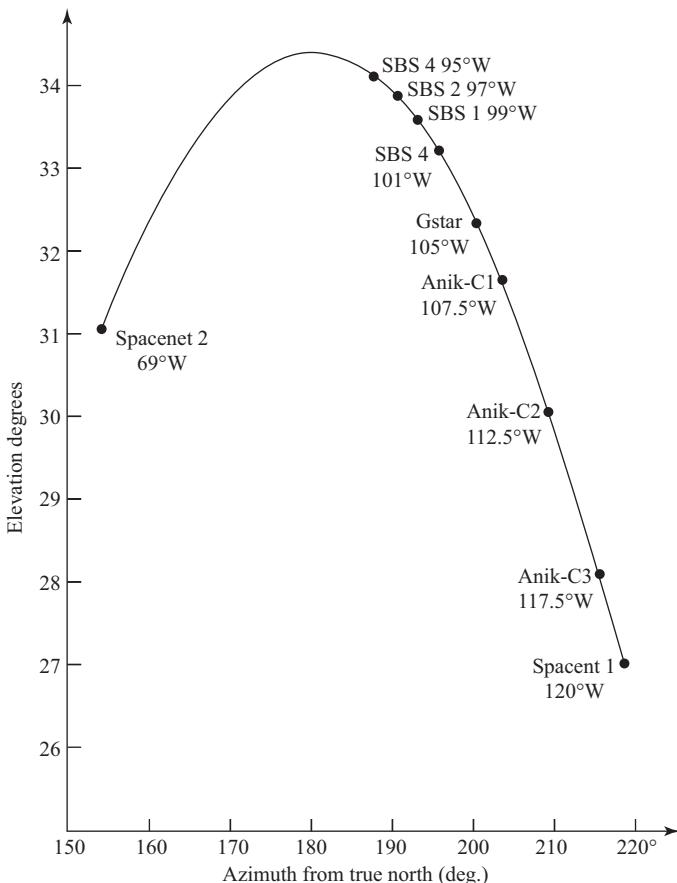


Figure 19.10.4 Azimuth-elevation angles for an earth station location 48.42°N, 89.26°W (Thunder Bay, Ontario). Ku band satellites are shown.

as shown, and its complement is $\alpha = 90^\circ - \lambda_E$. Since the polar axis is parallel to the earth's axis, the angle α also appears between the polar axis and the normal to the local horizontal plane. This is shown in detail in Fig. 19.10.6(b). It follows therefore that the angle between the polar axis and the local horizontal plane is

$$\begin{aligned}\beta &= 90^\circ - \alpha \\ &= \lambda_E\end{aligned}\tag{19.10.13}$$

Thus the first adjustment to make when installing the antenna is to point the polar axis to the true north, and adjust the elevation of the *polar axis* to be equal to the earth station latitude.

As shown, this points the boresight parallel with the equatorial plane. The antenna dish is now tilted to make the boresight intersect the geostationary arc. For this intersection, there will be zero pointing error, but errors will be introduced for satellites at either side. The intersection can be at any point on the arc, but to spread the error evenly on either side, it will be assumed to meet the arc at a point directly south of the earth station. The angle of tilt δ is readily found for this situation. (The tilt angle is sometimes referred to as the *declination*,

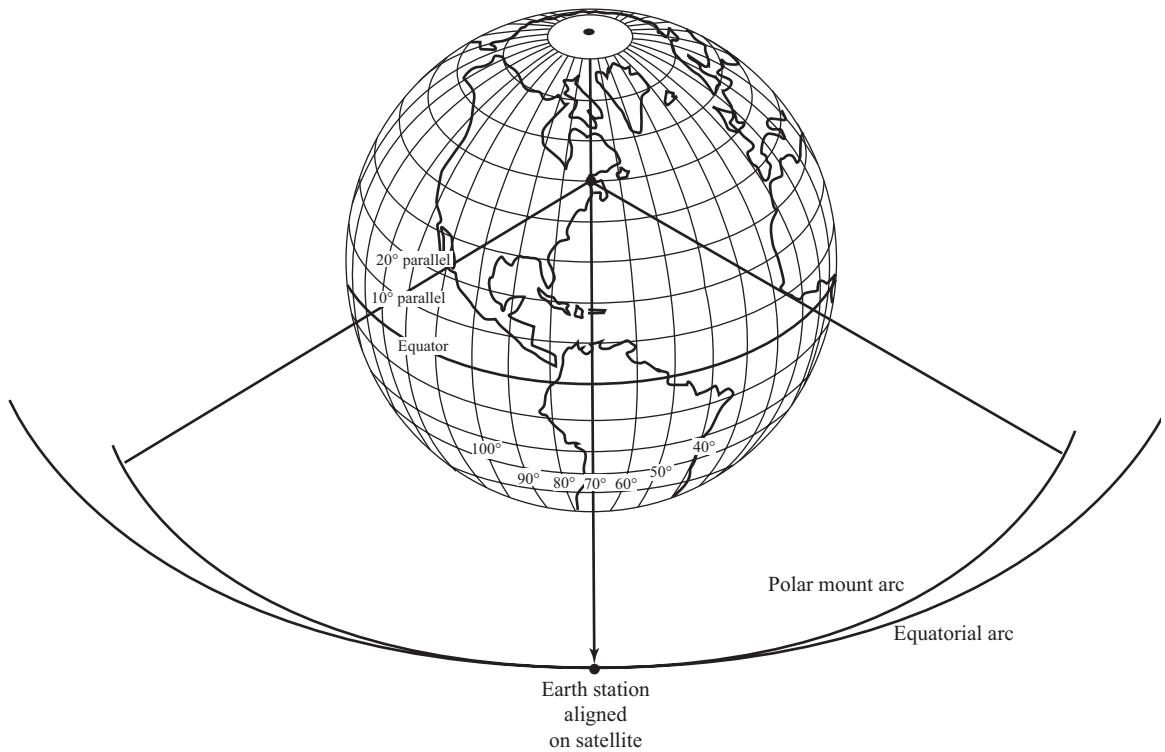


Figure 19.10.5 (a) Geostationary arc and (b) arc followed by a polar mount antenna.

but the term tilt will be used here to avoid confusion with magnetic declination used in compass correction.) The geometry is shown in Figure 19.10.7. From this, the angle of tilt is seen to be given by

$$\delta = 90^\circ - El - \lambda_E \quad (19.10.14)$$

But from Eq. (19.10.3) for a point due south, the angle $B = 0$, and hence from Eq. (19.10.10) $b = \lambda_E$. This in turn can be substituted into Eq. (19.10.13) to get

$$\cos El = \frac{a_{\text{gso}}}{d} \sin \lambda_E \quad (19.10.15)$$

Thus, in terms of earth station latitude and distances only, the angle of tilt is given by

$$\delta = 90^\circ - \arccos \left(\frac{a_{\text{gso}}}{d} \sin \lambda_E \right) - \lambda_E \quad (19.10.16)$$

Evaluation of s for a range of latitudes is left as Problem 19.11. It should be noted that in practice, rather than calculating s , the installation may be optimized by adjusting the antenna for maximum received signal from a satellite nearest to the due south point. This will introduce some asymmetry into the error curve, as a function

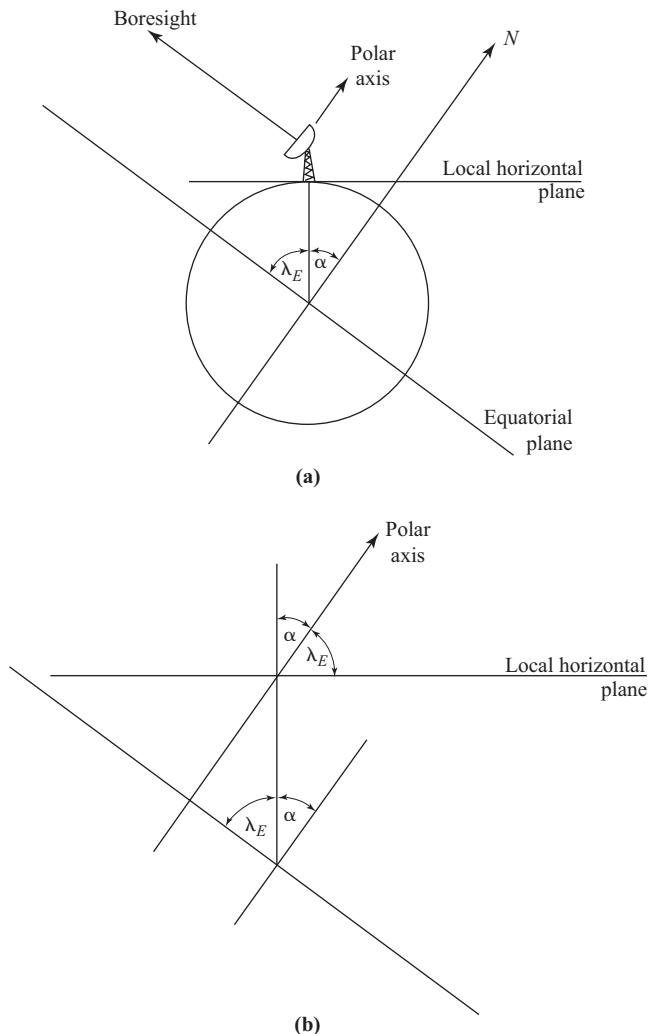


Figure 19.10.6 Installation of a single-drive antenna. (a) Antenna boresight adjusted to be parallel to the equatorial plane. (b) Geometrical relationships.

of displacement east or west of the earth station longitude. In any case, the attenuation resulting from the pointing error is generally quite small and may only be significant at the limits of visibility for the earth station.

19.11 Limits of Visibility

For any given earth station, the curvature of the earth will set the limits to the farthest satellite that can be “seen” east or west of the earth station longitude. At the limit set by the earth’s curvature, the earth station antenna will point along the horizontal, or the elevation will be zero. In practice, the noise picked up from the earth by the antenna at zero elevation is excessive, so that an angle of elevation of 5° is generally assumed as being the usable minimum. The plane triangle of Fig. 19.10.2(c) now becomes as shown in Fig. 19.11.1.

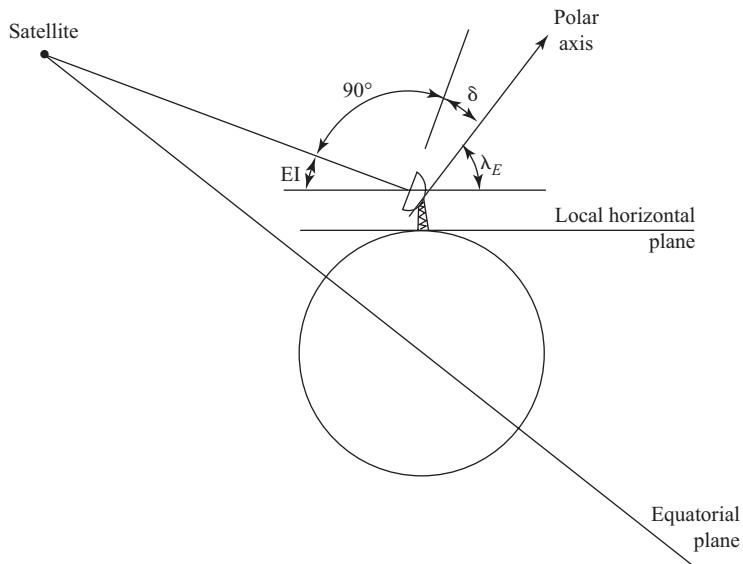


Figure 19.10.7 Angle of tilt δ : needed to make the boresight intersect the geostationary arc at a point due south of the earth station.

From the triangle

$$\sin S = \frac{R}{a_{\text{gso}}} \sin 95^\circ \quad (19.11.1)$$

This enables angle S to be calculated, and hence angle b as

$$\begin{aligned} b &= 180^\circ - 95^\circ - S \\ &= 85^\circ - S \end{aligned} \quad (19.11.2)$$

Knowing b , the longitude difference angle B can be determined from Eq. (19.10.10), which is

$$\cos B = \frac{\cos b}{\cos \lambda_E} \quad (19.11.3)$$

These calculations are illustrated in the following example worked out using Mathcad.

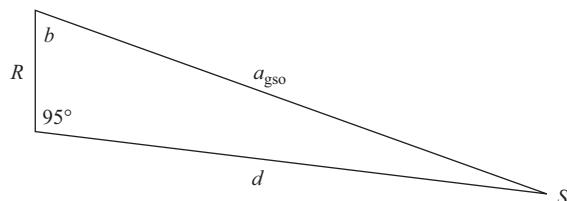


Figure 19.11.1 Plane triangle used in the calculation of limits of visibility.

EXAMPLE 19.11.1

The coordinates for an earth station are 43° south, 30° east. Calculate the limits of visibility.

SOLUTION Constants:

$R := 6378 \text{ km}$	mean radius of the earth
$a_{\text{gso}} := 42,164 \text{ km}$	geostationary arc radius
$\phi_E := 30^\circ$	ES longitude, degrees
$\lambda_E := -43^\circ$	ES latitude, degrees

Computation of limits of visibility:

$$S := a \sin \left(\frac{R}{a_{\text{gso}}} \sin(95^\circ) \right) \quad \text{Eq. (19.11.1)}$$

$$b = 85^\circ - S \quad \text{Eq. (19.11.2)}$$

$$B := a \cos \left(\frac{\cos b}{\cos \lambda_E} \right) \quad \text{Eq. (19.11.3)}$$

$$B = 71^\circ$$

The limits of visibility are

$$\phi_E - B = -41^\circ$$

$$\phi_E + B = 101^\circ$$

19.12 Frequency Plans and Polarization

There are well-defined frequency bands allocated for satellite use, the exact frequency allocations depending on the type of service (for example, mobile communications and broadcast). The frequency bands also differ depending on the geographic region of the earth in which the earth stations are located. Frequency allocations are made through the International Telecommunication Union (ITU). The most widely used bands at present are the C band and the Ku band. Uplink transmissions in the C band are nominally at 6 GHz and downlink transmissions nominally at 4 GHz. The band is sometimes referred to as the 6/4 GHz band. Uplink transmissions in the Ku band take place in the region of 14 GHz and downlink in the region of 12 GHz, this being referred to as the 14/12 GHz band. (The designation Ku arises from the fact that this frequency is under a microwave band known as the K band, and the u is sometimes shown as a subscript.) For each band, the bandwidth available is 500 MHz.

For each band mentioned, the higher-frequency range is used for the uplink (very rarely the situation is reversed, the higher frequency being used for the downlink). The reason for using the higher frequency on the uplink is that losses tend to be greater at higher frequencies, and it is much easier to increase the power from an earth station rather than from a satellite to compensate for this.

To make the most of the available bandwidth, *polarization discrimination* is used. Adjacent transponder channels can be assigned alternate polarizations, for example horizontal and vertical. Figure 19.12.1

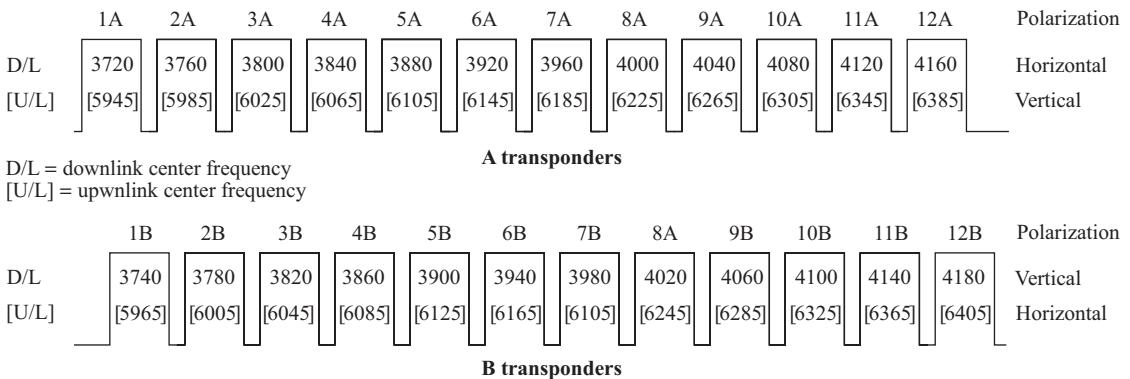


Figure 19.12.1 Anik-E frequency and polarization plan for the C band. (Courtesy Telesat Canada.)

shows the frequency and polarization plan for the C band in the *Anik-E* satellite. The 24 transponder channels are first of all formed into two groups of 12, labelled A and B transponders. The downlink signals for group A are horizontally polarized and for group B vertically polarized. Thus, although there is some overlap in the transponder bandwidths, the different polarizations prevent interference from occurring. For example, transponder 2A has a center frequency of 3760 MHz, and its bandwidth (including guard bands) extends from 3740 to 3780 MHz. Transponder 2B has a center frequency of 3780 MHz, and its bandwidth extends from 3760 to 3800 MHz. The use of polarization to increase the available frequency bandwidth is referred to as *frequency reuse*. It will also be observed from Fig. 19.12.1 that the uplink signals in each group are polarized in the opposite sense to the downlink signals.

Right-hand circular (RHC) and left-hand circular (LHC) polarization may also be used in addition to vertical and horizontal polarization, which permits a further increase in frequency reuse. The Intelsat series of satellites utilize all four types of polarization.

19.13 Transponders

The word *transponder* is coined from *transmitter-responder* and it refers to the equipment channel through the satellite that connects the receive antenna with the transmit antenna. The transponder itself is not a single unit of equipment, but consists of some units that are common to all transponder channels and others that can be identified with a particular channel. Figure 19.13.1(a) shows in block schematic form typical transponder channels for a C band satellite, and Figure 19.13.1(b) the typical frequency assignments.

Typically, a basic bandwidth of 500 MHz is available at the C band frequencies encompassing all the transponder channels and corresponding to an input (uplink) frequency range of 5.925 to 6.425 GHz, as shown in Fig. 19.13.1(a). This input range of signals is passed through a wideband, bandpass filter (BPF) to limit noise and interference and then on to a wideband receiver, which provides a frequency down-conversion common to all channels. The wideband receiver also provides the common low-noise amplification needed at the input to maintain a satisfactory signal-to-noise ratio, as described in the Section 4.11. The output frequency range is 3.7 to 4.2 GHz, which is the downlink frequency band. The wideband receiver is shown in more detail in Fig. 19.13.2. Typical signal levels are shown in decibels relative to the signal level at the receive antenna. The overall gain is provided in two sections, one at the input frequency range and the other at the output frequency range. This makes for a more stable arrangement and prevents oscillation, which might arise if the gain was provided all at one frequency range. Solid-state amplifiers are used throughout.

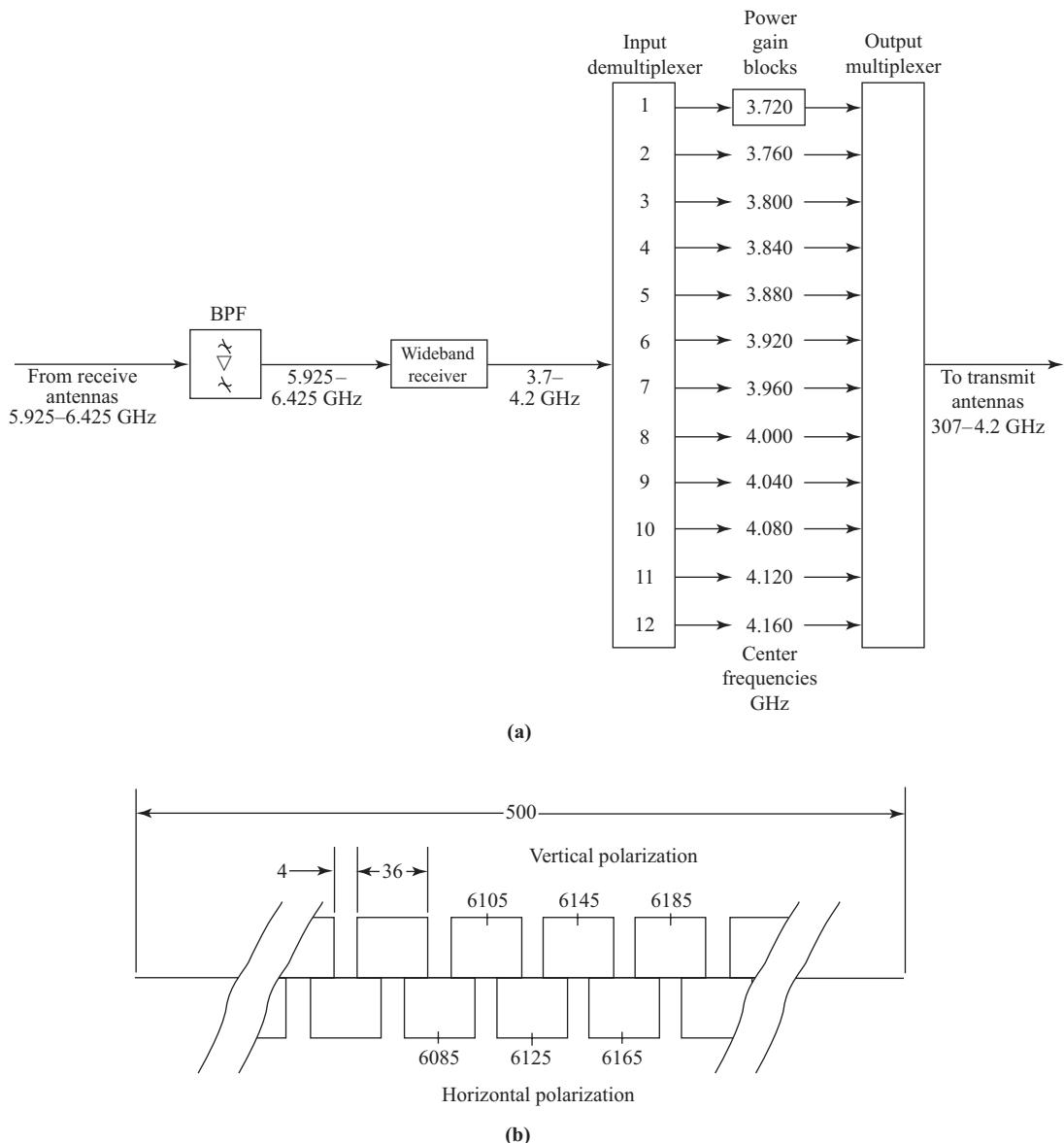


Figure 19.13.1 (a) C band satellite transponder channels. (b) Section of an uplink frequency and polarization plan. Numbers refer to frequency in megahertz.

Because the wideband receiver is critical to all transponders, a *redundant* receiver is provided. This is essentially a backup receiver that is switched in automatically if the other fails. An input *demultiplexer* following the wideband receiver is an arrangement of microwave circulators and filters that separates the 500-MHz band into the separate transponder channel bandwidths. A typical transponder bandwidth is 36 MHz, or 40 MHz including guard-bands, as shown in Fig. 19.13.1, although other values are commonly used. Following the demultiplexer, power amplifiers are provided for the individual transponder channels,

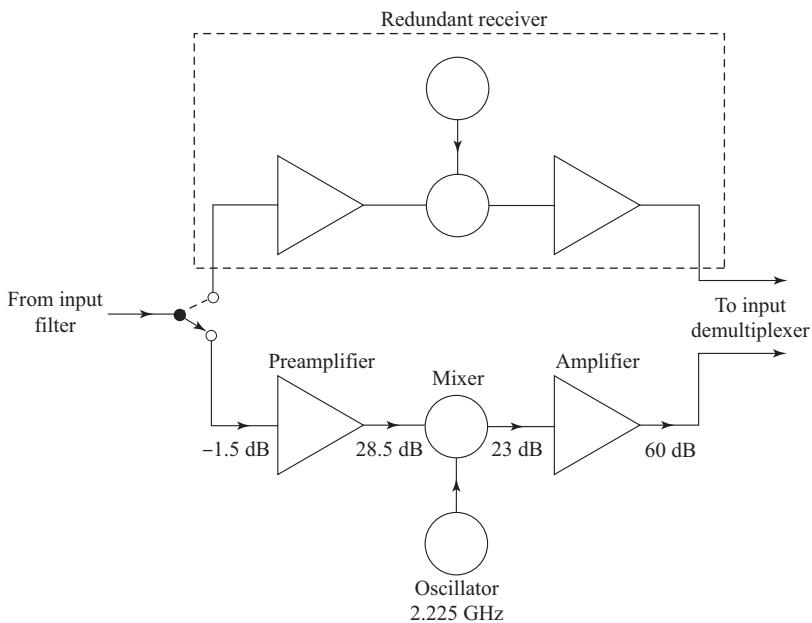


Figure 19.13.2 Satellite wideband receiver.

which brings the power levels up to those required for retransmission on the downlink. The power levels are shown in Fig. 19.13.3.

The transponder bandwidth of 36 MHz may be used by a single carrier modulated with a wideband signal, such as a TV signal or an FDM telephony baseband signal. It is also possible to divide the 36-MHz bandwidth into smaller bandwidths, which are then assigned to different carriers. This gives rise to a method of accessing a transponder known as *frequency division multiple access* (FDMA). Figure 19.13.4 shows one scheme in which about 800 one-way telephony channels may be assigned in the 36-MHz bandwidth.

A problem that arises with FDMA is that of nonlinearity in the power amplifier. The transfer curve for a typical high power amplifier is shown in Fig. 19.13.5. This is seen to be nonlinear, and operation in the nonlinear region near the peak results in a form of distortion known as *intermodulation distortion* when multiple carriers are present, such as occurs with FDMA. In earlier satellites, microwave tubes known as *traveling wave tubes* (TWTs) provided the required power amplification. These tubes continue to be used because they provide high power output at wide bandwidths, but gradually solid-state high-power amplifiers (SSHPAs) are being developed for this application. Compared to TWTAs, the SSHPAs cannot deliver as high a power output, but they produce less intermodulation distortion.

The peak point on the curve for single carrier operation is referred to as the *saturation point*. When multiple carriers are present, the power input is *backed off* from the saturation point to avoid the worst of the nonlinearity, which reduces intermodulation distortion to an acceptable level. Appropriately enough, the term *backoff* is used to describe this operation. There is an *input backoff* with a corresponding *output backoff* which typically is about 5 dB less than the input backoff. Where the carriers in an FDMA system transmit equal powers, the power in each carrier must be reduced by the input backoff amount, usually expressed in decibels.

Once the individual transponder signals have been amplified to the required power levels, they are combined in a multiplexer to form a wideband signal covering the downlink frequency range of 3.7 to 4.2 GHz, and this wideband signal is radiated by the transmit antenna or antennas.

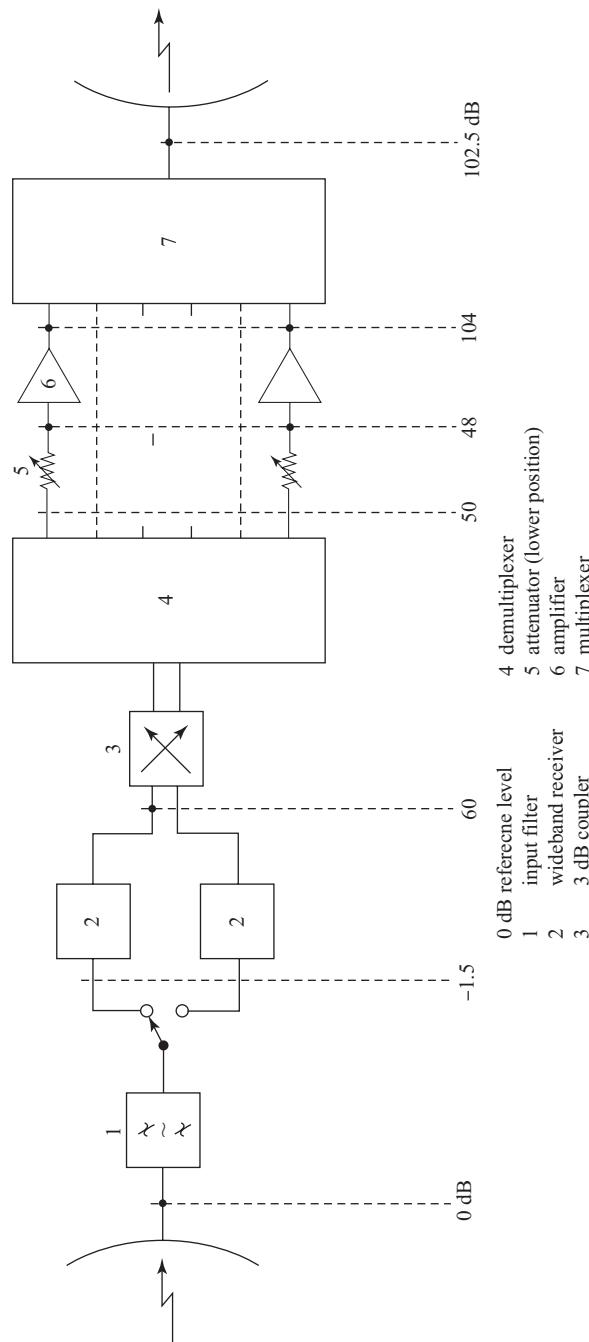


Figure 19.13.3 Typical relative power levels in a transponder. (Courtesy of CCIR. CCIR Fixed Satellite Services Handbook, p. 19, sect 4.2.2, final draft 1984).

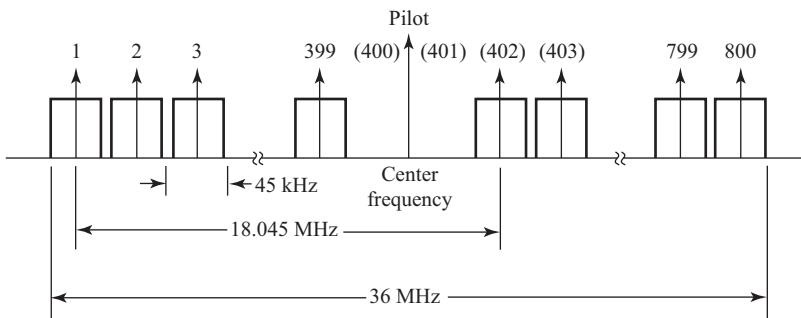


Figure 19.13.4 Channeling arrangements for Intelsat SCPC system.

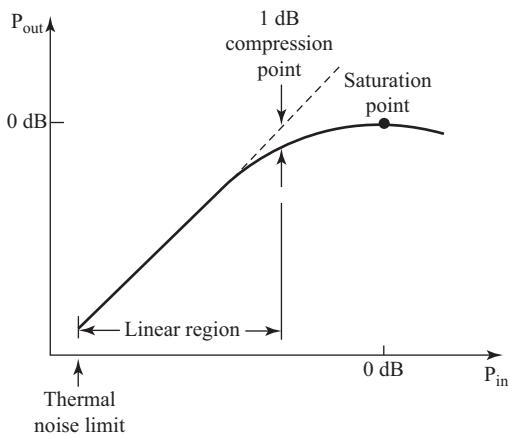


Figure 19.13.5 Power transfer curve for a satellite TWTA.

19.14 Uplink Power Budget Calculations

A power budget calculation simply shows how the transmitted power is accounted for across a communications link. Starting at the transmitter output, there will be various power gains in the system, which will increase the available power, and various losses, which will reduce it, and the received power will be the transmitted power plus the gains minus the losses. What is important is that decibel values must be used when adding and subtracting these quantities. Because decibel (dB) values are so frequently used, square brackets will be used to signify these. A power ratio X expressed in decibels is, by definition,

$$[X] = 10 \log X \quad (19.14.1)$$

Strictly speaking, the term *decibel* applies only to the power ratios, but for a detailed discussion of the extended use of decibels, see Appendix A. For the uplink, the transmitter power is generated by a high-power microwave amplifier (usually a klystron or traveling wave tube), and this power will be denoted by P_{HPA} watts. When expressing this in decibels, it is essential that the reference power be stated, and this will be

assumed to be 1 W. The power is then expressed in decibels referenced to 1 W, which is denoted by dBW. The transmitter power is transferred to the transmit antenna through a feeder, which will have certain losses. Denoting these at [TFL] decibels, this loss must be subtracted from $[P_{HPA}]$ to get the actual power, in dBW, radiated. This will be increased in the direction of maximum radiation by the power gain of the transmit antenna. Denoting the isotropic power gain of the earth station antenna by G_{ES} and using the subscript U to identify uplink quantities where necessary (similar downlink quantities will be introduced shortly, identified by a subscript D) the *equivalent isotropic radiated power* (EIRP) is

$$\text{EIRPdB}_U := \text{PdB}_{HPA} - \text{TFLdB} + \text{GdB}_{ES} \quad (19.14.2)$$

In Section 15.2 it is shown that the free space loss is given by FSL of Eq. (15.2.8). There will be additional losses, amounting to a few decibels, that have to be added to the free space loss. On the uplink, the transmit antenna boresight may not be pointing exactly at the satellite (see Section 19.10). This is referred to as the *antenna misalignment loss* (AML) or sometimes as the *antenna pointing loss*. This will be denoted by [AML] dB. There may also be a loss as a result of polarization misalignment, referred to as the *polarization loss* and denoted in decibels as [PL]. Losses also occur in transmission through the earth's atmosphere as a result of energy being absorbed by the atmospheric gases. This energy goes mainly into vibrational energy of certain molecules, which is subsequently lost as heat. It will be denoted by [AA] dB for *atmospheric absorption losses*. Note that this term does not take into account rain attenuation, which must be allowed for separately. The *transmission path loss* or [TPL] is conveniently defined as

$$\text{TPLdB}_U := \text{FSLdB}_U + \text{AMLDdB}_U + \text{PLdB}_U + \text{AAdB}_U \quad (19.14.3)$$

At the satellite receiver input, the power available at the receive antenna terminals must take into account the isotropic power gain of the satellite antenna [G_{SAT}] and any receive feeder losses [RFL]. The received power in dBW is therefore given by

$$[P_{UR}] = [\text{EIRP}]_U - [\text{TPL}]_U + [G_{SAT}] - [\text{RFL}]_U \quad (19.14.4)$$

As pointed out in Chapter 4, it is not power by itself that is significant in a receiving system, but rather the signal-to-noise power ratio. In this context, the signal power is also the received carrier power as given by Eq. (19.14.4), and the ratio is referred to as the *carrier-to-noise ratio* or C/N. Denoting the equivalent noise temperature of the satellite receiving system, referred to the satellite receiver input by T_{SAT} , the carrier-to-noise ratio is

$$\begin{aligned} \left[\frac{C}{N} \right]_U &= \left[\frac{P_{UR}}{kT_{SAT}B_N} \right] \\ &= [\text{EIRP}]_U - [\text{TPL}] + [G_{SAT}] - [\text{RFL}]_U - [k] - [T_{SAT}] - [B_N] \end{aligned} \quad (19.14.5)$$

where B_N is the noise bandwidth and k is Boltzmann's constant. An important figure of merit of a satellite receiving system is the ratio of receive antenna gain to system noise temperature. In decibel-like units, this is

$$\left[\frac{G}{T} \right]_U = [G_{SAT}] - [T_{SAT}] \quad (19.14.6)$$

Note carefully, however, that the ratio of G to T involves different kinds of quantities. It must be understood that G is dimensionless, while T has dimensions of temperature. The reference temperature is taken as 1 K,

and the decibel notation for $[G/T]$ becomes dBK^{-1} , which is sometimes written as dB/K . This must *not* be interpreted as decibels per kelvin. Equation (19.14.5) can now be written as

$$\left[\frac{C}{N} \right]_U = [\text{EIRP}]_U - [\text{TPL}]_U - [\text{RFL}]_U + \left[\frac{G}{T} \right]_{\text{SAT}} - [k] - [B_N] \quad (19.14.7)$$

The subscript SAT is used with $[G/T]$ to signify that this is the satellite figure of merit. Frequently, the carrier-to-noise power density ratio, denoted by C/N_0 , is used instead of $[C/N]$. The noise power density is given by $N_0 = kT_{\text{SAT}}$, and it is left as an exercise for the student to show that, on substituting the numerical value for k , Eq. (19.14.7) can be rearranged as

$$\text{CNoRdB}_U := \text{EIRPdB}_U - \text{TPLdB}_U - \text{RFLdB}_U + \text{GTRdB}_{\text{SAT}} + 228.6 \quad (19.14.8)$$

An uplink power budget calculation is given in the following example, worked in Mathcad. The notation is changed to conform to Mathcad requirements. For example $[C/N_0]_U$ is shown as CNoRdB_U for “uplink carrier-to-noise density ratio in decibels relative to 1 Hz.”

EXAMPLE 19.14.1

The high-power amplifier in an earth station delivers 600 Watts to the antenna feeder. The transmit feeder loss is 1.5 dB, and the antenna gain is 50 dB. The uplink free-space loss is 200 dB, the antenna misalignment loss is 0.5 dB, the polarization loss is 0.5 dB, and the atmospheric absorption loss is 1 dB. At the satellite, the receiver feeder loss is 1 dB, and the G/T ratio is -8 dB. Calculate the carrier-to-noise density ratio received at the satellite.

SOLUTION Given data:

$$\begin{aligned} P_{\text{HPA}} &:= 600 \text{ W} & \text{TFLdB} &:= 1.5 & \text{GdB}_{\text{ES}} &:= 50 \\ \text{RFLdB}_U &:= 1 & \text{GTRdB}_{\text{SAT}} &:= -8 \\ \text{FSLdB}_U &:= 200 & \text{AMLDdB}_U &:= 0.5 & \text{PLdB}_U &:= 0.5 & \text{AAdB}_U &:= 1 \end{aligned}$$

Note, $\text{GTRdB}_{\text{SAT}}$ denotes the G/T at the satellite, the units being decibels relative to 1 K for the temperature.

Computations: The HPA output in decibels relative to 1 Watt (dBW) is

$$\text{PdB}_{\text{HPA}} := 10 \log \left(\frac{P_{\text{HPA}}}{1 \text{ W}} \right)$$

$$\text{EIRPdB}_U := \text{PdB}_{\text{HPA}} - \text{TFLdB} + \text{GdB}_{\text{ES}} \quad \text{Eq. (19.14.2)}$$

$$\text{TPLdB}_U := \text{FSLdB}_U + \text{AMLDdB}_U + \text{PLdB}_U + \text{AAdB}_U \quad \text{Eq. (19.14.3)}$$

$$\text{CNoRdB}_U := \text{EIRPdB}_U - \text{TPLdB}_U - \text{RFLdB}_U + \text{GTRdB}_{\text{SAT}} + 228.6 \quad \text{Eq. (19.14.8)}$$

The C/N_0 ratio in decibels relative to 1-Hz bandwidth is

$$\text{CN}_0 \text{RdB}_U = 93.9$$

The *saturation flux density* is the power flux density (in watts per square meter) at the satellite receive antenna needed to drive the TWT into saturation (see Fig. 19.13.5), and uplink calculations are often made in terms of this quantity. This can be factored into the equations by noting that the power available at the terminals of an isotropic antenna is given by Eq. (19.14.4) on setting $G_{\text{SAT}}=1$, or equivalently

$[G_{SAT}] = 0 \text{ dB}$. Since it is the power at the antenna terminals that is being considered, the receiver feeder losses do not enter into the calculation, and therefore $[RFL]_U = 0$ also in Eq. (19.14.4). Hence the isotropic power received is

$$[P_{ri}] = [EIRP]_U - [TPL]_U \quad (19.14.9)$$

The effective area of an isotropic antenna is given by Eq. (16.8.4) on setting the gain equal to unity as

$$A_o = \frac{\lambda^2}{4\pi} \quad (19.14.10)$$

With λ in meters, the area in decibels referred to 1 m^2 is

$$[A_o] = 10 \log \frac{\lambda^2}{4\pi} \quad (19.14.11)$$

Denoting the power flux density in general by ϕ , the power received by the isotropic antenna is

$$[P_{ri}] = [\phi] + [A_o] \quad (19.14.12)$$

Hence, equating Eqs. (19.14.9) and (19.14.12) gives

$$[EIRP]_U - [TPL]_U = [\phi] + [A_o]_U \quad (19.14.13)$$

This can now be substituted in Eq. (19.14.8) to give

$$\left[\frac{C}{N_o} \right]_U = [\phi] + \left[\frac{G}{T} \right] - [RFL]_U + [A_o]_U + 228.6 \quad (19.14.14)$$

In this equation, ϕ represents power flux density in general. When the saturation value ϕ_s is specified, along with the input backoff $[BO_i]$, Eq. (19.14.14) becomes

$$\left[\frac{C}{N_o} \right]_U = [\phi_s] - [BO_i] + \left[\frac{G}{T} \right]_{SAT} - [RFL]_U + [A_o]_U + 228.6 \quad (19.14.15)$$

The link budget equations developed so far are for what is termed *clear sky* conditions, meaning that no precipitation effects are taken into account. Rain can degrade the carrier-to-noise ratio, mainly by attenuating the signal, although it may also cause a small rise in the effective noise temperature. By monitoring signal levels, the effect of rain fades can be compensated for as they arise, although this can require expensive monitoring and control equipment.

An uplink power budget calculation for clear sky conditions is illustrated in the following example. (See the note relating to Mathcad notation, Example 19.14.1.)

EXAMPLE 19.14.2

An uplink operates at a frequency of 14 GHz at a backoff of 10 dB. The flux density required to saturate the satellite transponder is -98 dB relative to 1 W/m^2 . The satellite G/T ratio is 3 dB relative to 1 K and the receiver feeder losses are 1 dB. Calculate the received C/N_o ratio.

SOLUTION Given data:

$$f = 14 \times 10^9 \text{ Hz} \quad \text{BOdB}_i := 10 \quad \text{GTRdB}_{\text{SAT}} := 3 \quad \text{RFLdB}_U := 1 \quad \phi dB_S := -98 \quad c := 3 \times 10^8 \text{ m/s}$$

Computations:

$$\lambda := \frac{c}{f}$$

$$A_0 dB_U = 10 \log \left(\frac{\lambda^2}{4\pi \cdot 1 \text{ m}^2} \right)$$

Eq (19.14.11), note the need to include
1 m² in the denominator to maintain
correct dimensionality

$$\text{CNdB}_U := \phi dB_S - \text{BdB}_i + \text{GTRdB}_{\text{SAT}} - \text{RFLdB}_U + \text{AodB}_U + 228.6 \quad (19.14.15)$$

The C/N_o ratio is

$$\text{CNdB}_U = 78.2$$

19.15 Downlink Power Budget Calculations

The downlink power budget calculation is similar to that for the uplink, and Eq. (19.14.8) can be used with usually be a TWT amplifier (although solid-state amplifiers are becoming more common), and the transmitter power P_{HPA} in Eq. (19.14.2) will be replaced by P_{TWT} . Using the subscript D to identify downlink quantities, the equation for EIRP therefore becomes

$$[\text{EIRP}]_D = [P_{\text{TWT}}] + [G_{\text{SAT}}] - [\text{TFL}]_D \quad (19.15.1)$$

This equation applies in general, but if the output power is specified for saturation conditions, then the EIRP will also be the saturation value. As mentioned in connection with uplink calculations, an input backoff is often employed. Under these conditions an output backoff must also occur. The output backoff in general will not be equal to the input backoff, and as a rule of thumb the output backoff is about 5 dB less than the input backoff. Suppose, therefore, that the saturation output P_{TWTS} along with the output backoff [BO_o] in decibels is specified; then

$$[\text{EIRP}]_D = [P_{\text{TWTS}}] + [G_{\text{SAT}}] - [\text{TFL}]_D - [\text{BO}_o] \quad (19.15.2)$$

The transmission path loss for the downlink is given by

$$[\text{TPL}]_D = [\text{FSL}]_D + [\text{AML}]_D + [\text{PL}]_D + [\text{AA}]_D$$

The carrier-to-noise ratio at the earth station is therefore given by

$$\left[\frac{C}{N_o} \right]_D = [\text{EIRP}]_D - [\text{TPL}]_D - [\text{RFL}]_D + \left[\frac{G}{T} \right]_{\text{ES}} + 228.6 \quad (19.15.3)$$

where the $[G/T]$ ratio is that of the receiving earth station, denoted by subscripts ES. There is no need to rearrange this downlink equation in terms of a saturation flux density (as was done for the uplink) since the concept of saturation flux density has no significance for earth stations.

As with the uplink, the equations developed for the downlink apply for clear sky conditions, and degradation of carrier-to-noise resulting from rain must be accounted for separately.

19.16 Overall Link Budget Calculations

A full communication circuit consists of an uplink and a downlink, and the overall C/N ratio will be the combined effect of both. Let the power gain from the receiver input terminals of the satellite to the receiver input terminals of the destination earth station be denoted by γ . If P_{CSAT} is the received carrier power at the satellite, the corresponding carrier power at the destination earth station receiver will be $P_C = \gamma P_{CSAT}$. If P_{NSAT} is the noise power density at the input to the satellite receiver, the noise power density at the input to the destination earth station receiver will be the sum of the noise power already present at the earth station, P_{NES} , and that “carried down” from the satellite. The total noise at the earth station is $P_N = P_{NES} + P_{NSAT}$, and the destination carrier-to-noise ratio is

$$\frac{C}{N_o} = \frac{P_C}{P_N} \quad (19.16.1)$$

To determine this in terms of the individual link values, it is easiest to work in terms of reciprocal values, but note carefully that power ratios, not decibels, must be used (P_N can be considered to be the noise power in unit bandwidth). Hence

$$\begin{aligned} \frac{N_o}{C} &= \frac{P_N}{P_C} \\ &= \frac{P_{NES} + \gamma P_{NSAT}}{\gamma P_{CSAT}} \\ &= \frac{P_{NES}}{\gamma P_{CSAT}} + \frac{P_{NSAT}}{P_{CSAT}} \\ &= \left(\frac{N_o}{C} \right)_D + \left(\frac{N_o}{C} \right)_U \end{aligned} \quad (19.16.2)$$

In summary, the separate link values must first be converted to power ratios and then their reciprocals added to give N_0/C , and from this the overall C/N_0 ratio is obtained. The following example illustrates this calculation.

EXAMPLE 19.16.1

For a satellite communications channel, the uplink C/N_0 ratio is 88 dBHz, and the downlink value is 78 dBHz. Calculate the overall C/N_0 ratio in dBHz.

SOLUTION Given data: $CN_{dB_U} := 88$ $CN_{dB_D} := 78$

Denoting the noise-to-carrier ratios by N_0C , then

$$NoC_U := 10^{\frac{-CN_{dB_U}}{10}} \quad NoC_D := 10^{\frac{-CN_{dB_D}}{10}}$$

$$NoC := NoC_U + NoC_D$$

$$CN_{dB} := 10 \log\left(\frac{1}{NoC}\right), \quad CN_{dB} = 77.6$$

19.17 Digital Carrier Transmission

As shown in Section 12.9, the parameter of importance in digital carrier transmission is the ratio of average bit energy to the noise power density, or E_b/N_o . Denoting the received carrier power in general as P_R , and the bit rate as R_b , the average bit energy is

$$E_b = \frac{P_R}{R_b} \quad (19.17.1)$$

[This is simply a repeat of Eq. (12.9.11)]. The connection between E_b/N_o and C/N_0 is therefore

$$\begin{aligned} \frac{E_b}{N_o} &= \frac{P_R}{N_o R_b} \\ &= \left(\frac{C}{N_o}\right) \frac{1}{R_b} \end{aligned} \quad (19.17.2)$$

In decibel-like quantities, this becomes

$$\left[\frac{E_b}{N_o}\right] = \left[\frac{C}{N_o}\right] - [R_b] \quad (19.17.3)$$

For the bit rate expressed in decibels, the reference unit is 1 bps. This is illustrated in the following example.

EXAMPLE 19.17.1

The required E_b/N_o ratio for a digital satellite link is 9.6 dB, and the bit rate is 1.544 Mbps. Calculate the required C/N_0 ratio.

SOLUTION Given data: $EbNo \text{ dB} := 9.6$ $R_b := 1.544 \times 10^6 \text{ s}^{-1}$

The bit rate in decibels relative to 1 bps is

$$R \text{ dB}_b := 10 \log\left(\frac{R_b}{s^{-1}}\right)$$

(Note the need to include s^{-1} in the denominator for correct dimensionality.)

$$\text{CN}_o \text{ dB} := E_b N_o \text{ dB} + 2 \text{ dB}_b$$

The required C/N_o ratio is

$$\text{CN}_o \text{ dB} = 71.5$$

19.18 Multiple-access Methods

Because of the wideband available on satellite transponders, a number of carriers may utilize a transponder together. Mention has already been made of *frequency division multiple access* (FDMA), where carriers access the transponder simultaneously but each in their own frequency slot (see Section 19.13). Implementation of FDMA is relatively straightforward, but its main drawbacks are the need for the backoff already referred to in Section 19.13 and the fact that network management, for example changing frequency allocations, requires changes in hardware such as filters.

With *time division multiple access* (TDMA), the earth stations are assigned nonoverlapping time slots in a time frame. The stations access the transponder as sketched in Fig. 19.18.1.

A reference station transmits repetitive bursts that define the time frames, and the traffic stations transmit the traffic bursts during the assigned slots within the frames. At any given time, only one carrier is being amplified in the power amplifier (for example, the TWTA), so intermodulation is absent and the amplifier

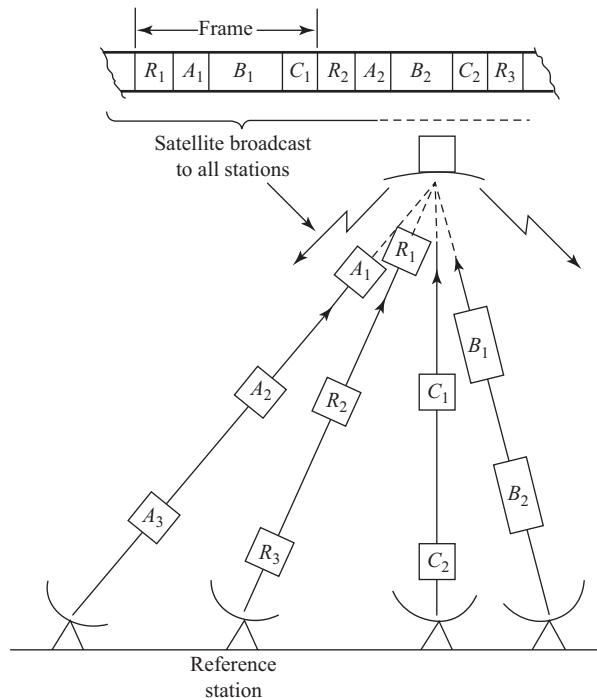


Figure 19.18.1 Time division multiple access.

can be operated at its saturation level. The modulated carrier occupies the full bandwidth of the transponder during this period.

Because of the bursty nature of the transmissions, TDMA is only suited to digitally modulated signals. Accurate synchronization of earth station transmissions is required to prevent collisions between bursts, and this makes the system technically more complex than FDMA. Its advantages are, however, that a higher power output can be achieved for the downlink since the power amplifier aboard the satellite is operated at its saturation level. Also, because of the digital nature of the signals, network management can be handled through software control.

A third method of access, known as *code division multiple access* (CDMA), is becoming more widely used in commercial applications. Initially, this method was largely restricted to military applications because of its cost and complexity. Briefly, all the carriers access the transponder at the same time and occupy the full bandwidth; thus they overlap in time and frequency. Each carrier, however, is modulated by its own digital codeword, a copy of which is stored at the destination earth station. This key enables the earth station receiver to detect the correct carrier even in the presence of overlapping signals.

In one commonly used method the digital codeword modulates the carrier at a high rate, which spreads the signal spectrum over the available bandwidth. This is referred to as *spread spectrum*, and the method is sometimes referred to as *spread spectrum multiple access*.

PROBLEMS

- 19.1. State and explain Kepler's laws in relation to artificial satellites orbiting the earth. An earth-orbiting satellite has a period of 12 h. Calculate the value of the semimajor axis. Given that the eccentricity is 0.002, calculate also the value of the semiminor axis.
- 19.2. A satellite orbiting in the earth's equatorial plane has an apogee height of 10,000 km and a perigee height of 7000 km. Given that the earth's equatorial radius is 6378.14 km, determine the semimajor axis for the satellite orbit and the period.
- 19.3. A satellite in inclined orbit crosses the ascending node at a longitude of 35° east. If the orbital period is 100 min, calculate the longitude at which the satellite next crosses the ascending node, assuming a perfectly spherical earth. (One effect of the earth's equatorial bulge is to cause a shift in the orbital plane of a satellite, an effect that is absent for a perfectly spherical earth.)
- 19.4. Explain what is meant by the *geostationary orbit* and why there is only one such orbit. Calculate the minimum delay time for a signal transmitted from a geostationary satellite to reach the earth.
- 19.5. A satellite in a circular equatorial orbit moves in an easterly direction and has an orbital period of 11 h. Calculate the time interval between successive appearances at a given longitude.
- 19.6. Explain what is meant by *attitude control* and *station keeping* and why these maneuvers are necessary for geostationary satellites.
- 19.7. An earth station at latitude 35°N, longitude 35°W is receiving from a geostationary satellite at longitude 25°W. Determine all the angles in the spherical triangle defined by these points.
- 19.8. An earth station at latitude 15°S, longitude 12°E is receiving from a geostationary satellite at longitude 25°W. Determine all the angles in the spherical triangle defined by these points.
- 19.9. Determine the range and the look angles for the conditions specified in Problems 19.7 and 19.8.
- 19.10. A satellite communications link is set up between earth stations at 48°N, 78°W, and 38°N, 40°W. Determine the one-way propagation delay time for the link.

- 19.11.** Explain what is meant by a *polar mount* antenna. Plot a graph showing the angle of tilt as a function of latitude for latitudes ranging from 0° to 70° .
- 19.12.** Explain what is meant by the *limits of visibility* in relation to satellite communications. Determine the limits of visibility for your home location. Assume a minimum antenna angle of elevation of 5° .
- 19.13.** Determine the limits of visibility for the earth stations of Problem 19.10.
- 19.14.** The borders of a certain country can be roughly represented by a triangle, the coordinates of which are

$\phi_E^\circ E$	39	43.5	48.5
$\lambda_E^\circ N$	33.5	37.5	30

If a geostationary satellite has to be visible from any point in the country, determine the limits of visibility as set by the satellite longitude. Assume a minimum angle of elevation of 5° for the earth station antenna. State clearly which geographic location fixes which limit.

- 19.15.** Determine the maximum possible longitudinal separation that can exist between a geostationary satellite and an earth station that maintain a line-of-sight communications link. Give the corresponding latitude of the earth station. Assume a minimum angle of 5° for the earth station antenna.
- 19.16.** An earth station has a high-power amplifier that has an output power of 300 W. This is fed through a feeder link to an antenna that has a gain of 48 dB, the feeder loss being 1.5 dB. Calculate the EIRP in decibels relative to (a) 1 W, (b) 1 mW, and (c) 1 kW.
- 19.17.** For a satellite uplink transmission, the EIRP is 32 dBW and the free space loss is 198 dB. The allowance for antenna misalignment loss is 0.5 dB, and for polarization loss and atmospheric attenuation combined it is 1 dB. Calculate the received power at the satellite if the satellite receiving antenna has a gain of 35 dB. The receiver feeder loss is 0.5 dB.
- 19.18.** For the uplink specified in Problem 19.17, the equivalent noise temperature of the satellite receiver is 500 K and the noise bandwidth is 36 MHz. Calculate the carrier-to-noise ratio in decibels at the satellite receiver.
- 19.19.** A satellite receiver has a G/T ratio of -7 dB/K , and the receiver feeder losses are 1 dB. The earth station transmits an EIRP of 50 dBW, and the transmission path losses amount to 205 dB. Calculate the carrier-to-noise power density at the receiver.
- 19.20.** An antenna has a gain of 48 dB at a frequency of 6 GHz. Calculate (a) its effective area and (b) the effective area of an isotropic antenna operating at the same frequency.
- 19.21.** Explain what is meant by *saturation flux density* in connection with satellite transmission. A satellite link operates at a frequency of 14 GHz. The transmission path loss is 210 dB, and the earth station EIRP is 52 dBW. Calculate the power flux density at the satellite receiving antenna.
- 19.22.** The saturation flux density specified for an uplink is $-118 \text{ dB} (\text{W/m}^2)$. The transmission path loss is 209 dB and the frequency is 14 GHz. Calculate the earth station EIRP required.
- 19.23.** The G/T ratio for the satellite receiver in Problem 19.22 is -5 dB/K , and the receiver feeder losses are 0.7 dB. Calculate the carrier-to-noise density at the satellite receiver for (a) no backoff employed and (b) 5-dB input backoff required.
- 19.24.** A satellite TWTA produces an output power of 10 W. The satellite antenna gain is 40 dB, and the transmit feeder losses are 1 dB. Calculate the EIRP in dBW.
- 19.25.** Explain why saturation flux density at the earth station is not a significant parameter in downlink power budget calculations. The saturation value of EIRP from a satellite is 23 dBW, and an output backoff of 6 dB is in operation. The transmission path losses amount to 200 dB. At the earth station receiver the G/T ratio is 41 dB/K and the receiver losses are 2 dB. Calculate the receiver carrier-to-noise density ratio.

- 19.26.** Considering the downlink only, determine the satellite EIRP required to produce a carrier-to-noise ratio of 26 dB at the earth station. The bandwidth is 36 MHz, the transmission path loss is 203 dB, and the earth station receiver feeder loss is 2 dB.
- 19.27.** For the downlink of Problem 19.26, an output backoff of 6 dB is employed at the satellite. The satellite antenna gain is 45 dB, and the transmit feeder loss is 2 dB. Calculate the saturation output power of the satellite power amplifier.
- 19.28.** A satellite circuit has an uplink C/N ratio of 28 dB and a downlink C/N ratio of 20 dB. Calculate the overall C/N ratio at the destination earth station.
- 19.29.** A satellite circuit has the following parameters:

Uplink

Saturation flux density, -68 dBW/m^2

Input backoff, 11 dB

Satellite G/T , -12 dB/K

Downlink

Satellite saturation EIRP, 26.5 dBW

Output backoff, 6 dB

Transmission path loss, 203 dB

Earth station G/T , 41 dB/K

Calculate the carrier-to-noise density ratio for uplink and downlink and the combined value.

- 19.30.** A satellite circuit has the following parameters:

	<i>Uplink</i>	<i>Downlink</i>
[EIRP] dBW	55	34
[G/T] dB/K	-2	12
[FSL] dB	200	198
[RFL]dB	2	1.5
[AMLJdB]	0.5	0.5
[PL]dB	0.7	0.7
[AAJdB]	1	1

Calculate the overall $[C/N_o]$ value.

- 19.31.** A satellite circuit has the following parameters:

	<i>Uplink</i>	<i>Downlink</i>	
Frequency, GHz	14	Saturation [EIRP], dBW	28
Sat. flux density, dBW/m^2	-90	Output backoff, dB	3
Input backoff, dB	8	[TPL], dB	200
Satellite [G/T]	-3	Earth stn. [G/T], dB/K	26.4

Calculate the overall $[C/N_o]$ value.

- 19.32.** For the downlink of a digital satellite circuit, the transmission path loss is 207 dB and the $[G/T]$ at the receiver is 0 dB/K Calculate the satellite [EIRP] required to maintain a transmission rate of 60 Mbps at a $[E_b/N_o]$ of 9 dB.
- 19.33.** Repeat the calculations in Problem 19.32 given that the BER on the downlink must not exceed 10^{-7} . The polar curve of Fig. 12.5.1 may be used.
- 19.34.** For an FDMA uplink, the frequency is 6 GHz nominal. The range is 39,000 km and the losses are [PL], 0.2 dB; [AA], 0.7 dB; [RFL], 1.5 dB; [TFL], 1.2 dB; [AML] 0.5 dB. The input backoff is 11 dB. If a total of five identical earth stations access the transponder in the FDMA mode, calculate the [EIRP] of each earth station, given that the saturation flux density is -122 dBW/m^2 .
- 19.35.** Derive, from fundamentals, the altitude of a *Geo-stationary* satellite.
- 19.36.** If a satellite has an orbiting time of 23 hours and 48 minutes, calculate the radius of orbit. Assume suitable data.
- 19.37.** Plot, using MATLAB/Mathematica/Octave, the trajectory of typical GEO satellite. Assume appropriate values for *apogee* and *perigee*.
- 19.38** In a satellite link the propagation loss is 200 dB, other losses are 3 dB, the receiver G/T is 12 dB/K, and $EIRP$ is 48 dBW. Calculate the received C/N for an FDM baseband consisting of 96 voice channels.
- 19.39.** Compute the effective input noise temperature of a receiver whose noise figure is 12 dB.
- 19.40.** A video signal has a *bandwidth*, BW of 4.6 MHz, and a deviation ratio of 2.56. Calculate the system BW required. Also calculate the SNR for $C/N=25$ dB.
- 19.41.** An FM system has a receiver threshold of 18 dB. How much received carrier power and RF bandwidth is needed to transmit a 4 kHz baseband signal with a demodulated SNR of 40 dB? Consider $N_0=10^{-10} \text{ W/Hz}$.



Fiber-optic Communications

20.1 Introduction

Optical fibers are increasingly replacing wire transmission lines in communications systems. Such optical fiber lines offer several important advantages over wire lines. First, since light is effectively the same as radio frequency radiation, but at a very much higher frequency (about 300 THz, or 3,000,000 GHz), the information-carrying capacity of a fiber is very much greater than for microwave radio systems. Next, the material used in fibers is silica glass, or silicon dioxide, which is one of the most abundant materials on earth, resulting in much lower material costs than with wire lines. The fibers are not electrically conductive, so they may be used in areas where isolation from electrical and electromagnetic interference is a problem. With the much higher information capacities, multiple channel routes using optic fibers can be compressed into much smaller cables, greatly reducing congestion in overcrowded cable ducts.

With present technology, fiber-optic communications systems are still more expensive than equivalent wire or radio systems, but this situation is changing rapidly. Fiber-optic systems will rapidly become competitive with other systems in price and eventually replace them.

20.2 Principles of Light Transmission in a Fiber

Propagation Within a Fiber

When light enters one end of a glass fiber under the right conditions, most of the light will propagate, or move, down the length of the fiber and exit from the far end. A small part of the light will escape through the side walls of the fiber, and some will also be lost due to internal absorption, but a portion of the light will be contained and guided to the far end. Such a fiber is called a *light pipe* or *light guide*.

The propagation of light in a fiber can be understood from an analysis process called *geometric ray tracing*, in which the paths of individual rays are geometrically traced along the guide path.

Light stays inside the fiber because it is totally reflected by the inside surface of the fiber. Light entering the end of the fiber at a slight angle to the axis follows a zigzag path through a series of reflections down the length of the fiber. *Total internal reflection* at the fiber wall can occur only if two conditions are met. The first is that the glass inside the fiber core must have a slightly higher index of refraction n_1 than the index of refraction n_2 of the material (cladding) surrounding the fiber core. The second is that the light must approach the wall with an angle of incidence ϕ (between the ray path and the normal to the fiber wall) that is greater than the critical angle ϕ_c , which is defined as

$$\sin \phi_c = \frac{n_2}{n_1} \quad (20.2.1)$$

The reflected ray will leave the fiber wall at the same angle ϕ as it struck the wall before reflection. These conditions are illustrated in Fig. 20.2.1(a).

Refraction occurs when the angle of incidence is *less than* the critical angle. A ray approaching the inside of the core wall at an angle of incidence that is less than the critical angle will pass through the wall into the cladding region by refraction and become lost. This is illustrated in Fig. 20.2.1(b).

In Fig. 20.2.1(c), a ray of light enters the core n_1 through the end face from the n_0 launch region with an angle of incidence θ_0 and leaves the interface at an angle of refraction θ_1 , which is smaller than the angle of incidence. It is bent closer to the normal to the interface. Snell's law says that the incidence angle θ_0 is related to the refraction angle θ_1 by the relationship

$$n_0 \sin \theta_0 = n_1 \sin \theta_1 \quad (20.2.2)$$

Figure 20.2.2 shows a longitudinal cross section of the launch end of a fiber with a ray entering it. The core of the fiber has a refractive index n_1 and is surrounded by a cladding of material with a lower refractive index n_2 . Light is launched into the end of the fiber from a launch region with a refractive index n_0 . If the launch region is air, then $n_0 = 1$. The ray enters with an angle of incidence to the fiber end face of θ_0 to the fiber axis (the normal to the end face). This particular ray enters the core at its axis point A and proceeds at the refraction angle θ_1 from the axis. It is then reflected from the core wall at point B at the internal incidence angle ϕ .

The entry incidence angle θ_0 can be related to the internal reflection angle ϕ by the right triangle ABC and Snell's law as follows. First, from the triangle ABC

$$\theta_1 = 90^\circ - \phi \quad (20.2.3)$$

Now substituting from Snell's law,

$$\sin \theta_0 = \frac{n_1}{n_0} \sin(90^\circ - \phi) = \frac{n_1}{n_0} \cos \phi \quad (20.2.4)$$

As long as the light enters the fiber at an incident angle such that the internal reflection angle ϕ is not less than the critical angle ϕ_c , then the light will be contained within the fiber and will propagate to the far end by a series of reflections. However, if the internal reflection angle is less than the critical angle, the light will be refracted into the cladding and lost. The critical value of the entrance incident angle θ that must not be exceeded is found as follows.

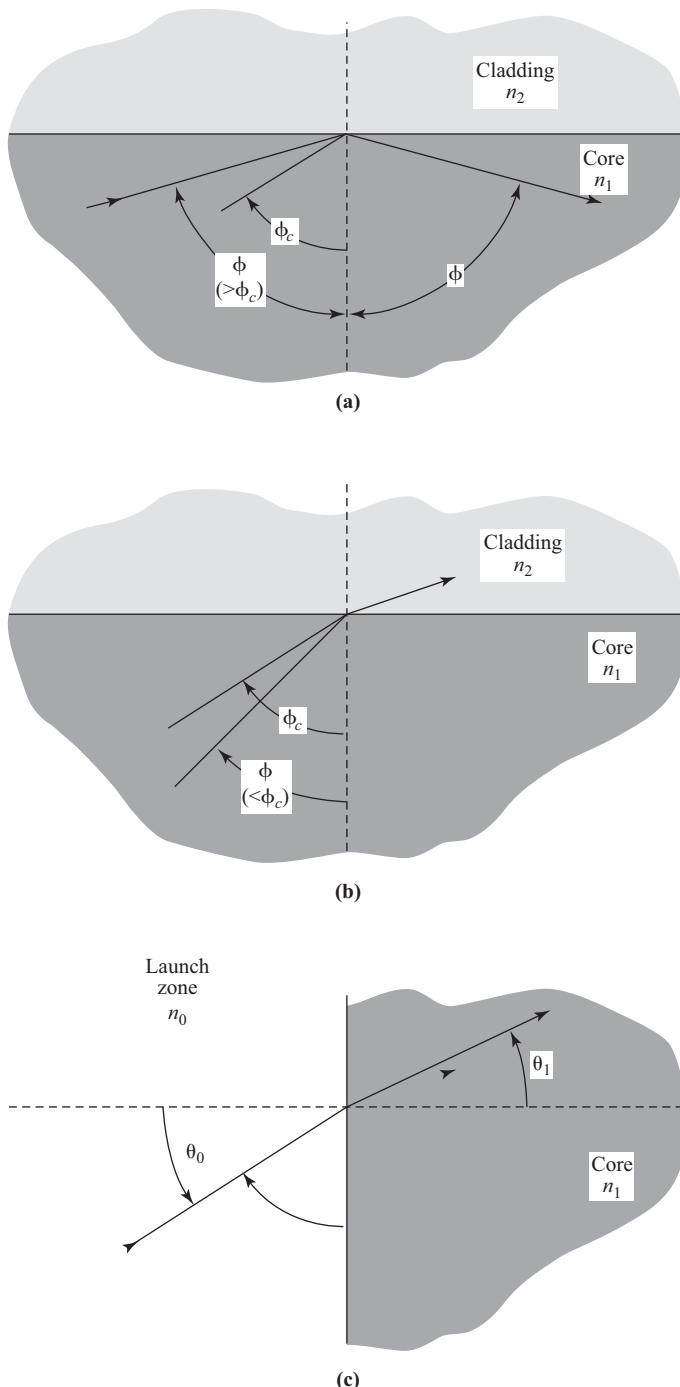


Figure 20.2.1 (a) Reflection from the inside of the core wall. (b) Ray escaping through the core wall by refraction. (c) Ray entering the fiber end by refraction.

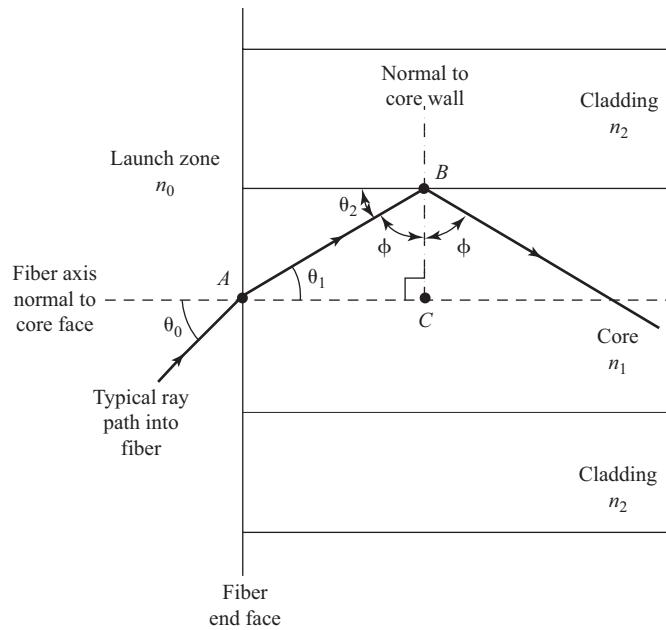


Figure 20.2.2 Path of a typical light ray launched into a fiber.

Equation (20.2.1) is represented graphically in Figure 20.2.3 by the right triangle DEF through the definition for the sine of an angle. Applying Pythagoras' theorem and the cosine definition gives

$$\cos \phi_c = \frac{\sqrt{n_1^2 - n_2^2}}{n_1} \quad (20.2.5)$$

Substituting Eq. (20.2.5) into Eq. (20.2.4) gives the maximum value of the external incidence angle for which light will propagate in the fiber as

$$\theta_0(\max) = \sin^{-1} \left(\frac{\sqrt{n_1^2 - n_2^2}}{n_0} \right) \quad (20.2.6)$$

This maximum angle is called the *acceptance angle* or the *acceptance cone half-angle*. Rotating the acceptance angle about the fiber axis as shown in Fig. 20.2.4 describes the *acceptance cone* of the fiber. Any light

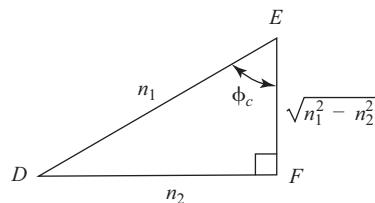


Figure 20.2.3 Pythagoras' theorem related to the critical angle.

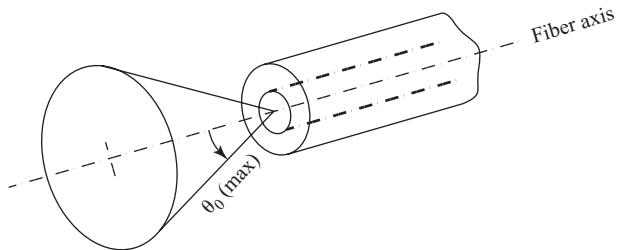


Figure 20.2.4 Acceptance cone obtained by rotating the acceptance angle about the fiber axis.

aimed at the fiber end within this cone will be accepted and propagated to the far end. Larger acceptance angles make easier launching.

The *numerical aperture* (NA) of the fiber is used as a figure of merit and is defined as the sine of the maximum acceptance angle, or

$$\text{NA} = \sin \theta_0(\text{max}) = \frac{\sqrt{n_1^2 - n_2^2}}{n_0} \quad (20.2.7)$$

If the light in the fiber is launched from air, as is often the case, $n_0=1$ and the numerical aperture becomes

$$\text{NA} \approx \sqrt{n_1^2 - n_2^2} \quad (20.2.8)$$

The normalized difference Δ between the indexes of the core and cladding is

$$\Delta = \frac{n_1 - n_2}{n_1} \quad (20.2.9)$$

Substituting this in Eq. (20.2.7) and noting that $n_1 \approx n_2$ for all practical fibers, the numerical aperture becomes

$$\text{NA} \approx \frac{n_1 \sqrt{2\Delta}}{n_0} \quad (20.2.10)$$

which if $n_0 = 1$ reduces to

$$\text{NA} \approx n_1 \sqrt{2\Delta} \quad (20.2.11)$$

It should be noted that the numerical aperture is effectively dependent only on the refractive indexes of the core and cladding materials and is not a function of the fiber dimensions.

EXAMPLE 20.2.1

An optic fiber is made of glass with a refractive index of 1.55 and is clad with another glass with a refractive index of 1.51. Launching takes place from air. (a) What numerical aperture does the fiber have? (b) What is the acceptance angle?

SOLUTION (a) By Eq. (20.2.9), the normalized difference between the indexes is

$$\Delta = \frac{n_1 - n_2}{n_1} = \frac{1.55 - 1.51}{1.55} = 0.0258$$

By Eq. (20.2.11), the numerical aperture is

$$NA \approx n_1 \sqrt{2\Delta} = 1.55 \sqrt{2 \times 0.0258} = \mathbf{0.352}$$

(b) By Eq. (20.2.7), the acceptance angle is

$$\theta_0(\max) = \sin^{-1} NA = \sin^{-1} 0.352 = \mathbf{20.6^\circ}$$

Fiber Index Profiles

The analysis in the previous section is based on a *step index profile* fiber, which is characterized by a core with a completely constant index of refraction n_1 throughout its bulk and a sudden transition of index to a lower value at the core wall.

Three types of step index fibers result depending on the material used to surround the core. The first case is that of an unclad fiber core surrounded by air with an index of refraction n_0 of unity. This combination results in a large difference of index between the core and the surrounding air and correspondingly large acceptance angles. However, small-diameter cores make mechanically weak fibers, so unclad fibers typically have core diameters in excess of 200 μm .

An *index profile* for a fiber is produced by plotting the index of refraction on the horizontal axis against the radial distance from the core axis on the vertical. The index profile of an unclad core is shown in Fig. 20.2.5(a).

The second case is that of a glass-clad core as shown in Fig. 20.2.5(b), where the core glass is surrounded by a concentric layer of cladding glass with a uniform index of refraction n_2 that is only slightly less than that of the core. The clad core may be used bare, or it may be enclosed in an opaque protective sheath. This fiber structure makes it possible to obtain very small diameter cores (down to 3 μm) without sacrificing mechanical strength and the low differences of index between core and cladding required for single-mode propagation.

The *W profile* fiber is a variation of the glass-clad fiber. In this case, shown in Fig. 20.2.5(c), the cladding layer n_2 is made only thick enough to obtain the desired guiding characteristic for the core. This first cladding layer is then surrounded by a second thicker glass cladding layer with an index n_3 with a value midway between n_1 and n_2 to produce the W-shaped profile. The second cladding layer tends to strip out modes that have leaked from the core into the first cladding layer. These fibers are used for low-loss, long-distance, single-mode communication links.

The third type of step index fiber uses a plastic-clad core. The shape of its profile is the same as that in Fig. 20.2.5(b), but the plastic-clad cable has higher losses than its glass-clad counterpart. It is less expensive to manufacture and finds many uses requiring short runs of fiber, such as plant instrumentation.

The profile of a *graded index* fiber is shown in Fig. 20.2.5(c). In this type of fiber, the material in the core is modified so that the index of refraction has a maximum value n_1 at the axis and lesser values falling off according to a carefully chosen profile with distance from the axis. The fiber is double clad to give an overall W-profile shape.

Propagation of light in the core of a step index fiber is characterized by light rays following a zigzag path of straight line segments. Light traveling through the uniform core material continues in a straight line

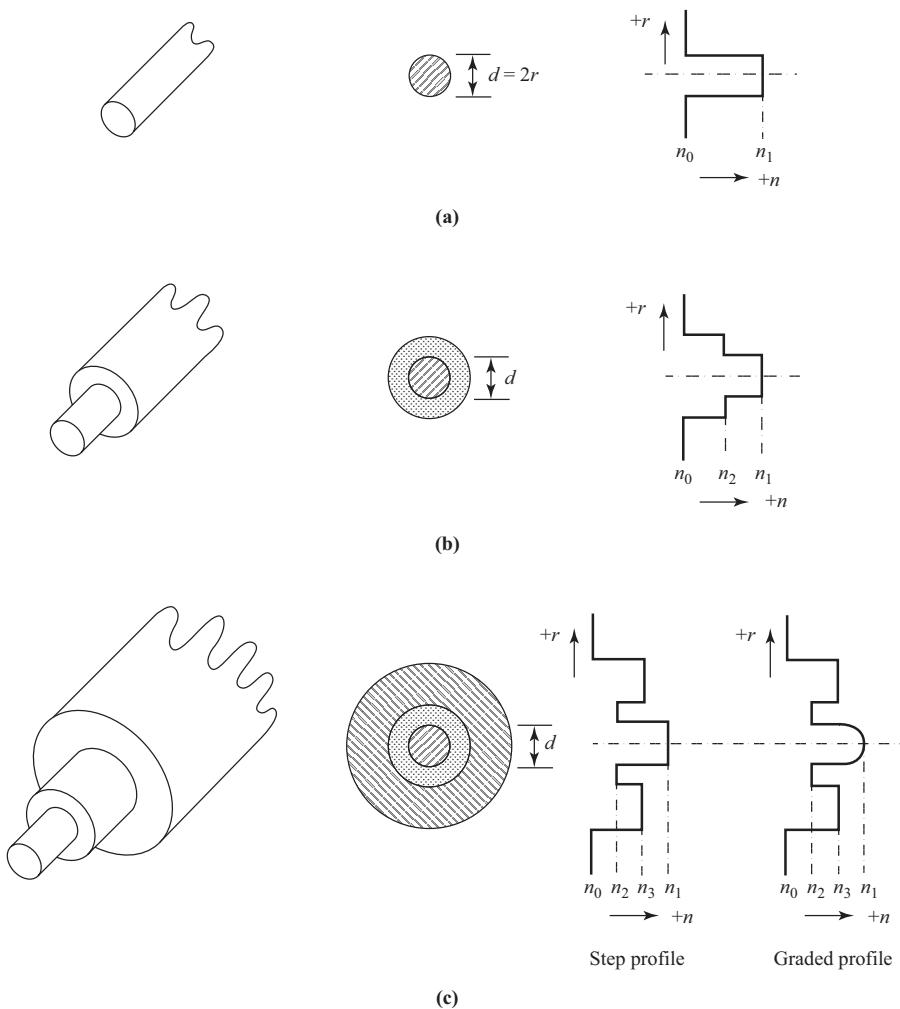


Figure 20.2.5 Core index profiles (a) for an unclad fiber, (b) for a single-clad step-index fiber, (c) for a double-clad fiber with a W-profile with either stepped or graded core profile.

until it encounters a reflecting or refracting surface such as the core-cladding interface. Figure 20.2.6(a) shows how two rays m_1 and m_2 launched within the acceptance cone are propagated down zigzag paths as they reflect off the core walls. Ray m_3 launched outside the acceptance cone escapes into the cladding and is lost.

Figure 20.2.6(b) illustrates how light propagates in a graded index fiber. The index is not uniform and decreases with distance from the axis, so light rays are curved toward the axis by refraction. Light rays periodically diverge and converge along the length of the fiber. Also, a somewhat larger acceptance cone results than with the step index fiber, and rays outside the acceptance cone will escape through the cladding.

By choosing the index grading profile carefully, it is possible to make a fiber in which the group velocities for all propagated ray paths average about the same, resulting in a large reduction of intermodal dispersion of transmitted pulses, making multimode fibers practical for longer runs.

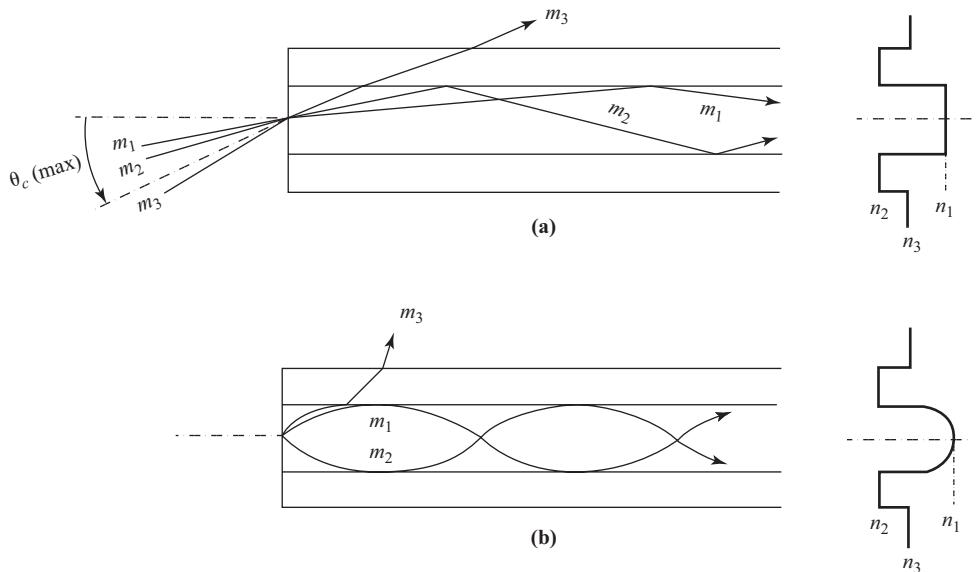


Figure 20.2.6 (a) Light ray propagation in a step-index fiber. (b) Propagation in a graded index fiber.

Graded index fibers may be made with a variety of different index grading profiles. One of the more popular profile functions is the *alpha profile function*, in which the index of refraction within the core is made to vary radially by the function

$$n(r) = n_1 \sqrt{1 - 2\Delta \left(\frac{2r}{d}\right)^\alpha} \quad (20.2.12)$$

where $n(r)$ is the core index at radius r from the core axis, n_1 is the index at the core axis, Δ is the normalized difference between core and cladding indexes [Eq. (20.2.9)], d is the core diameter, and α is the grading profile index number.

While any value of α between 1 and ∞ may be used, $\alpha = 2$, which produces a parabolic profile, is most often used for graded index fibers. The profile is relatively easy to manufacture reliably, and the value of $\alpha = 2$ simplifies several of the expressions involved in analyzing such a fiber. Figure 20.2.7 shows the index profiles plotted for graded index fibers with α 's of 1, 2, and ∞ . It should be noted that the infinite value of α corresponds to the step index fiber.

Modes of Propagation

Light propagates as electromagnetic waves in the same manner as do microwaves. It can be propagated in free space, or it can be guided in a duct in the same manner as microwaves. The frequencies of light are very much higher than those of microwaves. Microwaves occupy the frequency range from 3 to 100 GHz, with wavelengths between 10 cm and 3 mm. Visible light occupies a narrow range of wavelengths between 0.4 and 0.7 μm , or frequencies of 750 to 430 THz, which is six orders of magnitude higher than microwaves. Fiber optical communications presently use three bands of wavelengths in the infrared range, between 0.8 and 0.9 μm , between 1.2 and 1.3 μm , and between 1.4 and 1.5 μm . Modern silica fibers exhibit their lowest losses in these bands.

A plane electromagnetic wave propagates in free space as a transverse electromagnetic (TEM) wave as described in Appendix B. The TEM wave is characterized by having a component of its electric field and

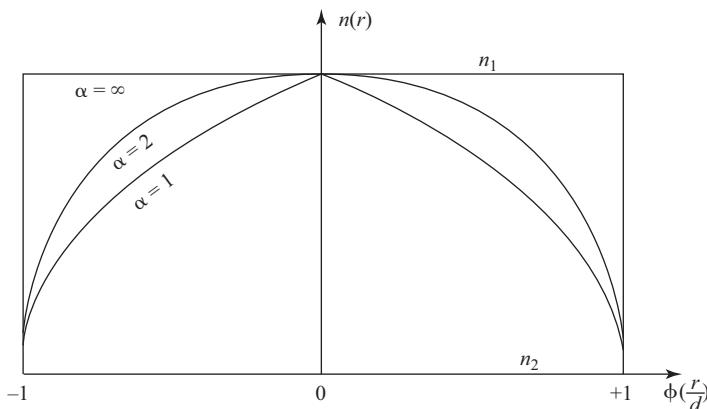


Figure 20.2.7 Normalized variation of index of refraction with radius for a graded index fiber.

a component of its magnetic field oriented at right angles to each other, and also each at right angles to the direction of propagation as shown by the vectors E_t and H_t in Fig. 20.2.8(a). The electric field vector E_t occupies the vertical axis (y), and the magnetic field vector H_t occupies the horizontal axis (x) of the field frontal plane. The fields (and the frontal plane) move in the direction of propagation (z) at the speed of light c . There are no electric or magnetic field components lying parallel to the direction of propagation; that is, $E_z = 0$ and $H_z = 0$.

When transverse electromagnetic waves are confined in a guide, they can propagate in several types of modes. Each of these mode types is characterized by having electric and magnetic field components lying normal to the direction of propagation, in a frontal plane that propagates (moves) in the z direction at the group velocity v_g but also has components lying parallel to the z direction.

One of these mode types, the TE mode, was described in Section 13.2 as supported in a metallic microwave guide. In this section, a ray analysis based on the interference between two interacting TEM waves within the guide was described. In this mode, the electric field vector E is at all points and times normal to the direction of propagation in the y direction, so that $E = E_t$. The magnetic field vector, however, has two components, H_t and H_z , which interact to form the traveling magnetic loop patterns. The $E-H$ frontal plane containing E and the complex H vectors shown in Fig. 20.2.8(b) corresponds to the frontal plane of one of the two interacting TEM waves, propagating at velocity c at the indicated angle to the z direction. However, the transverse components lie in the transverse plane, which move in the z direction at velocity v_g instead of at c .

In the transverse magnetic (TM) mode, the magnetic vector $H = H_t$ lies entirely in the transverse plane in Fig. 20.2.8(c), while the E vector is complex, containing a transverse component E_t and a longitudinal component E_z , so that $H_z = 0$ but E_z is finite. Again, the transverse plane propagates in the z direction at velocity v_g while the individual TEM fronts move at the angle to the z direction indicated by the $E-H$ plane at velocity c .

The TE modes and the TM modes in circular guides are described in the ray model as rays focused so that they pass through the fiber z axis on every reflection. These are called *meridional rays*. If these modes are to be excited, the transmitter must focus the injected light so it converges on the launch face of the fiber.

The *hybrid* modes (EH and HE) also propagate in cylindrical fibers. In fact, the HE_{11} mode is dominant and becomes the mode propagated in single-mode fibers. These modes are characterized by having both E_z and H_z field components in addition to the transverse components E_t and H_t . If the *transverse* electric field

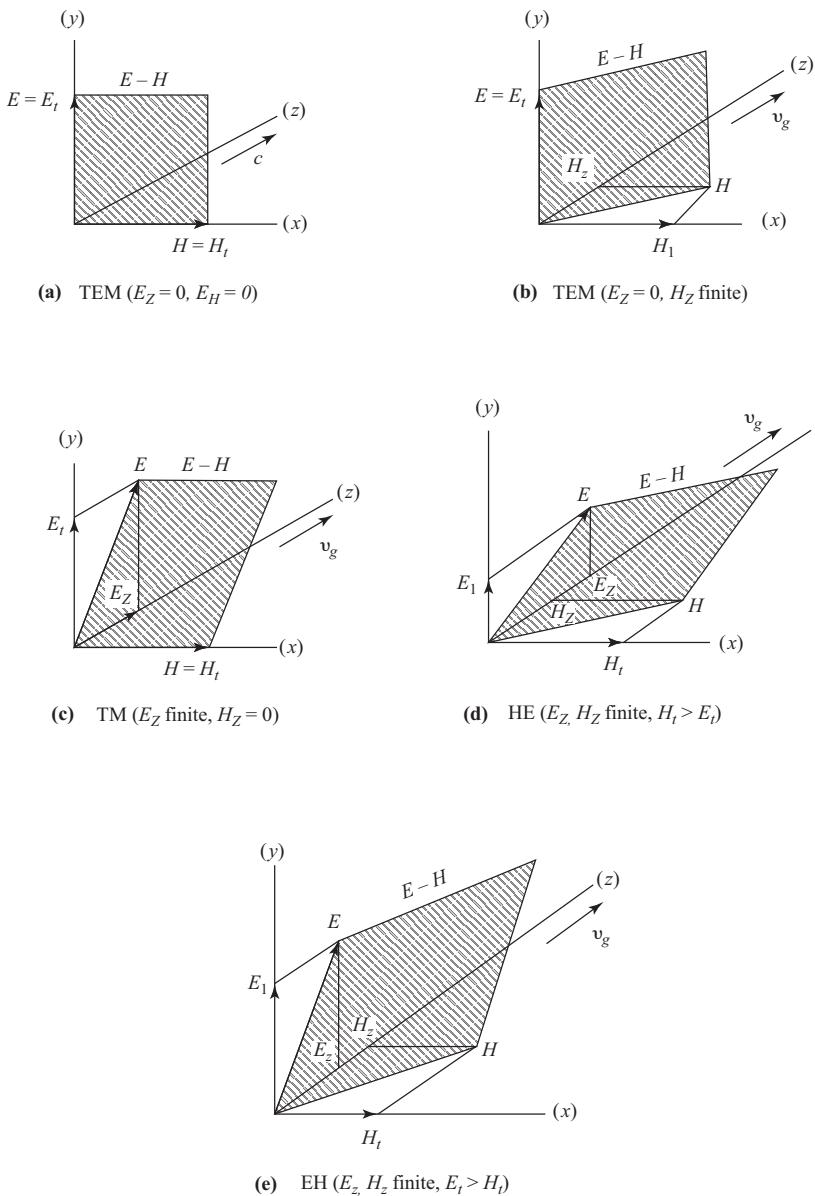


Figure 20.2.8 Electromagnetic wave modes. (a) TEM wave in space. (b) Guided TE mode. (c) TM mode. (d) HE mode. (e) EH mode.

component E_t is larger than the magnetic component H_t , the mode is called an EH mode. If H_t is greater than E_t it is called an HE mode. Figures 20.2.8(d) and (e) illustrate the $E - H$ plane orientation for the HE and EH modes.

The hybrid modes are characterized in the ray model by *skew rays*. These rays do not pass through the fiber axis, but spiral around the axis by a series of equal-spaced reflections in the manner of a corkscrew.

Number of Propagated Modes in Step-index Fibers

Within each of the classes of modes discussed above, only certain modes will propagate, and each of these will have a particular internal reflection angle ϕ with the normal to the fiber axis (or its complement angle ψ to the fiber wall) associated with it. Consider in Fig. 20.2.9 the case of a meridional light ray propagating by a series of reflections from the walls of a step index fiber. The light is assumed, to be monochromatic with a wavelength of λ and the fiber is assumed to have a core diameter d .

The ray enters reflection point A at angle ψ , is reflected and moves to B also at angle ψ and is reflected again at B and moves off at the same angle ψ . The path of the ray after the second reflection at B is parallel to the original path before A . The wavefront at point A is normal to the original path and is represented by the line AD . The wavefront at B must be parallel to the original wavefront at A and is represented by line BE . The path of the ray emerging from B is extrapolated back to meet the AD wavefront line to complete the right triangle ABD .

For the wavefront at BE to be supported, all rays must pass through this front with the same phase. For this to be, the distance DB between the two fronts AD and BE must be an integer number of wavelengths, a condition that can be satisfied by many values of the angle ψ .

Now examine the right triangle ABC formed by the reflected ray from A , the fiber wall, and the normal to the fiber axis. Where side AC is the fiber core diameter d , and side AB is the propagation distance between two successive reflection points on the ray path,

$$AC = AB \sin \psi \quad (20.2.13)$$

Therefore,

$$AB = \frac{d}{\sin \psi} \quad (20.2.14)$$

Then examine the right triangle ABD formed by the wavefront AD , the ray path BD and the reflected ray path AB . The angle ABD encloses an angle of 2ψ and the condition for propagation states that the distance between wave-fronts BD must be

$$BD = n\lambda, \quad \text{for } n = 1, 2, 3, \dots \quad (20.2.15)$$

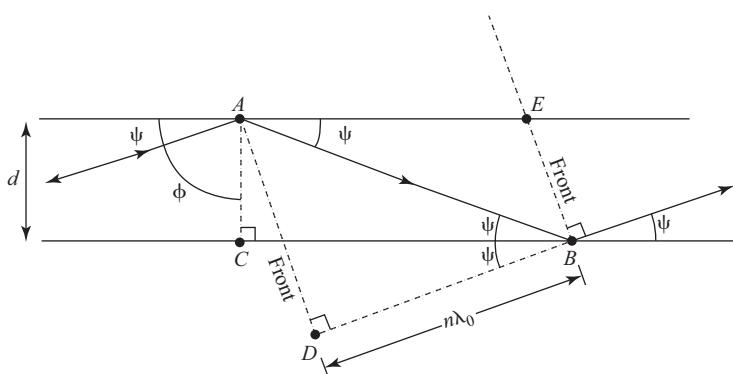


Figure 20.2.9 Condition for propagation of a meridional ray mode in a step-index fiber.

and from the triangle ABD ,

$$BD = AB \cos(2\psi) \quad (20.2.16)$$

which becomes on substitution from above,

$$d \cos(2\psi) = n\lambda \sin\psi, \quad \text{for } n = 1, 2, 3. \quad (20.2.17)$$

Only angles of ψ for which n is a positive integer will permit propagation of a wave in the fiber, and each such value will support one unique mode.

The value of n cannot be zero, since this would imply that there is no longitudinal separation between A and B , and the wave would reflect back and forth between the walls without progressing along the fiber. The upper limit of n for each type of mode is dictated by the acceptance angle at the entrance to the fiber and by the size of the core diameter.

The limits for meridional ray modes may be relatively easily computed, but for the skew modes this computation is very difficult and of limited usefulness. A different approach is used to approximate the number of modes supported by a given fiber. First, the *normalized frequency (cutoff parameter or V number)* is found for the fiber, given by

$$V = \pi \frac{d}{\lambda} \sqrt{n_1^2 - n_2^2} \approx \pi \frac{d}{\lambda} \text{NA} \quad (20.2.18)$$

This number is dependent only on the fiber characteristics and the wavelength of light being propagated.

When the number of modes is large, the number of modes a fiber will support is approximated by

$$n \text{ (modes)} \approx \frac{V^2}{2} \quad (20.2.19)$$

Each different mode type has a particular value of the normalized frequency number V below which that mode type is cut off. The first 12 of these, listed in ascending order, are given in Table 20.2.1. All modes except the first have normalized cutoff frequencies greater than unity. The HE_{11} mode has a normalized

TABLE 20.2.1 Cutoff V Numbers for the Lowest-order Modes a Fiber Will Support

Mode	Cutoff V Number
HE_{11}	0
$\text{TE}_{01}, \text{TM}_{01}$	2.405
HE_{21}	2.42
$\text{HE}_{12}, \text{EH}_{11}$	3.83
HE_{31}	3.86
EH_{21}	5.14
HE_{41}	5.16
$\text{TE}_{02}, \text{TM}_{02}$	5.52
HE_{22}	5.53

cutoff frequency of zero, indicating that it will guide all frequencies. The next modes, TE₀₁ and TM₀₁, propagate for any frequencies above that for which V = 2.405. The number of modes supported then can be increased by increasing the light frequency (lower wavelengths) or by increasing the core diameter of the fiber.

EXAMPLE 20.2.2

The fiber in Example 20.2.1 has a core diameter of 50 μm and is used at a light wavelength of 0.80 μm. Find its V number and the approximate number of modes it will support.

SOLUTION

The diameter/wavelength ratio is

$$\frac{d}{\lambda} = \frac{50 \text{ } \mu\text{m}}{0.8 \text{ } \mu\text{m}} = 62.5$$

From Example 20.2.1, the numerical aperture NA=0.352. Substituting in Eq. (20.2.18) gives the V number as

$$V = \pi \frac{d}{\lambda} \text{NA} = \pi \times 62.5 \times 0.352 = 69.1$$

Only modes with cutoff V numbers below this value will propagate. Their number is approximately

$$N \text{ (modes)} \approx \frac{V^2}{2} = \frac{69.1^2}{2} = 2390$$

This is truly a multimode fiber.

EXAMPLE 20.2.3

Another fiber has a core diameter of 5 μm and operates with infrared light at 1.3 μm. It has a numerical aperture of 0.35. Find the number of modes it will support.

SOLUTION

The V number is found as

$$V \approx \pi \frac{d}{\lambda} \text{NA} = \pi \times \frac{5}{1.3} \times 0.35 = 4.23$$

This is a small number within the range of Table 20.2.1, so the approximation is not valid. From the table it is seen that six modes have cutoff V numbers less than 4.23, so the fiber will support those six modes.

Graded index fibers generally will not support as many modes as the corresponding step index fibers. As noted above, an alpha profile is usually used for grading the fiber index. When α = is used, a step index fiber results, the normalized frequency parameter V_λ is given by Eq. (20.2.18) and the number of supported modes N_λ is given by Eq. (20.2.19). When a finite value of alpha (commonly α = 2) is

used, then the number of supported modes for a large number of modes is reduced according to the expression

$$N_\alpha = N_\infty \frac{\alpha}{\alpha + 2} \quad (20.2.20)$$

Graded index fibers are almost always used as multimode fibers. A carefully designed graded index fiber will give performance almost as good as a single-mode fiber for the same application.

EXAMPLE 20.2.4

The fiber used in Example 20.2.3 is redesigned as a graded index fiber with a grading profile index of 2, but with the same dimensions and indexes of refraction. Find the number of modes it will support.

SOLUTION From Example 20.2.3, the normalized cutoff frequency is

$$V_\infty = 69.1$$

and the number of modes supported as a step index fiber is

$$N_\infty = 2390$$

For the graded index fiber with $\alpha = 2$, the number of modes supported becomes, by Eq. (20.2.20),

$$N_\alpha = N_\infty \frac{\alpha}{\alpha + 2} = 2390 \times \frac{2}{2 + 2} = \mathbf{1195}$$

Single-mode Propagation in Step-index Fibers

A step-index fiber will become a single-mode fiber if its normalized frequency V can be made less than 2.405. As noted in Table 20.2.1, for V greater than 2.405, the TE₀₁ and TM₀₁ and HE₁₁ modes will be supported, while for V less than 2.405 only the HE₁₁ mode will be supported.

Examining Eq. (20.2.18), it can be seen that the V number can be reduced by reducing the core refractive index n_1 or by reducing the normalized difference between core and cladding indexes, Δ , or by reducing the diameter of the core d , or by increasing the wavelength of light used λ , or by a combination of these.

Practically, fibers with cores as small as about 5 μm can be made, which gives a d/λ ratio of about 4 for 1.3- μm light. Normalized index differences as low as 0.0001 have been obtained in practice, but these low differences result in extremely small acceptance angles. The result is that only high-quality laser sources that produce a very focused beam of nearly monochromatic light can be used for single-mode operation. Furthermore, because of the small core diameter and acceptance angle, it is especially difficult to get good coupling between the fiber and its transmitter or receiver. However, because of the superior transmission characteristics (low loss and low dispersion), such fibers are extensively used for long-distance applications that must have a minimum of repeater stations, such as submarine cables.

EXAMPLE 20.2.5

A single-mode fiber is made with a core diameter of $10 \mu\text{m}$ and is coupled to a laser diode that produces $1.3\text{-}\mu\text{m}$ light. Its core glass has a refractive index of 1.55. (a) Find the maximum value required for the normalized index difference, (b) Find the refractive index required for the cladding glass, (c) Find the fiber acceptance angle.

SOLUTION (a) The diameter-to-wavelength ratio is

$$\frac{d}{\lambda} = \frac{10}{1.3} = 7.69$$

The numerical aperture NA for cutoff is found by Eq. (20.2.18) to be

$$\text{NA(max)} = \frac{V(\text{max})}{\pi(d/\lambda)} = \frac{2.405}{\pi \times 7.69} = 0.995$$

Now the normalized index difference is found by Eq. 20.2.11 to be

$$\Delta = \frac{1}{2} \left(\frac{\text{NA}}{n_1} \right)^2 = \frac{1}{2} \left(\frac{0.955}{1.55} \right)^2 = \mathbf{0.0206}$$

(b) The cladding index required is, by Eq. (20.2.9),

$$n_2 = n_1(1 - \Delta) = 1.55(1 - 0.00206) = \mathbf{1.547}$$

(c) By Eq. (20.2.7), the maximum acceptance angle for the fiber is

$$\theta_0(\text{max}) = \sin^{-1}(\text{NA}) = \sin^{-1}(0.09950) = \mathbf{5.712^\circ}$$

20.3 Losses in Fibers

Rayleigh Scattering Losses

The glass in optical fibers is an amorphous (noncrystalline) solid that is formed by allowing the glass to cool from its molten state at high temperature until it freezes. While it is still plastic, the glass is drawn out under tension into its long fiber form. During this forming process, submicroscopic variations in the density of the glass and in doping impurities are frozen into the glass and then become reflecting and refracting facets to scatter a small portion of the light passing through the glass, creating losses. While careful manufacturing techniques can reduce these anomalies to a minimum, they cannot be totally eliminated.

It is also found that the losses induced because of the scattering vary inversely with the fourth power of the light wavelength used, so their effects are less than about 0.3 dB/km at a wavelength of $1.3 \mu\text{m}$. Figure 20.3.1 shows the intrinsic minimum Rayleigh scattering losses of silica fibers plotted against wavelength over the usable portion of the spectrum from 0.7 to $1.6 \mu\text{m}$.

Absorption Losses

Three different mechanisms contribute to absorption losses in glass fibers. These are ultraviolet absorption, infrared absorption, and ion resonance absorption.

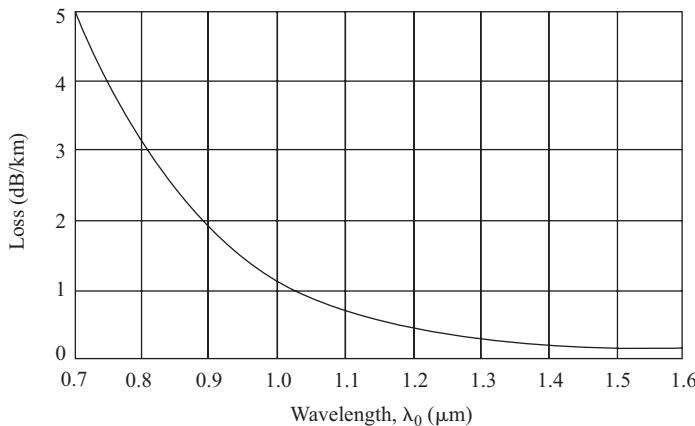


Figure 20.3.1 Rayleigh scattering losses in silica fibers.

Ultraviolet absorption takes place because, for pure fused silica, valence electrons can be ionized into conduction by light with a center wavelength of about $0.14 \mu\text{m}$, corresponding to an energy level of about 8.9 eV. The energy for this ionization is drawn from the light fields being propagated and constitute a transmission loss. The absorption loss does not only occur at this fixed wavelength, but occurs over a broad band that extends up into the visible part of the spectrum, with losses decreasing as wavelength increases. This absorption *tail* becomes negligible in the 1.2 - to $1.3\text{-}\mu\text{m}$ band.

The introduction of impurities such as germanium dioxide to modify the refractive index causes some increase in the magnitude of the UV absorption tail because of an upward shift of the wavelength at which peak absorption takes place. Figure 20.3.2 shows that for most fibers the UV absorption loss at $1.2 \mu\text{m}$ remains less than 0.1 dB/km .

Infrared absorption takes place because photons of light energy are absorbed by the atoms within the glass molecules and converted to the random mechanical vibrations typical of heating. This IR absorption also exhibits a main spectral peak that for silicon occurs at $8 \mu\text{m}$, with minor peaks at 3.2 , 3.8 , and $4.4 \mu\text{m}$. The peaks are broad and tail off into the visible part of the spectrum. Losses at $1.5 \mu\text{m}$ are typically less than 0.5 dB/km , as shown in Fig. 20.3.2.

Minute quantities of water molecules trapped in the glass during manufacture contribute OH-ions to the material. These ions also absorb energy at peaks of 0.95 , 1.25 , and $1.39 \mu\text{m}$, with the main peak at $1.39 \mu\text{m}$, as shown in Fig. 20.3.2. The water content of the glass must be kept below 0.01 ppm to prevent these peaks from spreading out and merging to eliminate the low-loss windows between peaks.

The presence of other impurities in the glass may also create unacceptable losses within the usable portion of the spectrum. Iron, copper, and chromium must especially be avoided. The same type of zone-refining techniques used to purify silicon for integrated circuits is used to make glass fibers.

Leaky Modes

For meridional modes in which all rays pass through the core axis, if the axial angle of incidence is greater than critical at each reflection point, it will be reflected and propagate. If the angle of incidence is less than critical, the rays of the mode will be refracted out of the core and lost. Such modes are either propagated or completely lost.

For skew modes, however, each incident ray has two components of its angle of incidence, one axial and the other radial. If both the radial and axial components of the angle of incidence are less than critical as

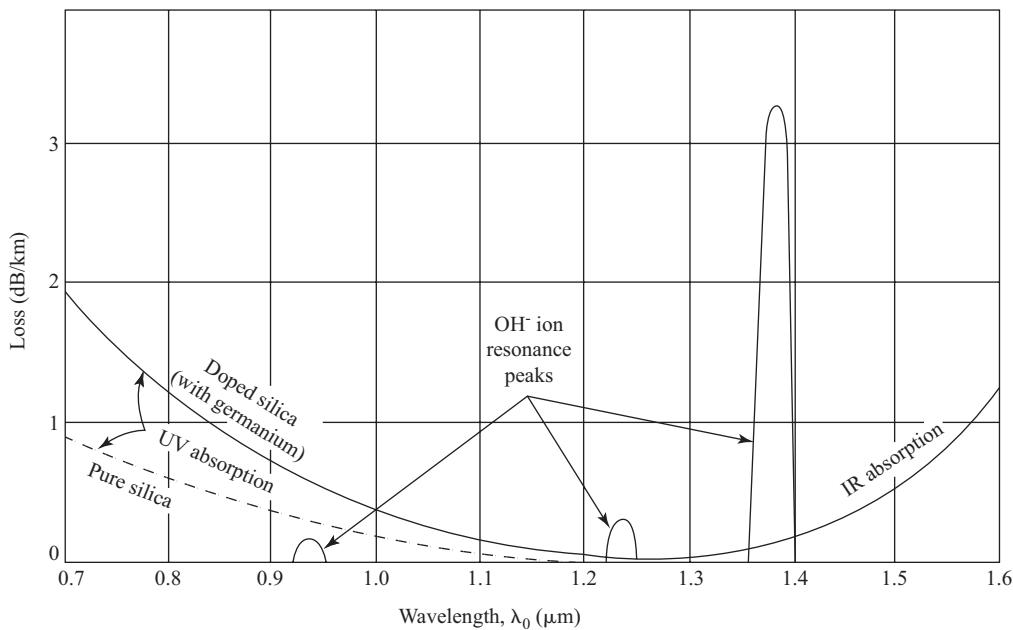


Figure 20.3.2 Absorption loss effects in fused silica glass fibers.

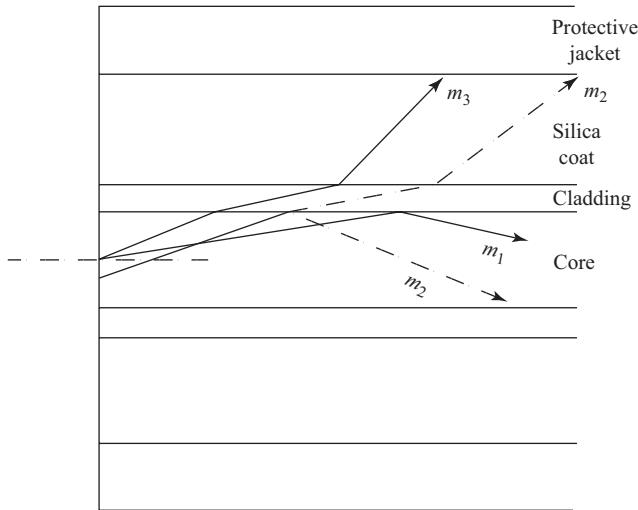


Figure 20.3.3 Leaky mode removed by an additional silica cladding.

In higher-order modes, the mode will be totally refracted into the cladding and lost. If both the radial and axial components are greater than critical as in lower-order modes, then the mode will be totally reflected and propagated within the core.

However, for intermediate-order skew modes, it is possible for the axial component of the incidence angle to be greater than critical while the radial component is less than critical. In this case, some of the mode rays will be refracted into the cladding and lost while the rest are propagated. These modes are called *leaky modes*.

In the W profile fibers, the cladding layer is only very thin, and this is in turn surrounded by a second cladding with a higher index of refraction (but still less than that of the core). This second cladding acts to remove the leaky modes that escape from the core into the first cladding by ensuring that they are refracted when they reach the second cladding interface and get absorbed by the opaque protective sheath. This stripping of the leaky modes from the cladding is done because, if they remain in the cladding and reach the receiver, they will contribute to the dispersion of the signal. Ray m_2 in Fig. 20.3.3 illustrates a leaky mode, while m_1 propagates and m_3 is completely removed.

The leaky modes introduced at the transmitting end of the fiber usually contain only a small fraction of the total guided power, and these are rapidly attenuated near the transmitter end of the fiber. If the fiber is uniform over its length, this leakage loss is fixed and occurs only once. However, if the fiber is spliced or bent or has changes of diameter along its route, each occurrence will cause more leakage to occur in the section following each discontinuity because of mode coupling into the leaky modes.

Mode Coupling Losses

Power that has been launched successfully into a propagating mode may be later coupled into a leaky or radiating mode because of some discontinuity in the fiber. Any variations in the distribution of impurities within the core can cause internal refractions to occur. Any variation in diameter because of splices or bending can cause a shift in the angle of incidence at reflection points. Any of these mechanisms can cause energy to be shifted from a fully propagating mode into one of the leaky modes and ultimately lost through leakage.

Bending Losses

Two types of bending can affect a fiber. These are microbending and large radius bending. *Microbending* is a microscopic bending of the core of the fiber that may result from different thermal contraction between core and cladding or because of kinking during handling. These microbends act as scattering facets within the fiber and cause energy from fully propagated modes to be cross-coupled into leaky modes and subsequently lost. Since microbends are randomly distributed over the length of the fiber, losses resulting from them are uniformly distributed and a total figure for the fiber can be obtained. Care in manufacture and handling will minimize microbending losses.

Large radius bending is caused by several things. Fibers are generally combined in multifiber cables, where they are spiraled about a central cable core. The spiral creates a constant radius bend that extends the full length of the cable. Aerial cables are hung from poles, and each pole hanger introduces a short, relatively sharp bend in the fiber. Buried ducts or ducts in buildings may be required to negotiate relatively sharp turns. These large radius bends also introduce loss by mode coupling into leaky modes.

Modes that are fully guided in straight sections of fiber are either only partially guided or not guided at all over the curved portion of fiber. Consider the ray in Fig. 20.3.4 as it approaches a bend in the fiber core. It is fully guided before it reaches point A in the straight section. It is reflected from point A at the axial angle ϕ_A , and if the core were not bent, it would arrive at the next reflection at point B_1 . In the bend, however, the ray encounters the fiber wall early at the outer edge of the bend at a smaller angle of incidence ϕ_B . If ϕ_B is less than critical, a portion of the energy from the mode will be lost into the cladding. Sharper bends will cause more lower-order modes to be lost, and care must be taken during installation to be sure that bends do not have radii less than some minimum.

Mode stripping is used to remove those modes that are near the cutoff point and may contribute to leakage losses. This technique makes use of bending losses and consists of several meters of fiber wrapped on a relatively small radius spool at the transmitter end. Higher-order modes near cutoff are removed through bending losses, leaving only the lower-order, fully propagated modes in the core. Subsequent bends or discontinuities then cannot cause cross coupling into the leaky modes further down the fiber.

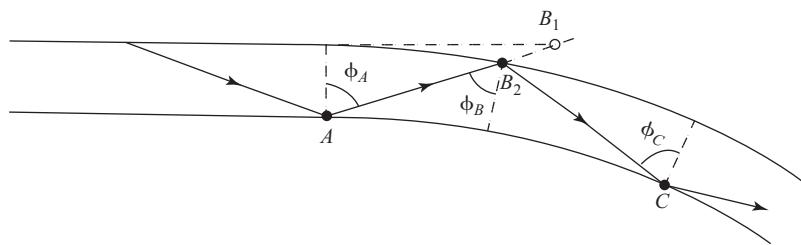


Figure 20.3.4 Ray propagation in a bent fiber.

Combined Fiber Losses

Four types of losses that must be reduced to a minimum during manufacture are Rayleigh scattering, material absorption, leaky modes, and scattering. Rayleigh scattering and material absorption are the predominant factors, and every effort is made to minimize these during manufacture.

Figure 20.3.5 shows the losses in a typical multimode fiber as a function of wavelength, with the contributions of the components shown for comparison. This fiber has intermediate loss levels compared to some fibers, which may be as high as 20 dB/km at 0.8- μm wavelength, and good quality silica fibers, which will have losses below 1 dB/km at 1.3 μm .

All fibers are characterized by a loss spectrum curve of this general shape, although the actual loss value and peaking wavelengths will vary depending on the type of fiber. In the region between 0.8 and 1.3 μm the losses are dominated by the Rayleigh scattering effect, which can be controlled to a degree with

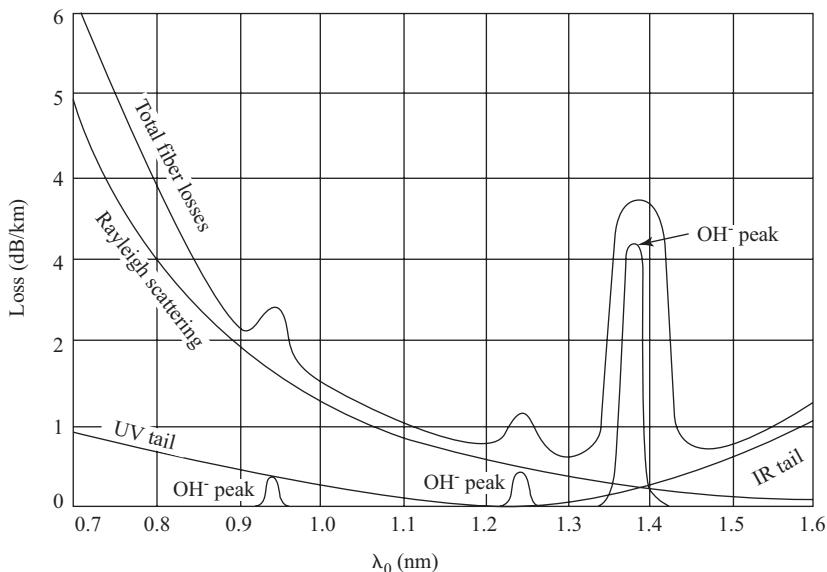


Figure 20.3.5 Total loss spectrum for an optical fiber.

careful manufacturing. Even at best, however, they will be an order of magnitude above the UV losses, which become negligible above about 1.2 μm.

Above 1.3 μm the IR tail becomes significant. A crossover between the Rayleigh scattering curve and the IR tail in silica fibers occurs at about 1.55 μm, where a minimum loss of less than 0.2 dB/km can be realized. Fiber-optic links can operate satisfactorily with total fiber losses of not more than about 30 dB, which means that unrepeated fiber links in excess of 100 km are practical. New fiber materials based on fluorides may move the IR tail further up the scale, making it possible to realize fibers with losses as low as 0.001 dB/km at wavelengths somewhere above 1.6 μm.

The OH⁻ peaks are reduced in amplitude and width by careful manufacturing techniques designed to eliminate water molecules. Good quality fibers presently reduce all these peaks except the 1.4 μm peak to negligible levels.

The result with the silica fibers then is that there are three usable spectrum windows within which optical communications systems may operate. These are at 0.9, 1.2, and 1.5 μm. Some fiber-optic links have already been *wavelength division multiplexed* (WDM) using three light sources of different wavelengths to more fully utilize the fiber's capacity.

EXAMPLE 20.3.1

A silica fiber has measured losses of 2 dB/km at 0.9 μm and 0.3 dB at 1.5 μm. If a total fiber loss of 25 dB can be tolerated in a single link, determine appropriate repeater spacings for (a) operation at 0.9 μm and (b) at 1.5 μm.

SOLUTION

(a) At 0.9 μm wavelength, repeater distance z is

$$z = \frac{A_{\max}}{A} = \frac{25 \text{ dB}}{2 \text{ dB/km}} = \mathbf{12.5 \text{ km}}$$

(b) At 1.5 μm wavelength,

$$z = \frac{A_{\max}}{A} = \frac{25 \text{ dB}}{0.3 \text{ dB/km}} = \mathbf{83 \text{ km}}$$

20.4 Dispersion

Effect of Dispersion on Pulse Transmission

A pulse of light with a given width and amplitude transmitted into one end of a fiber should theoretically arrive at the far end with its shape and width unchanged and only its amplitude reduced by losses. However, several effects contribute to time dispersion of the pulse during transmission, which tend to widen out and flatten it, further reducing its amplitude. Besides reduced amplitude, the widening of the pulse may cause it to overlap adjacent pulses, causing intersymbol interference and reducing the upper limit on the pulse transmission rate. At low bit transmission rates the required repeater spacing will be

dictated by the loss limits for the fiber. However, at some higher rate the dispersion effects will become predominant and further reduce the repeater spacing. The product of bandwidth (the maximum allowable transmission rate) and dispersion, or *bandwidth-dispersion product* (BDP), is used as a quality factor for the fiber.

Three separate dispersion mechanisms exist in a fiber. These are inter-modal dispersion, material or chromatic dispersion, and waveguide dispersion.

Intermodal Dispersion

Each mode that a step-index fiber supports has a different effective group velocity, even though the phase velocity in each ray path may be identical. This occurs because the total path followed by guided rays is zigzag in nature and has a different length for each mode. The shortest path coincides with the axial length of the fiber for the HE₁₁ mode, while the longest path occurs for the mode nearest the cutoff limit. A pulse coupled into several modes at the transmitting end becomes several pulses traveling in the several modes at different velocities, which arrive at the receiver at slightly different times. The received pulse is the summation of these mode pulses, each delayed by a different time. Figure 20.4.1 shows the effect of dispersion on an ideal pulse transmitted by two modes. The effects of loss are omitted, and only dispersion is accounted for. The received pulse is the summation of the two received modal pulses and has a lower amplitude and wider pulse width than would be the case without dispersion.

The shortest delay is for a wave propagating parallel to the axis in the HE₁₁ mode. The longest delay occurs for a wave propagating in the highest-order supported mode, just below cutoff, with an angle of incidence just slightly higher than critical. The difference in delay times Δ_t between the maximum and minimum propagation delay times can be readily derived for the step-index fiber as follows.

Let the total fiber length be z . Since it is convenient to state losses and delays in terms of a standard unit length, let z be 1 km. Figure 20.4.2 shows two meridional rays of different modes following their zigzag paths down a fiber at incident angles ϕ . The total zigzag path length for each ray is found as

$$z_t = \frac{z}{\sin \phi} \quad (20.4.1)$$

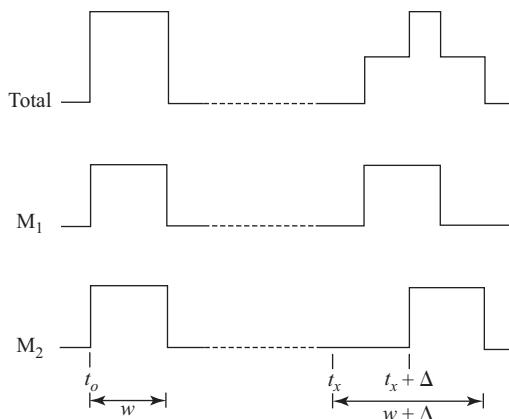


Figure 20.4.1 Effect of intermodal dispersion on a transmitted pulse.

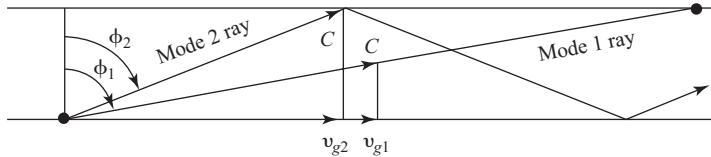


Figure 20.4.2 Group velocities for two modes.

In the lowest-order mode, the maximum angle of incidence is 90° and in the highest-order mode it is almost critical, or, from Fig. 20.2.3,

$$\phi(\max) = \phi_c = \sin^{-1} \left(\frac{n_2}{n_1} \right) \quad (20.4.2)$$

Now the shortest path is found to be

$$z_t(\min) = \frac{z}{\sin \phi(\max)} = \frac{z}{\sin 90^\circ} = z \quad (20.4.3)$$

and the longest path is

$$z_t(\max) = \frac{z}{\sin \phi(\min)} = z \frac{n_1}{n_2} \quad (20.4.4)$$

The maximum time dispersion now becomes

$$\Delta z = z_t(\max) - z_t(\min) = z \left(\frac{n_1}{n_2} - 1 \right) \quad (20.4.5)$$

Substituting from Eq. (20.2.9) then gives the dispersion in terms of the normalized index of refraction difference as

$$\Delta z = z \frac{\Delta}{1 - \Delta} \quad (20.4.6)$$

Since the light rays in the fiber travel through a dielectric medium with a dielectric constant ϵ_r , which is greater than unity, the rays travel more slowly than they would in free space. The relative permeability μ_r in the dielectric is about unity. Now, from equation (B.11), the phase velocity in the dielectric is

$$v_p = \frac{c}{\sqrt{\mu \epsilon}} = \frac{1}{\sqrt{\mu_0 \epsilon_0} \sqrt{\mu_r \epsilon_r}} \quad (20.4.7)$$

The speed of light is found when $\epsilon_r = 1$ and $\mu_r = 1$ to be

$$c = \frac{1}{\sqrt{\mu_0 \epsilon_0}} \quad (20.4.8)$$

Substituting c and $\mu_r = 1$ into Eq. 20.4.7 gives the phase velocity in the dielectric as

$$v_p = \frac{c}{\sqrt{\epsilon_r}} \quad (20.4.9)$$

which is the same as the velocity derived for transmission lines in Chapter 12.

It can be shown that the dielectric constant is related to the refractive index as

$$\epsilon = n^2 \quad (20.4.10)$$

so that in the glass core with a refractive index n_1 the velocity becomes

$$v_p(\text{glass}) = \frac{c}{n_1} \quad (20.4.11)$$

Finally, the dispersion in a step-index fiber is the maximum difference of delay between the lowest- and highest-order modes, found by dividing the path length difference by the phase velocity in the glass core as

$$\Delta_t = \frac{\Delta_z}{v_p(\text{glass})} = \frac{n_1 z}{c} \frac{\Delta}{1 - \Delta} \quad (20.4.12)$$

which is usually expressed in units of nanoseconds per kilometer (ns/km). Note that this intermodal dispersion is a characteristic of the fiber and is not affected by the wavelength of light used. It should also be noted that intermodal dispersion cannot occur in single-mode fibers.

EXAMPLE 20.4.1

For the step-index fiber of Example 20.2.1, find the intermodal dispersion per kilometer of length and the total intermodal dispersion in a 12.5-km length of fiber.

SOLUTION From Example 20.2.1, $n_1 = 1.55$ and $\Delta = 0.0258$.

(a) For a 1-km length of fiber, the intermodal dispersion is

$$\Delta_t (\text{per km}) = \frac{n_1 z}{c} \frac{\Delta}{1 - \Delta} = \frac{1.55 \times 1000 \text{ m/km} \times 0.0258}{0.3 \text{ Gm/s} \times (1 - 0.0258)} = 136.8 \text{ ns/km}$$

(b) For a length of 12.5 km, the total intermodal dispersion becomes

$$\Delta_t (\text{imd}) = \Delta_t (\text{per km}) \times z (\text{km}) = \frac{136.9 \text{ ns/km} \times 12.5 \text{ km}}{1000 \text{ ns}/\mu\text{s}} = 1.71 \mu\text{s}$$

Multimode graded index fibers have a much lower intermodal dispersion than do corresponding step-index fibers. This is so because as a ray passes from the center of the core of a graded index fiber out toward its outer edge on the zigzag path it passes through a zone of lower refractive index. As noted by Eq. (20.4.9), the phase velocity (and the group velocity) in this lower index region is higher than at the center of the core

and increases with distance from the center. As a ray zigzags across the core it experiences an alternately rising and then falling velocity. Careful choice of the profile index will cause the average velocity for all modes to be approximately the same, resulting in a much lower dispersion than occurs in a step-index fiber. It has been shown that an alpha-graded fiber profile with an alpha slightly less than 2 will have an intermodal dispersion approaching a theoretical minimum of

$$\Delta t_{\text{graded}} = \frac{n_1 z \Delta^2}{8} c \quad (20.4.13)$$

where n_1 is the refractive index at the core center and Δ is the normalized refractive index difference between center and cladding. While this is a theoretical minimum, commercial graded index fibers with intermodal dispersions of much less than 1 ns/km have been made. These low dispersion factors make the multimode graded index fibers very competitive with the single-mode fibers for long links.

EXAMPLE 20.4.2

Assume that the fiber in Example 20.4.1 has been made with an optimally graded core index profile and find its intermodal dispersion.

SOLUTION From Example 20.4.1, $n_1=1.55$ and $\Delta=0.0258$.

$$\Delta_t \text{ (graded)} = \frac{n_1 z \Delta^2}{8c} = \frac{1.55 \times 1000 \text{ m/km} \times 0.02558^2}{8 \times 0.3 \text{ Gm/s}} = \mathbf{0.430 \text{ ns/km}}$$

which is very much less than the 137 ns/km for the step-index fiber. The total dispersion for 12.5 km is

$$\Delta_t = \Delta_t \text{ (per km)} \times z \text{ (km)} = 0.43 \times 12.5 = \mathbf{5.38 \text{ ns}}$$

Material (or Chromatic) Dispersion

The index of refraction of the core glass is not the same for lights of different wavelengths, but varies across the spectrum. Practical light sources do not put out pure monochromatic light, but produce a spectrum distributed about a central wavelength λ_0 as shown in Fig. 20.4.3 and having a spectral bandwidth $\lambda_{3 \text{ dB}}$. Light components of a pulse with shorter wavelengths will experience more delay than will those components of the same pulse with longer wavelengths. The result will be a time dispersion of the pulse at the receiver end of the fiber. A narrow-bandwidth source will produce less dispersion than will a wideband source.

The chromatic dispersion has been shown to be proportional to the second derivative of the index of refraction with respect to wavelength ($d^2 n/d\lambda^2$), giving a dispersion of

$$\Delta_t = D_m \lambda_{3 \text{ dB}} \quad (20.4.14)$$

where the material dispersion factor D_m is given by

$$D_m = -\frac{z \lambda_0}{c} \frac{d^2 n}{d \lambda^2} \quad (20.4.15)$$

which is usually given in ps/nm-km. Typical values for pure and doped silica fibers are shown in Fig. 20.4.4.

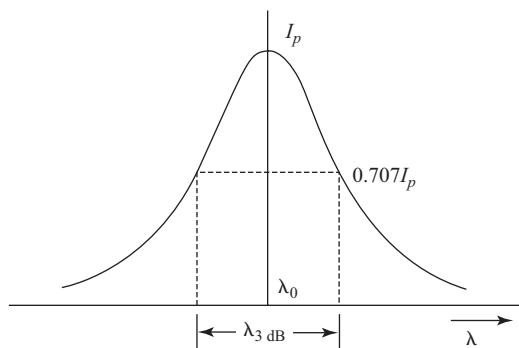


Figure 20.4.3 Spectral content of a light source.

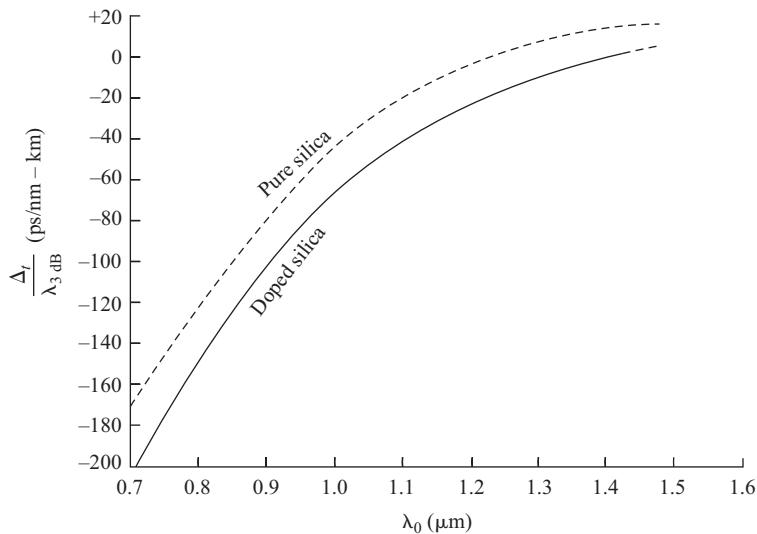


Figure 20.4.4 Material dispersion coefficient as a function of wavelength for silica fibers.

It should be noted that the curve of material dispersion factor crosses zero near a wavelength of 1.3 nm. As a result, if a fiber is used in single mode at a wavelength near 1.3 nm, there will be no intermodal dispersion and the chromatic dispersion will very nearly cancel out. For multimode fibers, however, the intermodal dispersion will generally be much larger than the chromatic dispersion.

EXAMPLE 20.4.3

The fiber in Example 20.4.1 is used with a 0.8- μm light source with a spectral bandwidth of 1.5nm. What value of material (chromatic) dispersion might be expected?

SOLUTION From Fig. 20.4.4, at $\lambda_0 = 0.8 \mu\text{m}$, the material dispersion coefficient is $D_m = -0.15 \text{ ns/nm} - \text{km}$. Dispersion per kilometer thus is

$$\Delta_t = D_m \lambda_{3 \text{ dB}} = -0.15 \text{ ns/nm} - \text{km} \times 1.5 \text{ nm} = -0.225 \text{ ns/km}$$

and total material dispersion for a 12.5-km fiber is

$$\Delta_t(\text{md}) = \Delta_t (\text{per km}) \times z (\text{km}) = -0.225 \times 12.5 = \mathbf{2.81 \text{ ns}}$$

This figure compares to an intermodal dispersion in the step-index fiber of 1710 ns and an intermodal dispersion in a graded index fiber of 5.38 ns (see Examples 20.4.1 and 20.4.2).

Waveguide Dispersion

If a fiber can be operated so that intermodal dispersion and material dispersion both disappear (as for a single-mode fiber operated near 1.3 μm), then a third dispersion mechanism will predominate. This is called *waveguide dispersion* and results only from the guiding characteristics of the fiber.

All practical light sources contain light components of different wavelengths distributed over a spectral bandwidth, as shown in Fig. 20.4.3. A slight spectral shift to higher wavelength will slightly lengthen the path length between successive reflection points (see Fig. 20.2.9) and increase the corresponding incidence angle for each supported mode. This in turn will increase the corresponding group velocity. Thus each supported mode will suffer a dispersion effect dependent on the spectral width of the source so that, even if the other effects cancel, this one still remains. It too would disappear if a truly monochromatic light source could be developed, but that is not possible.

It has been shown that for an ideal single-mode step-index fiber the dispersion coefficient due only to the waveguide dispersion mechanism has a peak value of about $D_w = 6.6 \text{ ps/nm-km}$.

EXAMPLE 20.4.4

A 12.5-km single-mode fiber is used with a 1.3- μm light source with a spectral bandwidth of 6 nm. Find the expected waveguide dispersion.

SOLUTION

$$D_w = 6.6 \text{ ps/nm-km}$$

$$\Delta_t (\text{wg}) = D_w z \Delta_{3 \text{ dB}} = 6.6 \times 12.5 \times 6 = \mathbf{495 \text{ ps}}$$

This compares to 1710 ps of intermodal dispersion and 2810 ps of material dispersion for the same fiber.

Total Dispersion and Maximum Transmission Rates

Three dispersion mechanisms have been discussed, each of which contributes to pulse broadening during fiber transmission to a greater or lesser extent. In multimode step-index fibers the intermodal dispersion will predominate, while in a single-mode fiber operated at the wavelength for minimum material dispersion the waveguide dispersion will predominate.

The dispersions result in an approximately Gaussian pulse shape at the receiver end of the fiber, for which the independent dispersion effects combine in a root-mean-square manner to give a total dispersion of

$$\Delta_t (\text{tot}) = \sqrt{\Delta_t^2 (\text{imd}) + \Delta_t^2 (\text{md}) + \Delta_t^2 (\text{wg})} \quad (20.4.16)$$

Note that this is only the dispersion effect and does not include the transmitted pulse width. It can be shown that if the transmitted pulse also has an approximately Gaussian shape, as is typically the case, then the received pulse width t_r can be approximated by the root-mean-square combination of the transmitted pulse width t_w and the total dispersion to give

$$t_r = \sqrt{t_w^2 + \Delta_t(\text{tot})^2} \quad (20.4.17)$$

Now the maximum allowed bit transmission rate B can be found. It can be shown that if a power penalty of less than 1 dB is allowed to obtain the same error rate as at lower bit rates, then the bit rate B must be less than

$$B \leq \frac{1}{2t_r} \quad (20.4.18)$$

which is sufficient to avoid significant intersymbol interference. This is based on a *unipolar nonreturn to zero* (UPNRZ) line code where two alternating bits constitute one cycle of the highest-frequency component to be transmitted, and the transmitted pulse condition lasts for a full bit period t_r . If a UPRZ code is to be used, then the maximum bit period must be doubled, since in the coding a 0 is placed between each pair of bits in time so that the effective bit rate is one-half the maximum bit rate. The use of a UPRZ code is effective on fibers with significant dispersion, where the inserted zero periods allow room for the pulses to spread without overlapping.

EXAMPLE 20.4.5

A single-mode fiber operating at $1.5 \mu\text{m}$ is found to have a material dispersion of 2.81 ns and a waveguide dispersion of 0.495 ns . Find the maximum allowed bit rate for the fiber with a pulse width of 2.5 ns .

SOLUTION

$$\Delta_t(\text{imd})=0, \quad \Delta_t(\text{md})=2.81 \text{ ns}, \quad \Delta_t(\text{wg})=0.495 \text{ ns}$$

The total dispersion is

$$\begin{aligned} \Delta_t(\text{tot}) &= \sqrt{\Delta_t^2(\text{imd}) + \Delta_t^2(\text{md}) + \Delta_t^2(\text{wg})} \\ &= \sqrt{0^2 + 2.81^2 + 0.495^2} = 2.85 \text{ ns} \end{aligned}$$

The received pulse width is

$$t_r = \sqrt{t_w^2 + \Delta_t(\text{tot})^2} = \sqrt{2.5^2 + 2.85^2} = 3.79 \text{ ns}$$

so the maximum allowed bit rate becomes

$$B \leq \frac{1}{2t_r} = \frac{1}{2 \times 0.00379 \mu\text{s}} = \mathbf{131.9 \text{ Mbits/s}}$$

When a bit stream is transmitted on a fiber at bit rate B , the maximum sinusoidal frequency component transmitted is determined as

$$f_{\max} = \frac{B}{2} \quad (20.4.19)$$

which occurs on an alternating pattern of 1 and 0 bits, with each bit forming a half-cycle of the sinusoid. If the transmitted pulse width is reduced to a very short pulse width (approaching an impulse), then the received pulse width approaches a minimum width $t_r = \Delta_t (\text{tot})$ or the total dispersion, which sets a limit on the maximum bit rate B (max) of

$$B (\text{max}) = \frac{1}{2t_r (\text{min})} \approx \frac{1}{2\Delta_t (\text{tot})} \quad (20.4.20)$$

Now the effective bandwidth of the fiber becomes the maximum value of f_{\max} expressed in hertz as

$$\text{BW} = f_{\max} (\text{max}) = \frac{B_{\max}}{2} = \frac{1}{4\Delta_t (\text{tot})} \quad (20.4.21)$$

This bandwidth is stated in terms of a sinusoidal signal. A binary bit rate *bandwidth* may also be expressed as the maximum bit rate B (max) expressed in bits per second.

Multiplying the bit rate bandwidth of the total fiber by its length gives a quality factor for the fiber called the *bandwidth-distance product* (BDP). This is equivalent to finding the bit rate bandwidth for a 1-km unit length of fiber. The BDP then is

$$\text{BDP} = B_{\max} \times z = \frac{1}{2\Delta_t (\text{tot per km})} \quad (20.4.22)$$

Longer fibers may be used, but the bandwidth must be reduced. The dispersion-limited length of a fiber for a given bit rate then can be found as

$$z_{\max} (\text{disp}) = \frac{\text{BDP}}{B} \quad (20.4.23)$$

EXAMPLE 20.4.6

A multimode step-index fiber has a dispersion of 4 ns/km and is to be operated at a bit rate of 10 Mbps.

- (a) Find the bandwidth-distance product for the fiber. (b) Find the dispersion limited length of fiber for the given bit rate.

SOLUTION

$$(a) \text{ BDP} = \frac{1}{2\Delta_t (\text{tot per km})} = \frac{1}{2 \times 0.004 \mu\text{s/km}} = \mathbf{125 \text{ Mbps-km}}$$

$$(b) z_{\max} (\text{disp}) = \frac{125 \text{ Mbps-km}}{10 \text{ Mbps}} = \mathbf{12.5 \text{ km}}$$

20.5 Light Sources for Fiber Optics

Introduction

Light sources for fiber optics act as light transmitters and must meet certain requirements if they are to be acceptable for the purpose. First, the light produced must be as nearly monochromatic (single frequency) as possible. Most light sources are not single frequency, but emit light at many frequencies distributed over a band or portion of the spectrum, which may be quite broad. A few sources such as gas ionization lamps, light emitting diodes, and lasers emit light over a much narrower band. But even these sources are not truly monochromatic and do emit at several frequencies over a narrow band. The emission spectra of some typical light transmitters are compared in Fig. 20.5.1.

Next, the light source should have a high-intensity output so that sufficient energy is transmitted on a fiber to overcome the losses encountered during transmission. Also the sources must be capable of being easily modulated. Although most sources presently available are capable of analog modulation (for example, amplitude modulation), binary on/off modulation with PCM is usually used since it gives good results with much better noise immunity than other methods.

Finally, the devices must be small and easily coupled to fibers so that excessive coupling losses do not occur. They must also be relatively inexpensive to manufacture.

Light emitting diodes and semiconductor lasers are both extensively used for this application. Both emit narrow-band light at fixed center wavelengths as the result of the recombination of hole-electron pairs in the junction area of the diode. Each such recombination is accompanied by the release of a photon of light with a fixed energy content that corresponds to the wavelength of light emitted and to the energy required to free a valence electron from its parent atom in the semiconductor.

Each photon contains an amount of energy that is related to the corresponding electromagnetic frequency by the expression

$$E = hf \quad (20.5.1)$$

where E is the energy in joules, h is Planck's constant (6.625×10^{-34} J-s) and f is the frequency in hertz. Light is usually designated by its wavelength instead of frequency. Wavelength is related to frequency by

$$f = \frac{c}{\lambda} \quad (20.5.2)$$

where c is the velocity of light in free space (300 Mm/s) and λ is the wavelength in meters. The energy is usually stated in electron voits (eV) found by dividing the energy in joules by the electronic charge q (1.602×10^{-19} C/electron), or

$$E_q = \frac{E}{q} \quad (20.5.3)$$

Combining these relationships gives the eV energy in terms of the wavelength as

$$E_q = \frac{hc}{q} \frac{1}{\lambda} = \frac{1.241}{\lambda (\mu\text{m})} \quad (20.5.4)$$

The energy content of a photon released in a semiconductor is related to the energy band gap of the semiconductor material. The energy band gap is the amount of energy needed to excite a valence electron sufficiently to free it from the valence bond so it becomes available to conduct electricity. The width of the energy band gap in a device can be precisely tailored during manufacture by careful choice of the materials and doping levels used.

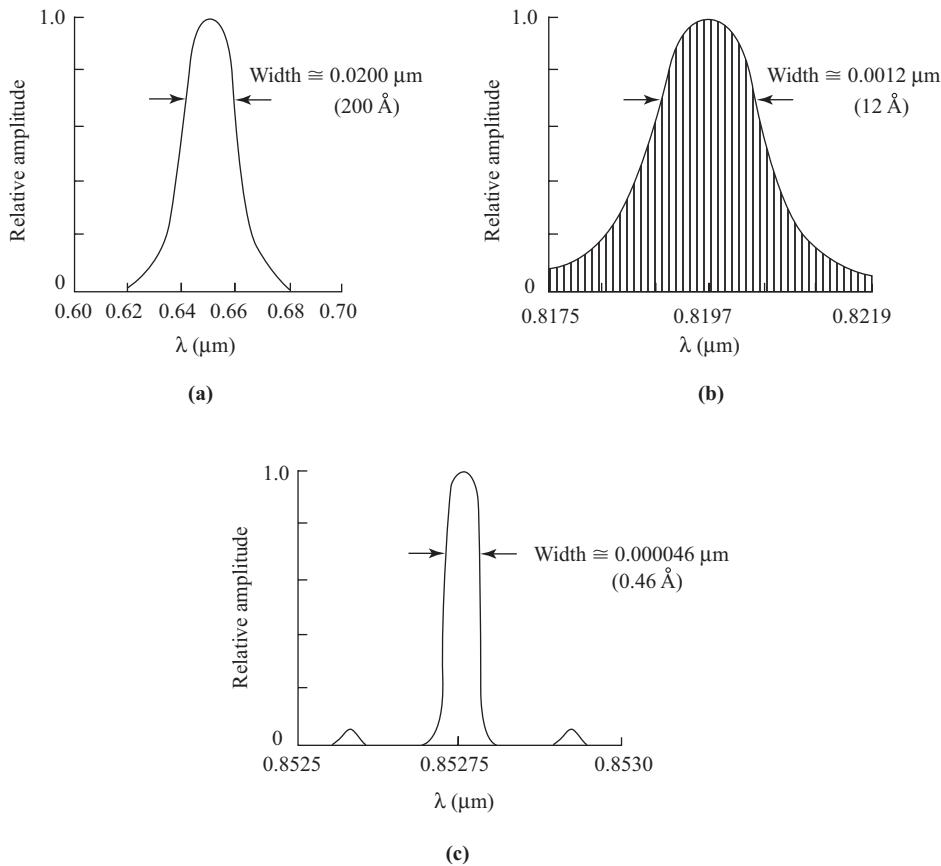


Figure 20.5.1 Light emission spectra for typical light transmitters. (a) LED operating at 0.65 μm. (b) Broad-spectrum laser diode at 0.8197 μm. (c) Narrow-spectrum laser diode at 0.85275 μm.

EXAMPLE 20.5.1

Three semiconductor diodes are made using materials that have energy band gaps of 1.9, 1.46, and 0.954 eV. Find the wavelengths and frequencies of the light produced by them.

SOLUTION (a) The wavelength is

$$\lambda = \frac{1.241}{\text{eV}} = \frac{1.241}{1.9} = 0.6532 \text{ } \mu\text{m} \quad (= 6532 \text{ angstroms})$$

which emits in the orange-red portion of the visible spectrum. The frequency of this light is

$$f = \frac{c}{\lambda} = \frac{300 \text{ Mm/s}}{0.6532 \text{ } \mu\text{m}} - 459.3 \text{ THz}$$

- (b) $\lambda = 0.850 \text{ } \mu\text{m}$ and $f = 352.9 \text{ THz}$, which produces light in the 0.8-μm loss window of silica fibers.
- (c) $\lambda = 1.300 \text{ } \mu\text{m}$ and $f = 230.6 \text{ THz}$, which produces light in the 1.3-μm loss window of silica fibers.

Light Emitting Diodes

A light emitting diode (LED) works by the process of spontaneous emission when it is forward biased and conducting current. One side of the diode junction is *p*-type material containing mostly holes (broken covalent bonds with missing electrons). The other side of the junction is *n*-type material containing mostly free electrons.

At zero bias a depletion zone separates the *p* and *n* regions as shown in Fig. 20.5.2(a). The depletion zone has had all free electrons and holes removed, uncovering two layers of fixed charges of opposite polarities that form a potential barrier across the depletion zone.

When forward bias is applied, the barrier potential is reduced and the depletion zone is narrowed until holes and electrons are free to cross the barrier to conduct current, as shown in Fig. 20.5.2(b). Holes injected into the *n* region quickly encounter free electrons and recombine. Electrons injected into the *p* region encounter holes and recombine. External current flow replenishes the lost holes and electrons.

When each hole-electron pair recombines, a single photon of light is released, which carries with it the amount of energy required to liberate an electron from a valence bond. The wavelength and frequency of light emitted are determined from this band-gap energy. The intensity of light emitted is proportional to the forward current conducted by the junction, which controls the number of holes and electrons crossing the junction to be recombined. Since the injected minority carriers have a very short lifetime, all recombination (and hence light emission) takes place in the near vicinity of the *pn* junction of the diode.

The LED is formed by diffusing a microscopically thin transparent layer of *p* material into the surface of an *n*-type substrate chip. Light is emitted within the junction and radiates randomly through the thin *p* layer in all directions, as illustrated in Fig. 20.5.3. The chip is usually arranged so that it is at the bottom of a well or behind a lens that concentrates the light onto the end of a fiber. The light generally is not concentrated, does not travel in well-defined directions, and has a very broad spectrum. As a result ordinary LEDs are generally useful only for short lengths of large-core diameter multimode fibers operating at low bit rates.

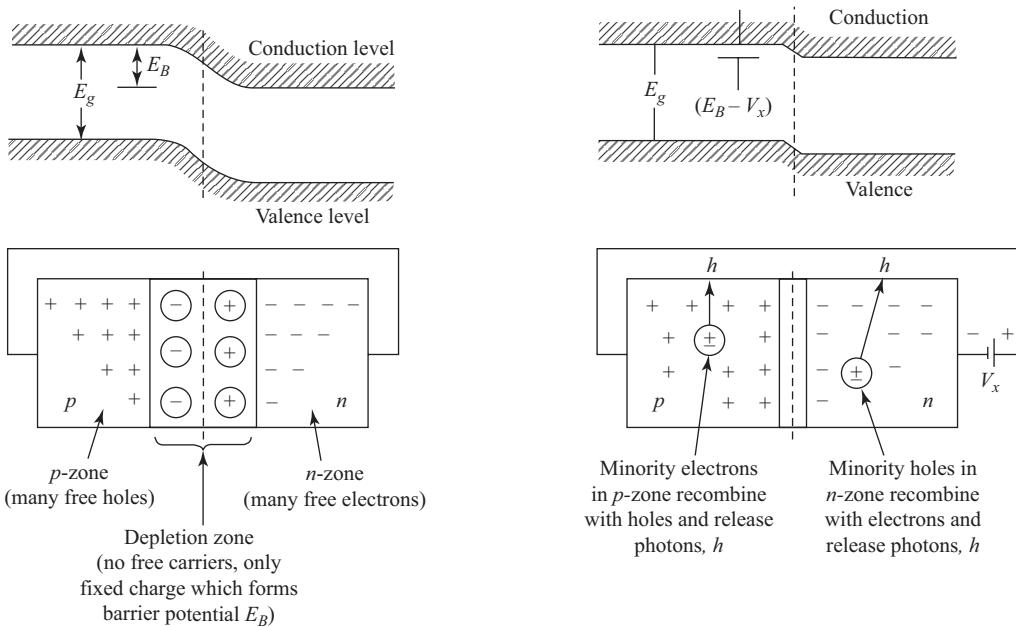


Figure 20.5.2 Energy band diagram and carrier distribution within a *pn* diode (a) with zero bias and (b) with forward conducting bias.

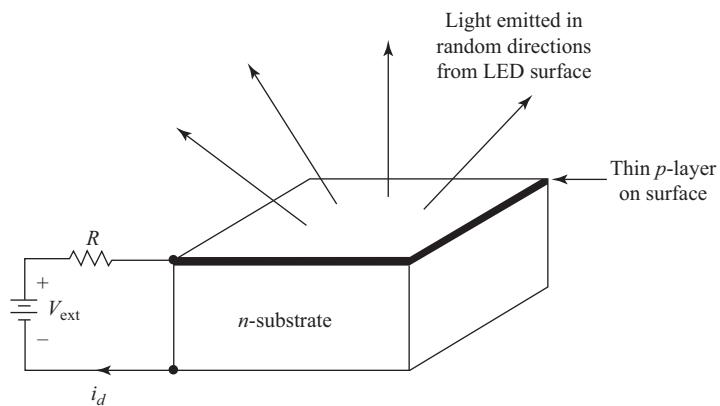


Figure 20.5.3 Forward-biased light emitting diode (LED) chip.

The LEDs used for fiber optics are usually of the gallium arsenide (GaAs) type, with various dopants added to shift the center wavelength of the radiation spectrum. Dopants used are phosphorus (P), indium (In), and aluminum (Al). Gallium arsenide phosphide (GaAsP) diodes can be made with band gaps in the range of 1.5 to 2.0 eV (0.62- to 0.83- μm wavelength). Gallium indium arsenide (GalnAs), indium arsenide phosphide (InAsP), or aluminum indium arsenide (AlInAs) diodes cover the range from 0.5 to 1.5 eV (0.83 to 2.5 μm). Some of these structures may also be used to form laser diodes.

The radiation spectrum of a typical LED is shown in Fig. 20.5.1(a). This is centered on a wavelength of 0.65 μm and has a spectral width of about 0.02 μm (200 angstroms), which is typical of the spectra of GaAsP diodes. The actual diode emission center frequency is chosen so that it falls in the middle of one of the loss windows of the fiber loss spectrum.

Semiconductor Laser Diodes

The term *laser* is an acronym for *light amplification by stimulated emission of radiation*. Laser action has been obtained using many different materials, including gases such as neon or carbon dioxide, liquids, and solids such as rubies. The semiconductor laser uses the solid semiconductor as the lasing material.

Stimulated emission occurs as follows. In the semiconductor diode, when a hole-electron pair is created by absorption of energy from a "pump" source, it is raised from the ground or valence energy state to a higher conduction energy state represented by the energy band gap of the semiconductor. An electron raised to the high-energy conduction state will remain in that state for a short period (the carrier mean lifetime), and then it will recombine with a hole and revert to the lower valence state, giving off a photon of energy. This recombination in the ordinary LED is purely spontaneous. In this spontaneous condition, the density of excited electrons in the conduction band will be lower than the density of electrons trapped in the valence state.

If the semiconductor is pumped hard enough, however, the equilibrium density of excited electrons in the conduction band can rise until it is higher than the density of valence electrons. This surplus of excited electrons forms a population inversion, which exists in a semistable condition. If light with photons of energy close to the excitation energy of these electrons is injected into this unstable inverted population of electrons, the light field will absorb energy from the unstable electrons and prematurely cause them to recombine, thus raising the amount of light generated above the spontaneous level. Any given injected photon may shake loose one, two, or more additional photons, depending on the density of the excess conduction electrons.

A small increase in pumping energy can cause a large increase in the density and an increase in the number of excess photons released by the incident light, creating a *light gain*. The pumping level at which the light gain becomes greater than unity is called the *laser threshold*. In a LED laser the pump is the forward conduction current, which can be externally controlled.

A three-level laser uses a material containing two different types of atoms with different energy band gaps. Certain of these will have a single valence energy band and two conduction bands. The lower conduction band is narrow and separated from the upper one. In operation, a surplus of conduction electrons is pumped to the higher conduction band, creating a higher electron density in this band than in the lower one (the population inversion). Injecting light with energy corresponding to the difference between the two conduction bands will cause laser action to occur and emit light in phase with the injected light as electrons fall from the upper conduction band to the lower conduction band. The electrons in the lower-energy band then will spontaneously and rapidly recombine and return to the valence state so that the intermediate state does not get loaded with excess electrons.

The laser action of the semiconductor diode can be enhanced by placing a reflecting surface at each end of the junction region to form a Fabry–Perot resonant cavity, as shown in Fig. 20.5.4. One reflecting surface is made partially reflecting so that a part of the incident light will pass out of the junction instead of being reflected back through it. The two surfaces are made parallel to each other so that light generated in the junction will bounce back and forth several times before escaping, thus increasing the chance that each photon will stimulate more emissions. The zone between the reflection surfaces can be compared to a resonant cavity in a microwave oscillator, with the cavity length being a multiple of the dominant wavelength mode of the laser junction. The parallel reflecting surfaces also tend to concentrate the generated light into a parallel beam emerging from one end of the chip, making it easier to couple to a fiber.

While the light beam in a laser such as that in Fig. 20.5.4 is contained within the junction by the opaque *p* and *n* layers, the junction width may be quite large depending on the chip width (for example, 2000 μm). The *stripe laser* or *gain guided laser* avoids this problem by confining the laser action to a narrow stripe across the chip between the reflecting surfaces, as shown in Fig. 20.5.5. The confining action results from having the pumping current injected into the junction from a narrow stripe contact deposited over the *p* layer. Current density within the junction only exceeds the laser threshold value within the stripe region under the contact, thus confining light generation to this stripe.

Light tends to be concentrated along the stripe under the contact, resulting in partial guiding. This guiding is not complete, however, and results in a wide beam width. Reducing the width of the stripe tends

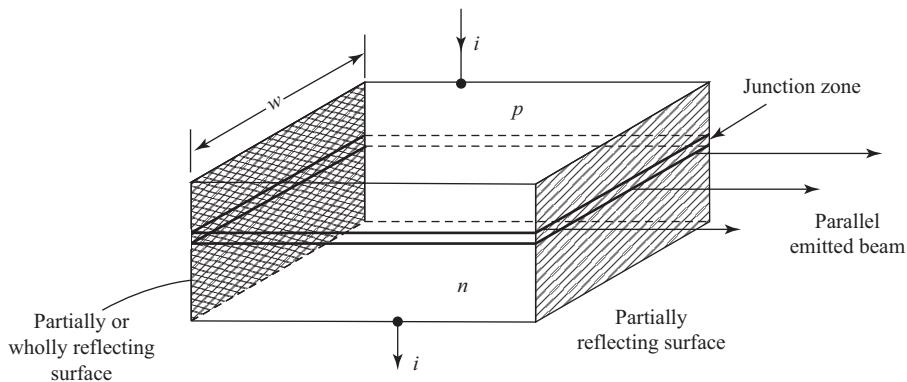


Figure 20.5.4 Unguided Fabry–Perot laser diode.

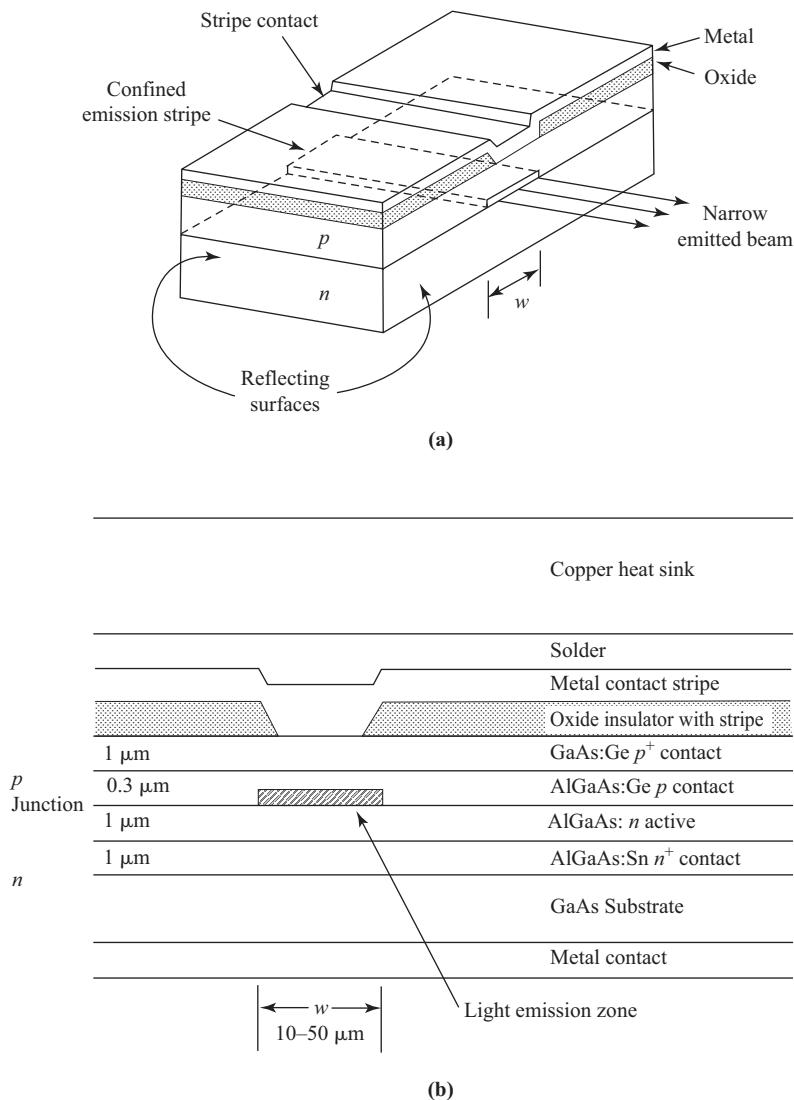


Figure 20.5.5 (a) Gain-guided stripe laser diode, (b) Cross section of a double-heterostructure laser showing the various layers.

to reduce, but not eliminate, the higher-order spatial modes, so the beam is wider than it should be. Furthermore, the cavity supports several spectral modes within its band pass (as with a low- Q single cavity resonator), so the laser tends to produce a series of harmonic lines with normally distributed magnitudes within its spectrum, centered on the dominant mode determined by the cavity length.

The *index guided laser diode* (ILD) or *buried heterostructure diode* provides a large improvement in both beam concentration and narrow spectrum. In this device, a stripe of the lasing material is surrounded

on all four sides by material with a different refractive index, resulting in a rectangular light guide, with Fabry–Perot reflectors at each end. The cross section of such a device is shown in Fig. 20.5.6. A light-confining layer of n -type GaAlAs forms the n zone of the laser diode. A thin ($0.03\text{ }\mu\text{m}$) strip as narrow as $2\text{ }\mu\text{m}$ turn of p -type GaAs is laid down on this to form the active zone light guide. A light-confining layer of p -type GaAlAs is laid over and around the stripe to complete the guide. The length of the guiding channel is typically less than $100\text{ }\mu\text{m}$, the length of the chip. A p -ion implantation on either side of the stripe serves to improve the guide lateral containment. Uniform current density is passed through the entire pn junction region, but lasing only occurs within the active guide zone.

The narrow stripe serves to eliminate the higher-order spatial modes, while the Fabry–Perot cavity tunes the laser to produce a very narrow spectrum consisting of a few narrow spectral lines distributed about the center wavelength. The center spectral line is typically 20 to 30 dB more intense than the adjacent lines. Beam widths of 40° lateral by 5° and spectral widths of $2\text{ }\mu\text{m}$ can be readily obtained from such laser diodes. This type of diode is much more efficient in light production than the preceding types, with the result of optical powers in the 1- to 20-mW range being typical.

The *distributed feedback* (DFB) laser diode is a modification of the index-guided laser diode. In this device, the exit surface of the stripe is made as nonreflecting as possible so that there is no Fabry–Perot cavity, as shown in Fig. 20.5.7. In addition, the boundary of the light guide adjacent to the active junction is etched to form a diffraction grating before the top layers are formed. This diffraction grating acts as a *Bragg reflector*, which reflects light traveling toward the back end of the guide and adds it in phase to reinforce the light traveling in the forward direction.

When a light beam is passed close to and parallel to a diffraction grating, the phenomenon of *Bragg reflection* occurs. For any given spatial mode, only light of a single wavelength, determined by the spacing of the grating, is reflected back on itself. Others are scattered or cancelled.

Within the light guide, both a forward wave and a reverse wave are present. Each facet of the grating reflects a portion of the reverse wave and adds it to the forward wave in phase, thereby reinforcing it. The grating has many modes, depending on the angle of incidence of the reverse wave on the grating plane, and the wavelength of the light source. However, only one that coincides with a dominant mode of the laser itself

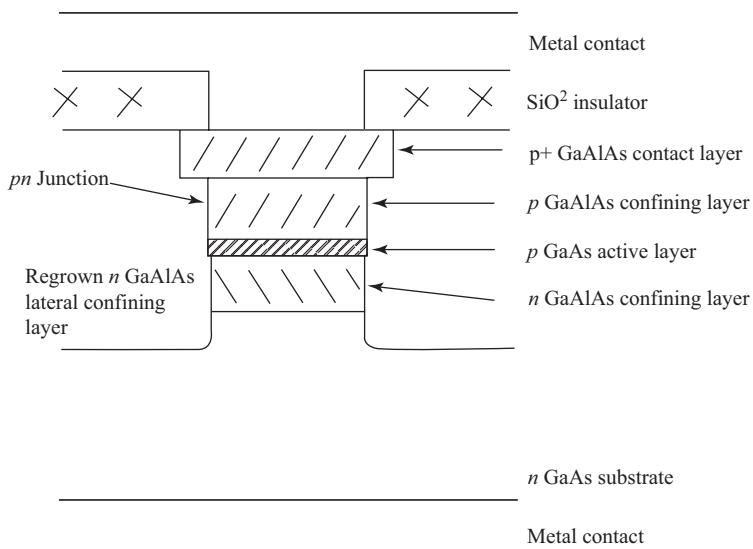


Figure 20.5.6 Index-guided or buried heterostructure laser diode.

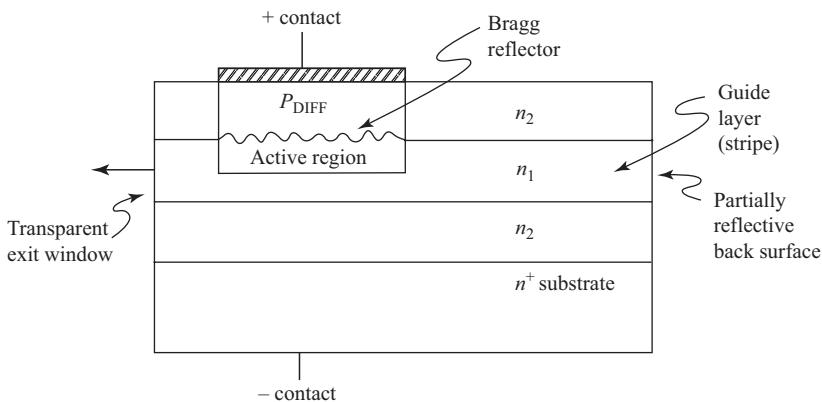


Figure 20.5.7 Distributed feedback laser diode.

will be excited, resulting in virtually monochromatic light. The grating acts very much in the manner of a series of resonant circuits tuned to the dominant-mode wavelength and sequentially coupled to the guide, providing a very narrow band-pass filter effect.

The wavelength of the dominant mode is related to the free-space wavelength by the Bragg relationship, as follows:

$$G = \frac{m\lambda_0}{2n_{\text{eff}}} \quad (20.5.5)$$

where m = Bragg reflection mode order ($0, 1, 2, \dots$)

n = guide layer index of refraction

n_{eff} = average guide layer index within the active layer between the junction and the grating

G = spacing between ridges of the grating.

The odd-numbered Bragg modes reinforce the forward wave without producing excessive scattering. However, the even-numbered modes produce more scattering and thus have lower reflection efficiency. The $m = 1$ mode results in a very fine grating that is hard to fabricate, but it provides the highest degree of reflection. The $m = 3$ mode is more often used although it isn't as efficient. The nonreflecting surface at the back end of the stripe is placed some distance from the active region so that reflections from it are minimized to reduce interference with the Bragg reflector. A transparent window at the other end allows coupling into the fiber for transmission.

Index guiding contains the light within the active region. The result of this is a device with a single spatial mode and a single spectral mode, producing a very narrow beam of essentially monochromatic light with a spectral width of less than 0.1 nm, with no significant side lines. Spectral line widths of less than 0.2 nm have been realized in such diodes. Using laser diodes of this type with mono-mode fibers results in very low material dispersion of the transmitted pulses and correspondingly higher transmission bandwidths.

The *distributed Bragg reflector* (DBR) laser diode shown in Fig. 20.5.8 is an improvement on the DFB laser diode above. In this case, the active region is made in the normal manner, with the guiding interfaces plane and parallel to the junction. Two Bragg reflectors are placed on the guide walls on either end of the

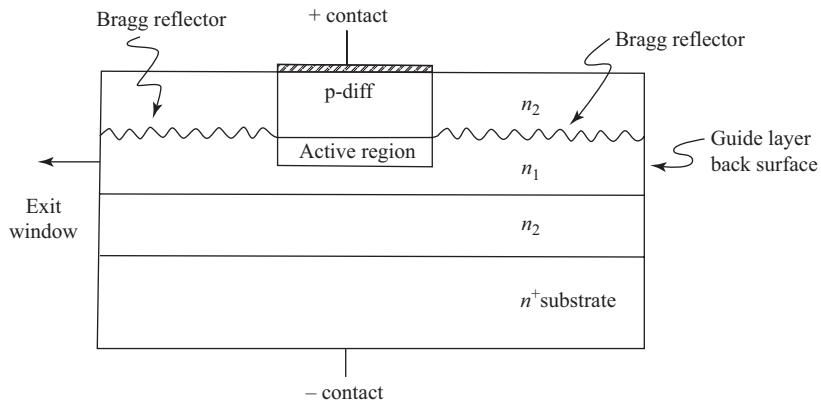


Figure 20.5.8 Distributed Bragg reflector laser diode.

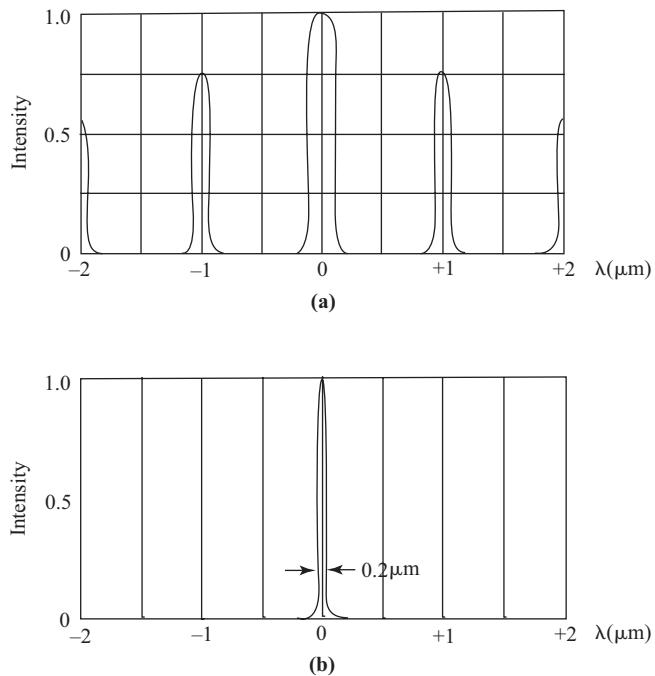


Figure 20.5.9 (a) Light spectrum of a Fabry-Perot laser diode. (b) Light spectrum of a distributed Bragg reflector laser diode.

active region so that light is reflected back and forth between them to reinforce the dominant wavelength, forming a resonant cavity. However, in this case the resonant wavelength is determined by the spacing of the grid on the Bragg reflectors and not by the guide length. The rest of the structure is the same as in the DFB laser. The result is a much more efficient diode with a virtually monochromatic light whose wavelength can be closely tailored during manufacture. Figure 20.5.9 compares the spectrum of light produced by a

Fabry–Perot laser diode to that of a DBR laser. The monochromatic nature of the light produced by the latter makes it ideal for use with linear analog modulation and for wavelength multiplexing applications.

20.6 Photodetectors

Introduction

Several types of photosensitive devices may be used as detectors for use with fiber optics. These include silicon photodiodes, phototransistors, and photore sistors. Not all of them have the speed of response and high sensitivity needed for useful communications application. Those that do include the *pin* diode and the avalanche diode.

pn Photodiode

An ordinary *pn* diode may be used as a photodetector. It has sufficient speed, but its sensitivity is very low. Figure 20.6.1(a) shows the structure of a *pn* photodiode. It has a *p* layer deposited on an *n* substrate so that light enters the junction through the *p* layer. The reverse-bias junction depletion zone is relatively thin. Photons of light entering the depletion zone ionize hole–electron pairs when they encounter atoms within the crystal structure. The mobile holes and electrons are swept out of the depletion zone by the electric field due to the reverse bias and contribute to the leakage current. The resulting leakage is proportional in magnitude to the incident light intensity.

Many of the photons entering the junction depletion zone of the *pn* diode pass through into the *n* region without ionizing an atom. Hole–electron pairs generated within the *n* region are not affected by the junction electric field and do not contribute to the photocurrent. As a result, the *responsivity* or conversion efficiency of the diode is quite low. However, since the depletion zone is quite thin, the carrier lifetime within the depletion zone is short so that the diode responds to rapid changes in light intensity.

pin Photodiode

The sensitivity of the *pn* photodiode can be improved by including a lightly doped (or almost intrinsic) *n* layer between the junction and the more heavily doped *n*-contact region to form the *pin* diode. The intrinsic layer, shown in Fig. 20.6.1(b), is made thick enough so that most of the photons that pass through the junction without ionizing are absorbed within this layer. The junction electric field extends deep into this region, and any holes produced by the photons are swept across the junction to add to the photocurrent. This more complete use of the incident photons results in a larger photocurrent than would be the case without the intrinsic layer and a much higher sensitivity. However, the carriers produced within the intrinsic layer have farther to travel to cross the junction, so the response of the *pin* diode is slower than that of the *pn* diode. Sensitivity is gained at the expense of speed.

Avalanche Photodiode (APD)

If the negative reverse-bias voltage applied to the *pin* diode is increased, a threshold will be reached beyond which the field intensity at the junction becomes high enough so that electrons being accelerated through the depletion zone will create secondary hole–electron pairs when they collide with atoms. One photoelectron can result in many additional secondary electrons being created, resulting in an *avalanche multiplication effect*. The number of carriers generated by the avalanche effect is exponentially related to the field intensity so that high gains can be obtained with modest reverse bias voltages. All the additional carriers

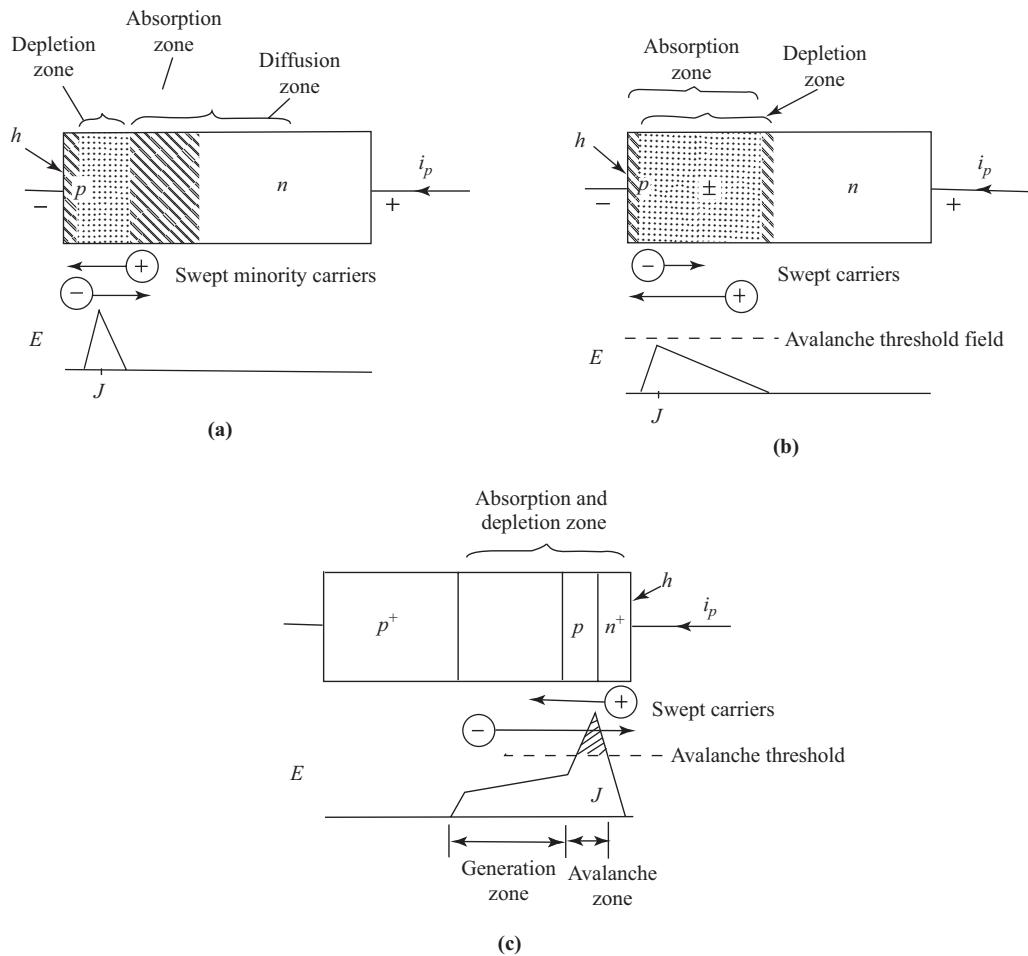


Figure 20.6.1 Photodiode structures with field distribution patterns. (a) pn photodiode without multiplication. (b) pin photodiode without multiplication. (c) $pipn$ avalanche photodiode with multiplication.

produced by the avalanche contribute to the photocurrent, so much higher conversion efficiencies are possible than without it.

The pin structure is not ideal for avalanche-mode operation since the electric field is distributed throughout the intrinsic region instead of being concentrated at the junction. As a result, high bias voltages are required to get field intensities over the avalanche threshold value. To correct this, the $pipn$ structure in Fig. 20.6.1(c) is used. In this case, light enters the diode through a thin heavily doped n layer, which forms an abrupt junction with the lightly doped p layer adjacent and passes into the thick intrinsic layer. The electric field is distributed as shown, with a high field intensity zone at the abrupt $n^+ - i$ junction and a long tail of lower field intensity extending into the intrinsic layer.

As before, the intrinsic layer serves to allow the field tail to sweep the photoelectrons back to the junction and into the avalanche region. The $p^+ - i$ on the other side serves to collect the holes from the intrinsic layer, and its doping is used to control the field distribution.

These diodes have the highest sensitivity of any of the presently available photosensors. However they do have disadvantages. First, carriers in the intrinsic region have a long transit time to the junction, which slows the response as in the *pin* diode. Second, the avalanche multiplication factor tends to fluctuate randomly to add noise to the signal, so a compromise must be struck between gain and noise.

A typical avalanche photodiode will produce avalanche multiplication factors of 100 to 300 with bias voltages of 100 to 400 V. The bias voltage must be closely regulated to minimize noise. Photocurrents are typically only a few microamperes in magnitude. Sensitivities of avalanche diodes are such that received photo powers of -70 dBm (at 1-Mbit rate) are usable. This corresponds to a received power of 0.1 nW .

Silicon *pin* and avalanche diodes typically have a wavelength response that extends from 0.6 and $1.2 \mu\text{m}$. Different types respond to wavelengths from 1 to $1.4 \mu\text{m}$. Figure 20.6.2 shows the spectral responsivity (a measure of conversion efficiency in $\mu\text{A}/\mu\text{W}$) for a typical silicon *pin* diode. The figure shows a rapid decrease of response for wavelengths above $1.1 \mu\text{m}$.

Optical Receiver Circuit

Figure 20.6.3 shows the block diagram of a practical optical receiver circuit that makes use of an avalanche photo diode (APD). Light transmitted over a fiber is on/off amplitude modulated so that when light is off only a few nanoamperes of dark (leakage) current flows in the APD, but when light is on a photocurrent of up to $250 \mu\text{A}$ flows. A low-noise transresistance amplifier with a gain of 4000Ω produces a maximum output voltage of about 1 V. The AGC amplifier provides a dynamic range before overload of about 37 dB. The output is bandwidth-limit filtered to reduce noise before being passed to a digital detection circuit.

A sample of the signal output from the AGC amplifier is peak detected to provide the feedback input to the AGC summing point, where it is compared to a reference voltage to set the AGC onset threshold. The AGC output is used to suppress the gain of the main AGC amplifier on high signals and also to reduce the gain of the APD (to limit noise in the detector). The receiver has a sensitivity to a minimum light signal of about -60 dBm (or 1 nW) for full output and a dynamic range of up to 37 dB. Maximum transmission speeds of about 50 Mb/s can be realized depending on the APD used and the noise characteristics of the circuit.

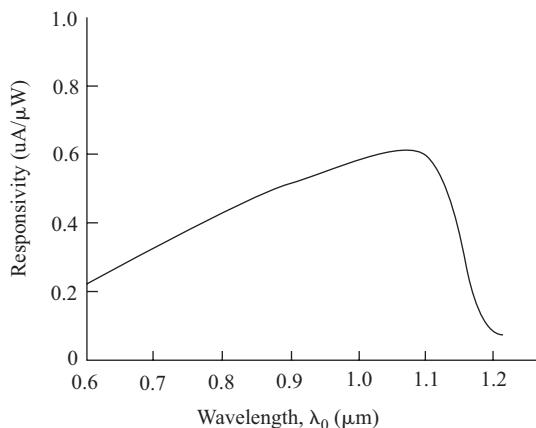


Figure 20.6.2 Spectral responsivity of a typical silicon *pin* photodiode.

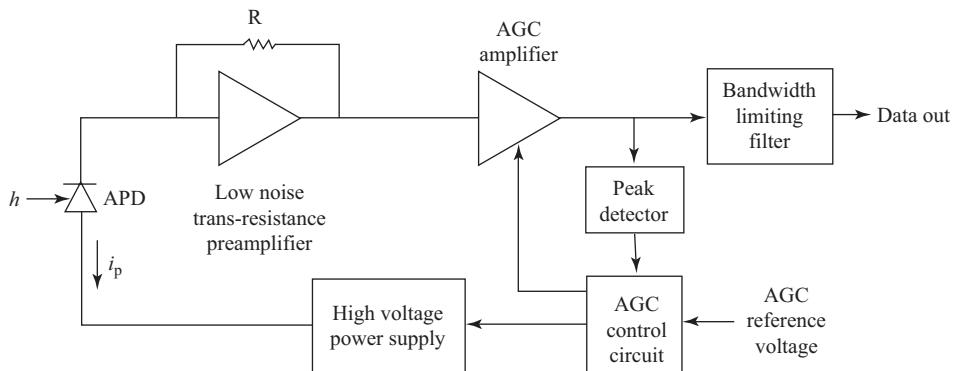


Figure 20.6.3 Practical optical receiver circuit block diagram. (S. D. Personick, M. K. Bamoski, eds., *Fundamentals of Optical Fiber Communications*, 2nd ed., Academic Press, 1981, p. 309, used with permission.)

20.7 Connectors and Splices

Losses in Connectors and Splices

Any fiber run will have a minimum of two connectors or splices since it must be terminated in a transmitter at one end and a receiver at the other. Usually, the transmitter and receiver each have a pigtail permanently connected during manufacture with a connector at the end, so that they may be removed for servicing. Matching connectors are fitted on the main fiber ends.

The main fiber run can be up to 100 km in length, requiring several inline splices to connect the sections during installation. Furthermore, the fiber may suffer accidental breaks during its lifetime, requiring the introduction of field repair splices. Each of these adds to the total loss of the fiber run and must be accounted for in the receiver and transmitter gains.

Several factors may affect the amount of loss a connector or splice will introduce into a fiber run. Some of these are discussed here.

Core Size Mismatch. Poor control of core diameter during fiber manufacture may result in trying to match two cores of different sizes in a connector. If the incoming core is smaller than the outgoing one, no problem is caused because all the light from the source fiber enters the other. However, if the outgoing fiber is smaller, then only part of the light from the incoming core will reach it. Part of the light will be lost, adding to the total loss of the fiber. This is illustrated in Fig. 20.7.1(a), which shows a cone of light escaping around the smaller outgoing core. The problem may also occur if an attempt is made to couple two different types of fiber.

Lateral Core Misalignment. If the cores are exactly the same size, but do not line up exactly on each other's axes (lateral displacement), then light will escape from the exposed portion of the incoming core face, as shown in Fig. 20.7.1(b). Such a misalignment may occur because the connector used does not line up the two outside diameters exactly, either because the outside cladding diameters are not exactly the same or because the cores are not exactly centered in the cladding. The result is more loss.

Some connectors are arranged so that lateral displacement may be corrected after assembly, using adjusting screws to center the fibers. This is crucial in the small-diameter core single-mode fibers where displacements of 10 μm could result in complete loss of coupling.

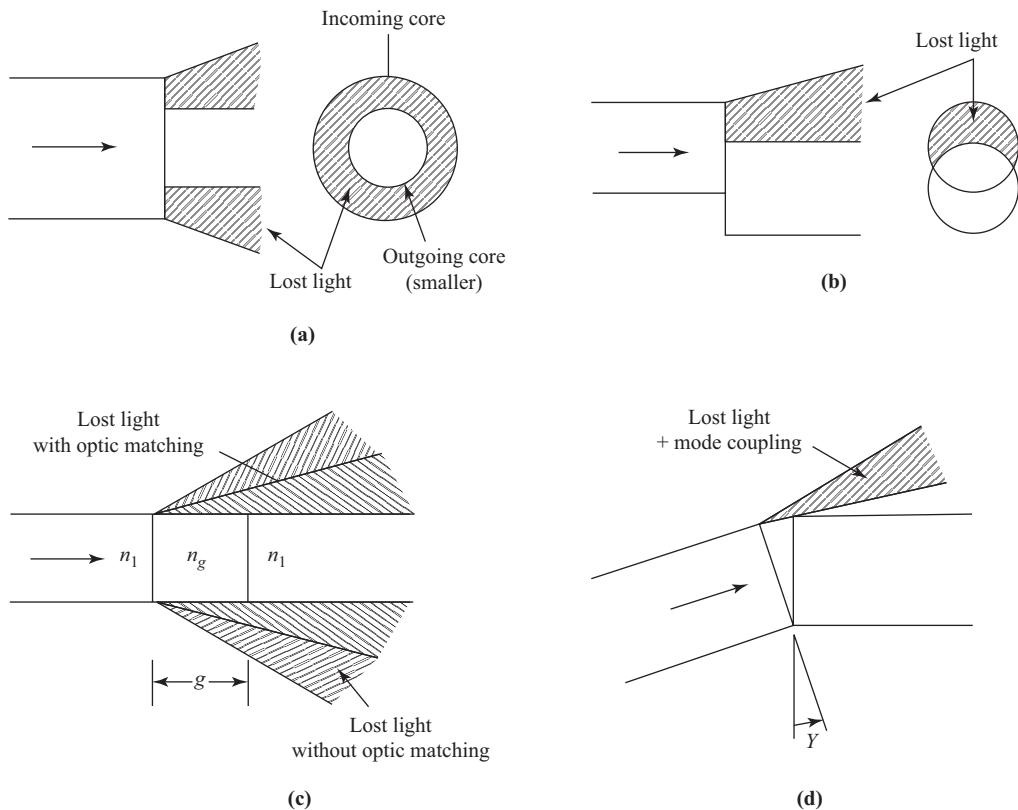


Figure 20.7.1 Connector loss factors. (a) Core size mismatch. (b) Lateral core misalignment. (c) Longitudinal gap and optical mismatch. (d) Angular misalignment.

Longitudinal Gap Separation. When light leaves the end of a fiber, it diverges in a cone that is determined by the acceptance angle of the fiber. If the mating fiber is not butted right up against the incoming fiber, light will be lost through this divergence. The amount of light lost increases as the length of the gap g increases, as shown in Fig. 20.7.1(c).

Optical Gap Losses. If the gap between the fibers contains air, light propagating through the gap must pass through two partially reflecting interfaces because of the change of refractive index in going from core n_1 to air n_g and then back to core n_1 . This causes what are called *Fresnel losses*, which are given by

$$\alpha = \left(\frac{n_g - n_1}{n_g + n_1} \right)^2 \quad (20.7.1)$$

Optical gap losses may be almost entirely eliminated by placing an optical matching cement in the gap that has the same refractive index as the core glass. In this case, $n_g = n_1$ and the loss becomes zero. Gap separation losses are also reduced by the index matching.

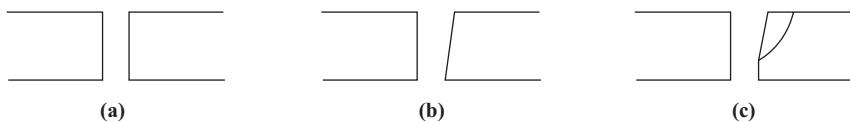


Figure 20.7.2 Fiber cleavage for joining. (a) Perfectly cleaved joint. (b) Joint with one fiber cleaved at an angle to the axis. (c) Joint with one fiber unevenly cleaved.

Angular Misalignment. If the two cores are misaligned so that they meet at an angle γ , as shown in Fig. 20.7.1(d), some light will escape through the open gap at one side of the joint. Moreover, light leaving one fiber may be coupled into lossy modes in the second fiber and be lost to leakage further along the fiber. Careful design and installation of connectors will minimize the angular misalignment losses.

Improper Fiber End Preparation. A properly prepared pair of fibers matched for joining is shown in Fig. 20.7.2(a). Both have been cleanly cleaved at right angles to the core axes and have smooth planar mating surfaces. Figure 20.7.2(b) shows a fiber cleaved at an angle to the core axis. Light striking this surface will reflect out of the fiber or be coupled into lossy modes, causing an increase in loss. Figure 20.7.2(c) shows a fiber that did not cleave evenly due to improper application of cleaving stress. The end of the fiber has an uneven scalloped surface or may even have spurs on it. These irregularities scatter light and increase losses.

Special jigs have been designed so that when a fiber is clamped into one of them it automatically scores the glass surface and applies just the right amount of force in the right direction to achieve clean cleavage. Such machines can be used successfully in field splicing in spite of poor environmental conditions.

Dirt. Any dirt or foreign substance that gets into a connector or splice during assembly may increase its losses or even completely block it. Extreme care has to be taken during installation to ensure that no dirt is included. In spite of this, successful field splicing in very dirty circumstances can be accomplished using special splicers and due care.

Connectors

The fiber connectors used at the points where fibers terminate in transmitters and receivers, either at the ends of the line or at intermediate repeaters, must have dismountable connectors to allow equipment to be removed for servicing. These connectors must automatically ensure a minimum connection loss and be easy to disconnect and connect.

A connector must eliminate the effects of angular and lateral misalignment and also ensure that the two fiber ends butt against each other in order to ensure a low-loss connection. Many arrangements have been used to build such connectors. Two of the more commonly used methods are illustrated here. Figure 20.7.3(a) shows the V-groove system, where the lateral and angular alignments are obtained by holding the fiber ends into the bottom of a V-groove cut in a block. The fibers must have the same dimensions or lateral core misalignment will occur. The fibers may be held in place either by a spring loading system or by sandwiching them between two V-blocks as shown. The V-blocks are made as parts of mating ferrules that press the fiber ends together when they are slid into each other.

The second method, shown in Fig. 20.7.3(b), uses a precision bored hole in a jewel bearing similar to the type used in watch making to provide self-alignment. One fiber is permanently fixed with its end face halfway through the bore hole of the jewel, which is mounted in one ferrule. The other fiber is held in a flexible rubber mount in the mating ferrule with its tip exposed. When the ferrules are pushed together, the

conical end of the jewel bore guides the mating fiber into place to butt against the fixed fiber. The rubber mount is pushed back and deformed to clamp the fiber in place. The result is a good low-loss connection that can be quickly disconnected as required.

Any good connector also provides for field adjustment of alignment. When connection is made in the field, received signal strength is monitored and the connector is adjusted to give maximum signal.

Connectors are rarely attached directly to the fibers in the field, but come with a short fiber attached. These pigtails can then be fusion spliced to the main fibers, eliminating critical field assembly procedures.

Fiber Splices

Fiber splicing procedures must meet two criteria. First, splices must be easy to make under the worst of environmental conditions. Second, they must provide a minimum of introduced loss. Two methods are discussed here.

The collapsed glass sleeve splice shown in Fig. 20.7.4(a) is easy to assemble in the field without the need for sophisticated equipment. The fibers are first prepared by removing short lengths of protective sheath from each and then cleaving both ends using a special hand-operated cleaving tool to provide smooth mating end faces. Next, a short piece of soft glass tubing (the sleeve) is slid onto one fiber end and softened with an oxy-hydrogen torch until it collapses around the fiber. After cooling, the mating fiber is pressed into the free end of the sleeve with some optical matching cement. After the cement has set, the entire splice is encased in protective heat-shrink tubing and replaced in the cable casing. Only simple hand tools are required, but care must be taken to avoid getting dirt or moisture into the splice. Splices of this type typically give about 0.5 dB of insertion loss.

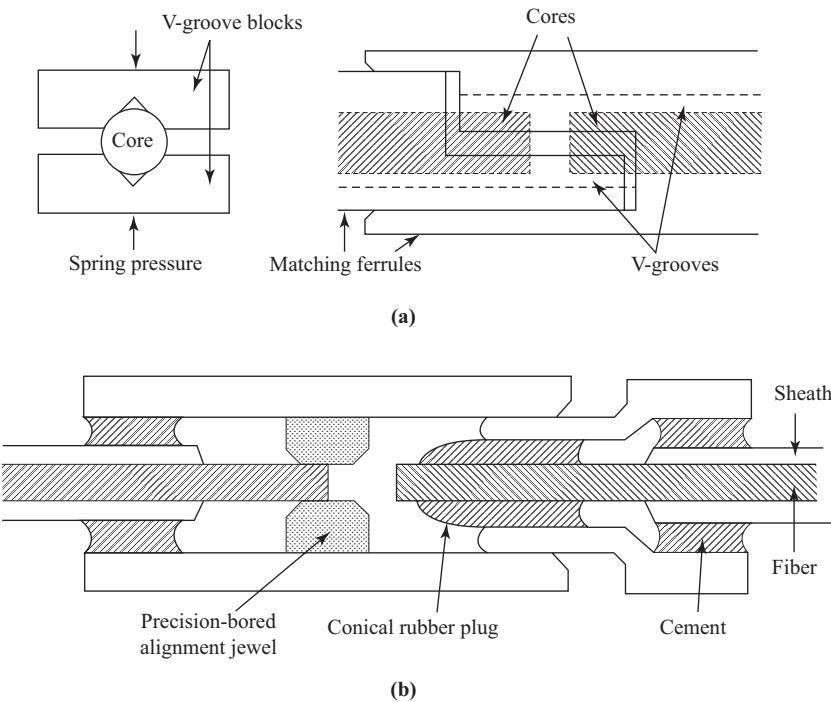


Figure 20.7.3 Coupling alignment methods. (a) V-groove alignment. (b) Precision-bored jewel alignment.

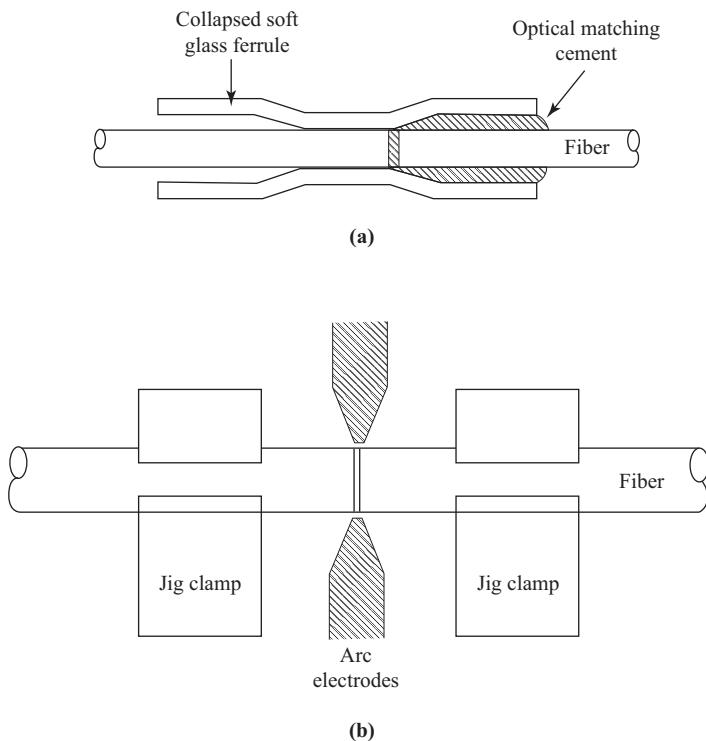


Figure 20.7.4 Fiber splices. (a) Collapsed glass sleeve splice. (b) Making a fusion splice.

Fusion splices give the lowest insertion losses of any method, but they are more difficult to make in the field. Figure 20.7.4(b) illustrates the fusion splicing process. The fiber ends are first stripped and cleaved in the manner described for sleeve splicing, or a special stripping and cleaving tool mounted in a splicing machine may be used. The prepared fiber ends are then placed in clamps on a jig that can be aligned using vernier screws. The fiber ends are butted and aligned using a binocular microscope, a process requiring considerable skill of the operator. Once aligned, the ends are heated either with a microtorch or with an electric arc until the two ends fuse together. This is a critical step since too much heat will cause the molten glass to flow and deform the fiber. Finally, the splice is encased in a protective sleeve and replaced in the cable. Low-loss splices are relatively easy to make, and typically insertion losses will be less than 0.1 dB.

20.8 Fiber-optic Communication Link

The length of a fiber-optic transmission line is limited almost entirely by its losses for low information rates. At higher rates, however, the length becomes limited by the amount of pulse dispersion that takes place in the fiber. Repeaters must be included in the fiber if either the loss limit or the dispersion limit is exceeded. The repeaters receive the signal from one link and amplify, reshape, retime, and retransmit the pulses into the next link.

The major components of a two-link fiber-optic line are shown in Fig. 20.8.1. At the sending end, an optical transmitter (a LED or laser diode) couples light into the fiber, turning it on and off according to

the bit stream to be sent. At the repeater, an attenuated and dispersed train of light pulses is detected by an avalanche photo diode (APD) or a *pin* diode and amplified, reshaped, and retimed before being sent on the second link. The receiver at the second terminal converts the light to electrical pulses for distribution after again being regenerated.

In a loss-limited fiber link, the total losses introduced by the fiber and all the connectors and splices along the fiber must not exceed the difference between the transmitted optical power P_t and the lowest acceptable received optical power P_r (both expressed in decibels) for a given type of detector and allowable error rate. In a single fiber link such as that shown in Fig. 20.8.1(b), total light flux or power P_t is emitted from the optical source device (a laser diode here). Because of the inefficiency in coupling the diode to the fiber end, which protrudes into an optical port on the diode, only part of the light from the source actually gets into the fiber. The result is an inserted port loss L_{pt} , which may be as high as a 6 or 8 dB. A short piece of fiber (a pigtail) is permanently attached to the optical port, with a quick-disconnect connector permanently attached to its end. This mates with another connector on the end of the line fiber. Typical connector insertion loss L_c is about 1 dB.

Light passing down the fiber encounters absorption and leakage losses at the rate of L_f (dB/km) for a total fiber loss of zL_f (dB), where z is the fiber length (km). It will also encounter N_s splices each with a loss L_s (dB) to give a total splice loss of $N_s L_s$ (dB).

At the receiving end (which may be a repeater), another pigtail, line connector, and port feed the light from the fiber into the avalanche photo diode, introducing a second port loss L_{pr} and a second connector loss L_c . A loss margin M (dB) of 5 or 10 dB is included to account for any unidentified losses or for increases of loss due to aging, bending, or extra splices introduced to repair accidental breaks.

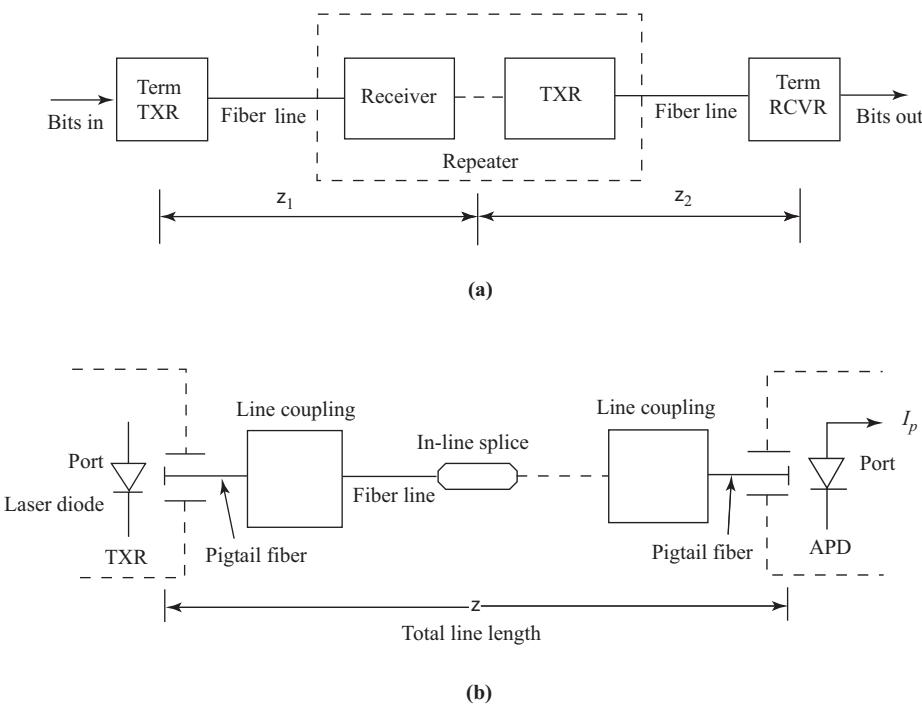


Figure 20.8.1 (a) A fiber optic line with two links and one repeater. (b) Elements in a typical fiber link.

The *loss budget* for the link, which is similar to that for a radio link, is found as the summation of all these components, with all units in decibels.

$$P_t - P_r = M + L_{pt} + L_{pr} + N_c L_c + N_s L_s + z L_f \quad (20.8.1)$$

Solving for z in this expression gives the maximum allowable link length as limited by losses.

If the bit rate of the link is to be high, then the fiber may be dispersion limited. In this case the total length can be determined by a conservative extension of Eq. (20.4.23) as

$$z = \frac{1}{(5B\Delta_t)} \quad (20.8.2)$$

where Δ_t is the time dispersion per unit length ($\mu\text{s}/\text{km}$) and B is the maximum allowed bit rate (Mbits/s) to give z in kilometers. The maximum allowed link length is the shorter of the loss-limited length and the dispersion limited length.

EXAMPLE 20.8.1

A fiber link is to have the following characteristics:

Transmitter: Laser diode minimum $P_t = 0 \text{ dBm}$ (1 mW)

Receiver: APD minimum $P_r = -57 \text{ dBm}$ at a maximum BER = 1 in 1,000,000,000

Port losses: $L_{pt}, L_{pr} = 6.0 \text{ dB}$ (each)

Connector losses (2): $L_c = 1.0 \text{ dB}$ (each)

Splices (5): $L_s = 0.5 \text{ dB}$ (each)

Fiber losses: $L_f = 2 \text{ dB/km}$

Loss margin: $M = 5 \text{ dB}$

Fiber total dispersion: $\Delta_t = 0.505 \text{ ns/km}$

Maximum bit rate for BER given: $B = 35 \text{ Mbits/s}$

- (a) Find the loss-limited fiber length.
- (b) Find the maximum bandwidth for the loss-limited length.
- (c) Find the dispersion-limited fiber length.

SOLUTION (a) From Eq. (20.8.1),

$$L_t - P_r = M + N_c L_c + N_s L_s + L_{pt} + L_{pr} + z L_f$$

$$0 - (-57) = 5 + (2 \times 1) + (5 \times 0.5) + 6 + 6 + 2z$$

$$z = \mathbf{17.8 \text{ km}}$$

(b) From Eq. (20.8.2),

$$B_{\max} = \frac{1}{5\Delta_t z} = \frac{1}{5 \times 0.000505 \times 17.8} = \mathbf{22.2 \text{ Mbps}}$$

which is less than the desired bit rate. The fiber is dispersion limited.

(c) The dispersion-limited length is given by

$$z = \frac{1}{5\Delta_t B} = \frac{1}{5 \times 0.000505 \times 35} = 11.3 \text{ km}$$

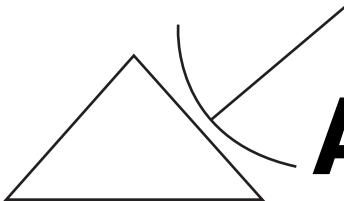
The repeater spacing may not exceed 11 km for this link.

PROBLEMS

- 20.1.** A glass-clad fiber is made with core glass of refractive index 1.500 and the cladding is doped to give a fractional index difference of 0.0005. Find (a) the cladding index, (b) the critical internal reflection angle, (c) the numerical aperture, and (d) the external critical acceptance angle.
- 20.2.** Core glass of index 1.6200 is to be used to make a step-index fiber with an acceptance cone half-angle of 5° . (a) What will the internal critical reflection angle be? (b) What should the cladding index be? (c) What fractional index difference does this give? (Beware of roundoff errors).
- 20.3.** (a) Explain the difference between a step-index fiber and a graded-index fiber, (b) What is a *W* profile fiber?
- 20.4.** (a) What is the advantage of using a graded-index core in a fiber? (b) Explain how that advantage is obtained.
- 20.5.** The fiber in Problem 20.1 has a core diameter of $50\text{ }\mu\text{m}$. (a) What is the *V* number of the fiber? How many modes will the fiber support?
- 20.6.** A step-index fiber is made with a core index of 1.52, a core diameter of $29\text{ }\mu\text{m}$, and a fractional index of 0.0007. It is operated at a wavelength of $1.3\text{ }\mu\text{m}$. Find (a) the fiber *V* number, (b) the number of modes the fiber will support, and (c) the name of each of these modes.
- 20.7.** A step-index fiber has a core index of 1.5000 and a fractional difference of 0.0005 and operates at a wavelength of $1.2\text{ }\mu\text{m}$. (a) What is the largest diameter the core may have if it is to operate as a single-mode fiber? (b) Which mode does a single-mode fiber propagate?
- 20.8.** A germanium doped silica fiber is to be used at a wavelength of $0.85\text{ }\mu\text{m}$. Find the approximate loss factors caused by (a) scattering and (b) absorption. (c) Assuming maximum mode coupling loss of 0.5 dB/km , what will be the total loss factor for the fiber? (d) If maximum loss in the fiber is to be less than 30 dB , how long can the fiber be?
- 20.9.** Explain how energy is lost from a leaky mode.
- 20.10.** (a) Explain how mode coupling can occur. (b) Explain how mode coupling causes losses.
- 20.11.** Explain how energy is lost at a sharp bend in a fiber.
- 20.12.** (a) At what wavelength is a glass fiber likely to have its lowest loss? (b) What would be a typical value for that lowest loss and what would be its main cause? (c) Why would a fiber not likely be used at a wavelength of $0.94\text{ }\mu\text{m}$ or $1.38\text{ }\mu\text{m}$?
- 20.13.** Explain the term *dispersion* as applied to optical fibers.
- 20.14.** A step-index multimode fiber has a core index of 1.5000 and a cladding index of 1.49800. Find (a) the intermodal dispersion factor for the fiber, (b) the total dispersion in a 18-km length, and (c) the maximum bit rate allowed, assuming dispersion limiting.
- 20.15.** If the fiber in Problem 20.14 is operated at $1.1\text{ }\mu\text{m}$ with a light source of 1.5 nm spectral bandwidth, what is its total material dispersion factor?
- 20.16.** A step-index fiber operated at $1.3\text{ }\mu\text{m}$ is found to have dispersion factors of 10 ns/km intermodal, 0.5 ns/km material, and 0.8 ns/km waveguide. The fiber is 15 km long and only propagates one mode. (a) What is the total dispersion? (b) What maximum bit rate will it handle?
- 20.17.** If a dispersion-limited fiber with a total dispersion factor of 3.3 ns/Km is used at 1000 Mbits/s , how long can it be without repeaters?
- 20.18.** A laser diode produces a dominant wavelength of $1.55\text{ }\mu\text{m}$. Find the spacing for a Bragg reflector using (a) the first Bragg mode or (b) the second Bragg mode, if the effective index of refraction within the guiding layer is 3.50.

- 20.19.** A light emitting diod is made with gallium arsenide doped with aluminum and has an energy band gap of 1.55 eV. What is its dominant emission wavelength?
- 20.20.** A laser diode has an energy band gap of 1.1 eV, but experiences a 20% increase in emitted wavelength when operated as a laser. What wavelength does it emit?
- 20.21.** Explain how a light emitting diode generates light.
- 20.22.** Explain the term *stimulated emission* as applied to laser diodes.
- 20.23.** What advantages do semiconductor laser diodes have over light emitting diodes when used for fiber transmission?
- 20.24.** What is a stripe laser? What advantage does it have?
- 20.25.** (a) Explain how a *pin* photodiode works. (b) How is this diode an improvement over the *pn* diode?
- 20.26.** Can a *pin* diode be used as an avalanche diode? (b) If so, why would it not be so used?
- 20.27.** (a) Explain how an avalanche diode works. (b) How does the avalanche diode compare to the *pin* diode for sensitivity? (c) What limits the ultimate sensitivity of an avalanche diode?
- 20.28.** (a) Describe the spectral response of a typical silicon *pin* diode, (b) What type of diode might be used for a 1.3- μm receiver?
- 20.29.** (a) List the factors that are most likely to affect the quality of a fiber connector. (b) List those that may affect the quality of a splice.
- 20.30.** (a) What precautions should be taken when installing or repairing fiber equipment, connectors, or splices? (b) Explain why these precautions are so important in terms of impairment of transmission.
- 20.31.** (a) Which splice technique gives the lowest insertion losses? (b) Which is the easiest done in the field? Why? (c) List typical insertion losses for the two types of splices.
- 20.32.** List the steps involved in making a collapsed sleeve splice in a manhole full of water.
- 20.33.** A fiber link is to be installed using the following items, (a) Find the maximum repeater spacing. (b) Is the spacing loss limited or dispersion limited?
- [1] A laser diode with power output of 250 μW
 - [2] A pin diode detector with a sensitivity of -45 dBm
 - [3] A transmitter port loss of 5 dB
 - [4] A receiver port loss of 2.5 dB
 - [5] Two terminal connectors with 1.5-dB loss each
 - [6] Ten splices with losses of 0.15 dB each
 - [7] Fiber losses of 3 dB/km
 - [8] Fiber dispersion of 0.9 ns/km
 - [9] System loss margin of 5 dB
 - [10] Maximum bit rate of 15 Mbit/s
- 20.34.** Plot, using MATLAB/Mathematica/Octave, the *normalized* variation of index of refraction with radius, for a *graded index* fiber. (Hint; Plot equation 20.2.12).
- 20.35.** Calculate the transmission distance over which the optical power will attenuate by a factor of 10 for three fibers of 0.2, 40, and 3000 dB/km. Assuming that the optical power decreases as $\exp(-\alpha L)$, calculate α (in cm^{-1}) and plot attenuation as a function of distance from source, for the above three cases.
- 20.36.** Assume that a digital communication system can be operated at a bit rate of up to 1% of the carrier frequency. How many audio channels at 64 kb/s can be transmitted over a microwave carrier at 5 GHz and an optical carrier at 1.55 μm ?

- 20.37.** A 2-hour lecture script is stored on the computer hard disc in 8 bit ASCII format. Calculate the total number of bits assuming a delivery rate of 300 words per minute and on average 6 letters per word. How long will it take to transmit the script at a bit rate of 1 Gb/s?
- 20.38.** A $1.55 \mu\text{m}$ digital communication system operating at 1 Gb/s receives an average power of -45 dBm at the detector. Assuming that zeros and ones are transmitted with equal probability, calculate the number of photons received within each one bit.
- 20.39.** A distribution network uses an optical bus to distribute the signal to 10 users. Each optical tap couples 10% of the power to the user and has 1 dB insertion loss. Assuming that the station 1 transmits 1 mW of power over the optical bus, calculate the power received by the stations, 7, 8, 9 and 10.
- 20.40.** Prove that the *rise time* T_r and the 3 dB bandwidth Δf of a RC circuit are related by $T_r \Delta f = 0.35$.



Appendix A

Logarithmic Units

A.1 The Decibel

If two powers P_1 and P_2 differ, their difference can be expressed in decibels as

$$D = 10 \log_{10} \left(\frac{P_1}{P_2} \right) \quad (\text{A.1})$$

A positive number of decibels indicates that P_1 is greater than P_2 , a negative number, that P_1 is less than P_2 . For example, 50 W is greater than 10 W by $D = 10 \log_{10}(50/10) = 7$ decibels (dB), whereas 10 W is less than 50 W by $D = 10 \log_{10}(10/50) = -7$ dB. The decibel is one-tenth of a larger unit, the bel, defined as $\log_{10}(P_1/P_2)$; in practice, the bel is inconveniently large, and the decibel is the preferred unit.

Although the decibel is based on power ratios, it can also be used to express voltage and current ratios. Let the power P_1 be developed in a resistor R_1 across which the voltage is V_1 and the current through which is I_1 ; let the corresponding quantities for P_2 be R_2 , V_2 , and I_2 . Then

$$\begin{aligned} D &= 10 \log_{10} \left(\frac{V_1^2/R_1}{V_2^2/R_2} \right) \\ &= 20 \log_{10} \left(\frac{V_1}{V_2} \right) + 10 \log_{10} \left(\frac{R_2}{R_1} \right) \end{aligned} \quad (\text{A.2})$$

By similar reasoning, it is easily shown that D can be expressed as

$$D = 20 \log_{10} \left(\frac{I_1}{I_2} \right) + 10 \log_{10} \left(\frac{R_1}{R_2} \right) \quad (\text{A.3})$$

Therefore, it can be seen that a knowledge of the resistance values is needed, in addition to that of voltage or current values, in order to determine the decibel value. Because of the widespread use of the decibel, its

meaning has been extended to cover voltage and current ratios. By *definition*, if two voltages V_1 and V_2 differ, their difference D_v in decibels is given by

$$D_v = 20 \log_{10} \left(\frac{V_1}{V_2} \right) \quad (\text{A.4})$$

Similarly, if two currents I_1 and I_2 differ, their difference D_i in decibels is

$$D_i = 20 \log_{10} \left(\frac{I_1}{I_2} \right) \quad (\text{A.5})$$

As already shown, D_i and D_v can only be related to D when the ratio R_1/R_2 is known.

It is meaningless to state absolute values of voltage, current, or power in decibels. The decibel represents a ratio, and therefore a reference level must also be stated if absolute values are required. For example, the phrase “a power of 20 dB” is meaningless, but “20 dB relative to 1 W” means 100 W. Very often the reference level is implicitly understood. A common example of this is where noise levels are quoted in decibels, the reference level being the threshold of hearing. Another common example, in telecommunications engineering, is the selectivity curve of a tuned circuit, where the resonance value is used as reference. (see Fig. 1.3.3).

Sometimes the reference level is indicated in abbreviated form, the best example of this being the dBmW (or dBm), meaning “decibels relative to 1 milliwatt.” Other examples are dB μ V and dBV, these being decibels relative to 1 microvolt and 1 volt, respectively.

A.2 The Decilog

In link power budget calculations, it is often necessary to take the logarithm of quantities that are not power (or voltage and current) ratios. The logarithm must always involve the ratio of two similar quantities. Let the ratio be denoted by X , and in *decilogs*, by $[X]$; then

$$[X] = 10 \log X \quad (\text{A.6})$$

Where the quantities do not occur as ratios, the unit quantity is taken as reference. For example, a bandwidth of 36 MHz expressed in decilogs is

$$[B] = 10 \log \frac{36 \times 10^6 \text{ Hz}}{1 \text{ Hz}} = 75.56 \text{ decilogs}$$

Because of the widespread use of the decibel, a clear distinction is not always observed between decilogs and decibels, and in the above example B may be written as 75.56 dBHz, meaning 75.56 decibels (or decilogs) relative to 1 Hz. Other common examples are temperature, where the unit is taken as 1 K, and bit rate, where the unit is taken as 1 bps.

Often, where the unit abbreviation proves to be clumsy, it is simply shown as dB. For example, Boltzmann's constant is $k = 1.38 \times 10^{-23} \text{ J/K}$ and expressed in decilogs is

$$[k] = 10 \log \frac{k}{1 \text{ J/K}} = 10 \log 1.38 \times 10^{-23} = -228.6 \text{ decilogs}$$

This is usually written as -228.6 dB, whereas it should be dBJpK.

A.3 The Neper

The neper is a logarithmic unit originally introduced to express the attenuation of current along a transmission line, using natural (or neperian, hence the name “neper”) logarithms rather than common logarithms. If two currents I_1 and I_2 differ, their difference N in nepers is

$$N = \ln\left(\frac{I_1}{I_2}\right) = \log_e\left(\frac{I_1}{I_2}\right) \quad (\text{A.7})$$

The relationship between decibels and nepers is easily established. Equation (A.5) gives

$$D_i = 20 \log_{10}\left(\frac{I_1}{I_2}\right)$$

and, on changing the logarithmic base,

$$\begin{aligned} D_i &= (20 \log_{10}e) \left(\ln \frac{I_1}{I_2} \right) \\ &= 8.686 N \end{aligned} \quad (\text{A.8})$$

EXAMPLE A.1

Express (a) 3 nepers in decibels and (b) 3 decibels in nepers.

SOLUTION (a) $D = 8.686 \times 3$

$$= 26.06 \text{ dB}$$

$$(b) N = \frac{3}{8.686}$$

$$= 0.345 \text{ N}$$

A.4 Logarithmic Scales

By graduating graph axes in lengths proportional to the logarithms of numbers rather than to the numbers themselves, a much wider range of values can be accommodated on a given scale. Scales based on common logarithms are usually employed. Figure A.1 illustrates graph paper that has one axis graduated logarithmically and the other linearly. This is known as *semilog* graph paper, and because the logarithmic scale pattern repeats itself four times, it is referred to as *four-cycle*. The logarithmic scale of Fig. A.1 can accommodate values ranging over four orders of magnitude (for example, 1 to 10^4 , 10 to 10^5 , 0.01 to 10^2 , and so on). It will be apparent that a zero origin cannot be shown on a logarithmic scale, and care must be taken when interpreting graphs plotted on logarithmic scales. Also, the slope of the graph must be carefully interpreted, as it involves the logarithms of numbers. Figure A.1 shows the decibel equations (D , D_i , and D_v) plotted

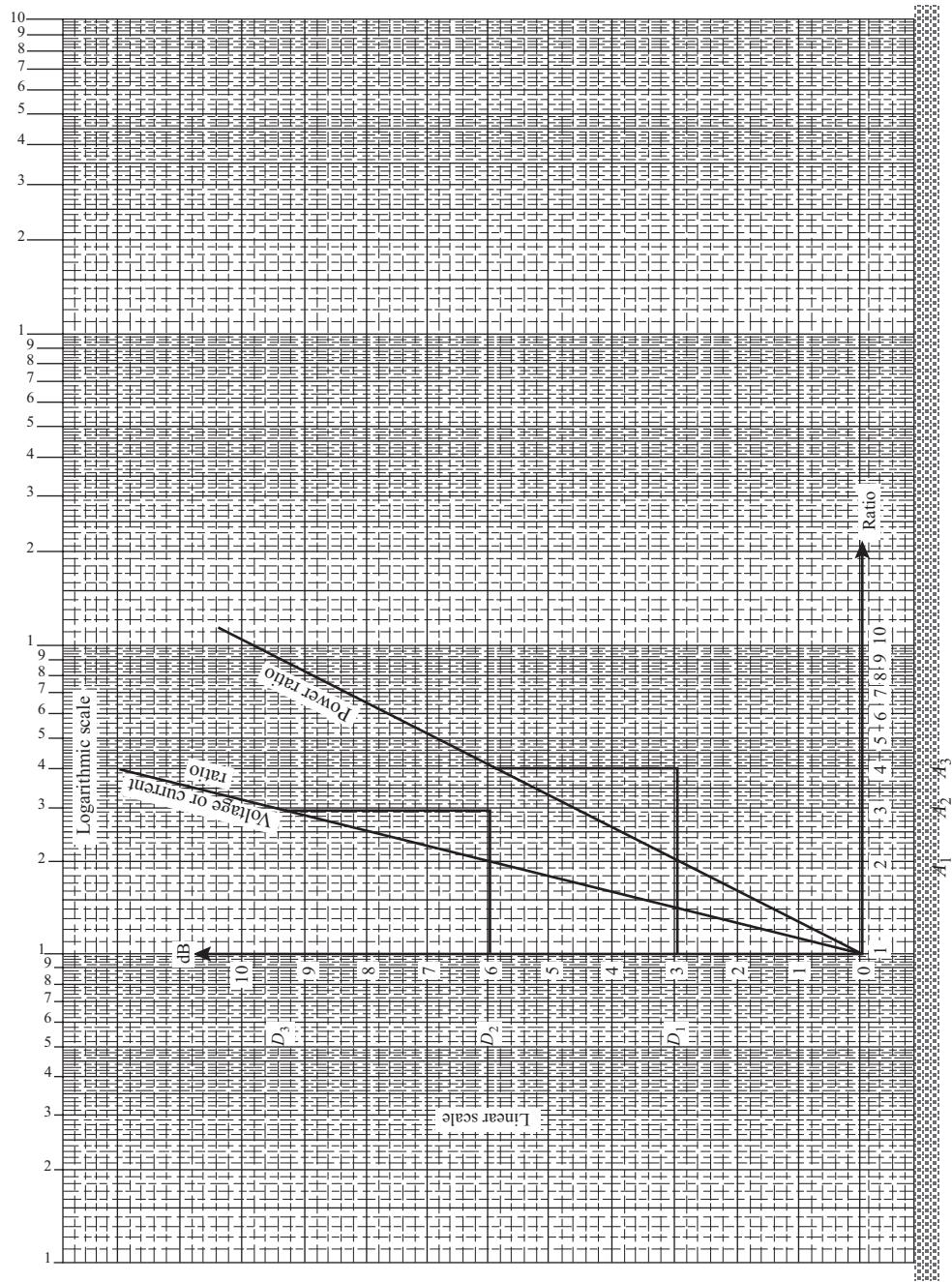


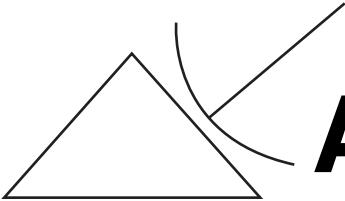
Figure A.1 Conversion chart (ratios to decibels) using semilog graph paper.

against a ratio scale, which is logarithmic. The slope of the *voltage or current ratio line* should be 20, and this is obtained from

$$\begin{aligned} \text{slope} &= \frac{D_3 - D_2}{\log_{10}X_2 - \log_{10}X_1} \quad \left(\text{not } \frac{D_3 - D_2}{X_2 - X_1} \right) \\ &= \frac{9.5 - 6}{0.4771 - 0.3010} \\ &= 20 \end{aligned} \tag{A.9}$$

Similarly, the slope of the *power ratio line* is

$$\begin{aligned} \text{slope} &= \frac{D_2 - D_1}{\log_{10}X_3 - \log_{10}X_1} \quad \left(\text{not } \frac{D_2 - D_1}{X_3 - X_1} \right) \\ &= 10 \end{aligned} \tag{A. 10}$$



Appendix B

The Transverse Electromagnetic Wave

Electromagnetic energy is propagated through space and is guided along transmission lines in the form of a *transverse electromagnetic wave* (TEM wave). For this type of wave, the electric field E (V/m), the magnetic field H (A/m), and the direction of propagation x , along which the wave travels with phase velocity v_p (m/s), are mutually at right angles, as shown in Fig. B.1(a). If the TEM wave is reversed in direction, then either the E or the H field [Fig. B.1(b)] must reverse. This is similar to the condition required when reversing direction of rotation of a dc generator while maintaining same polarity of induced emf.

To illustrate some of the basic properties of the TEM wave, a sinusoidal variation will be assumed; the field equations are then

$$e = E_{\max} \sin(\omega t - \beta x) \text{ V/m} \quad (\text{B.1})$$

$$h = H_{\max} \sin(\omega t - \beta x) \text{ A/m} \quad (\text{B.2})$$

where $\omega = 2\pi f = 2\pi/T$, and $\beta = 2\pi/\lambda$. The periodic time T and the wavelength λ are shown in Fig. B.1(c) and (d).

Clearly, since one cycle occurs in T seconds, the number of cycles in one second, which is the frequency f in hertz, is

$$f = \frac{1}{T} \quad (\text{B.3})$$

Also, since the wave generates f cycles in 1 s and covers a distance v_p meters in 1 s, the wavelength λ , which is the distance spanned by one cycle, is

$$\lambda = \frac{v_p}{f}$$

or

$$\lambda f = v_p \quad (\text{B.4})$$

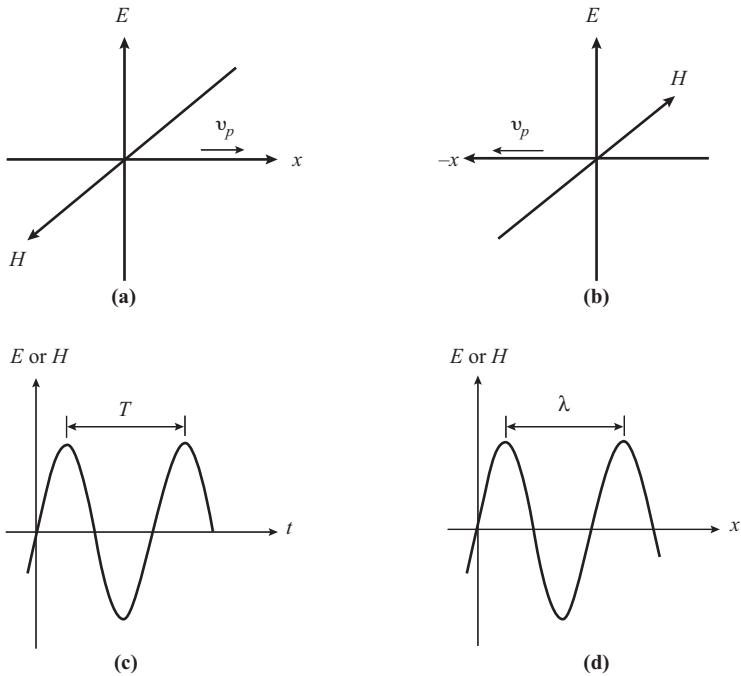


Figure B.1 Transverse electromagnetic (TEM) waves.

In the following equations, rms values, indicated by E , H , B , and D , will be used for simplicity.

The electric field E will be accompanied by an electric flux density D (C/m^2) given by

$$D = \epsilon E \quad (\text{B.5})$$

The magnetic field H will be accompanied by a magnetic flux density B (teslas), given by

$$B = \mu H \quad (\text{B.6})$$

Here, ϵ is the absolute permittivity and μ the absolute permeability of the medium in which the wave travels.

Two further experimental facts are (1) an electric field density D moving at velocity v_p will induce a magnetic field H given by

$$H = Dv_p \quad (\text{B.7})$$

and (2) a magnetic field density B moving at velocity v_p will induce an electric field E given by

$$E = Bv_p \quad (\text{B.8})$$

The vector directions in each case are shown in Fig. B.1(a). The fact that each type of field, when moving, generates the other in the proper direction gives rise to a self-sustaining electromagnetic wave, once the initial disturbance of either type of field is made.

The relationships in Eqs. (B.7) and (B.8) may not be familiar, but a common application of the latter is in the electromechanical generation of an emf by rotating armature conductors to cut a magnetic field at right angles.

From the four field equations (B.5), (B.6), (B.7), and (B.8), the following important relationships are easily obtained.

$$\text{Wave Impedance: } Z_0 = \frac{E}{H} \Omega \quad (\text{B.9})$$

$$= \sqrt{\frac{\mu}{\epsilon}} \quad (\text{B.10})$$

This should be compared with Eqs. (13.4.4) and (13.4.5).

$$\text{Phase velocity: } v_p = \frac{1}{\sqrt{\mu\epsilon}} \text{ m/s} \quad (\text{B.11})$$

Average power density in wave:

$$\begin{aligned} P_D &= EH \text{ W/m}^2 \\ &= \frac{E^2}{Z_0} \\ &= H^2 Z_0 \end{aligned} \quad (\text{B.12})$$

These relationships should be compared with the circuit equations relating V , I , R , and power P .

This page is intentionally left blank.

Index

Acceptance angle, 657

Acceptance cone half-angle, 657

Active filters

overview, 163

RC low-pass, 164–165

Switched capacitor, 165–169

Adaptive delta modulation, 354

All-pass filters, 37

Alpha profile function, 661

Amplitude demodulator circuits, 240–241

diagonal peak clipping, 241–243

negative peak clipping, 243–244

Amplitude limiters, 319–320

Amplitude modulation (AM)

amplitude demodulator circuits, 240–244

amplitude-modulated transmitters, 244–247

average power for sinusoidal, 231–232

circuits, 236–240

double-sideband suppressed carrier

(DSBSC), 235

effective voltage and current for sinusoidal, 232–233

frequency spectrum for sinusoidal, 228–231

index, 225–228

modulation index for sinusoidal, 228

and noise, 252–257

nonsinusoidal, 233–235

overview, 223–224

receivers, 247–252

transmitters, 244–247

Amplitude shift keying, 376

Analog communications, 362

Analog fax transmission, 587–592

Angle modulation, 223

amplitude limiters, 319–320

automatic frequency control, 318

average power in, 291–292

circuits, 297–305

detectors, 309–318

digital phase modulation, 297

equivalence between PM and FM, 294–296

FM, 283–285

FM broadcast receivers, 325–327

FM stereo receivers, 328–330

FM transmitters, 305–309

frequency spectrum for sinusoidal FM, 287–290

measurement of modulation index for sinusoidal

FM, 293

noise in FM systems, 320–324

non-sinusoidal modulation: deviation ratio, 292

overview, 283

phase modulation, 293–294

pre-emphasis and de-emphasis, 324–325

sinusoidal FM, 285–287

sinusoidal phase modulation, 296

Antenna misalignment loss (AML), 643

Antenna pointing loss, 643

Antennas

aperture, 505

broadside array of, 532

coordinate systems, 509–510

dielectric lens, 543–544

discone, 536

driven arrays, 530–534

effective area of, 515–516

effective length of, 516–517

end-fire array, 532–533

equivalent circuits, 505–509

ferrite-rod, 528–529

folded elements of, 526–527

grounded vertical, 524–526

half-wave dipole, 520–524

helical, 536

Hertzian dipole, 518–520

horns, 538

isotropic radiator, 512

linear array of, 530–531

log periodic, 531

long-wire, 529–530

look angles, 627

loop, 527–528

Marconi, 524

microwave, 538–545

nonresonant, 529–530

notch, 545

overview of, 505

paraboloidal reflector, 538–543

parasitic directors, 535

parasitic reflectors, 534

plane reflector arrays, 536

polarization, 510–512

power gain of, 513–515

- Antennas (*Contd.*)
radiation fields, 510
resonant, 505
rhombic, 530
slot, 544–545
turnstile, 534
unipole, 525
vertical, 523–526
VHF-UHF, 536–537
Yagi-Uda array, 535–536
- Antialiasing filter, 338
- Aperture antennas, 505
- Aperture effect, 340
- Aperture efficiency, 540
- Apogee, 622
- Armstrong method, 307
- Ascending node, 622
- Aspect ratio, 604
- Asynchronous transfer mode (ATM), 364
- Asynchronous transmission, 363–364
- Atmospheric absorption losses, 643
- Attenuation, 461
- Attenuation coefficient, 412
- Attenuator pads, 1–9
- Automatic frequency control (AFC), 318
- Automatic gain control (AGC), 212–214
- Automatic repeat request (ARQ), 393
- Available average power, 107
- Available power gain, 153–154
- Avalanche multiplication effect, 691
- Avalanche noise, 117
- Avalanche photodiode (APD), 691, 692
- Avalanche region, 117
- Backlobe radiation, 539–540
- Backoff, 642
- Balanced modulators, 162–164, 235
doubly balanced diode ring modulator, 265–266
FET singly balanced modulator circuit, 264–265
IC doubly balanced modulator, 265–266
- Balanced slope detector, 310
- Baluns, 408
- Band-pass filters, 36, 37
- Band-pass waveforms, 362
- Band-stop filters, 36, 37
- Bandwidth dispersion product (BDP), 674
- Bandwidth distance product (BDP), 681
- Barkhausen criterion, 173
- Baseband transmission, 364–368
- Baseband waveforms, 362
- Baudot, Emile, 83
- Bauds, 83
- Binary digital coding, 82
- Binits, 82
- Bipolar code, 90
- Bipolar junction transistor mixer, 160
- Bipolar transistor noise, 118
- Bit-error rate (BER), 205, 365, 391, 398
- Bits, 82
- Bit-timing recovery, 372
- Block codes, 392–394
- BORST, 552
- Bose-Chaudhuri-Hocquenghem (BCH) block code, 399
- Bragg reflector, 688
- Broadcast fading zone, 494–495
- Broadside arrays, 532
- Bulk acoustic waves, 44
- Buried heterostructure diode, 687
- Burst noise, 117
- Butterworth response, 38
- Capacitive tap, 29–32
- Carrier recovery circuits, 379, 386–388
- Carrier synchronization, 362
- Carrier-to-noise ratio (C/N), 643
- Carrier wave, 223
- Carson's rule, 292
- Cascode amplifier, 154–155
- Cassegrain feed system, 542
- Cass horn, 542
- CATV (cable television), 594
- Cauer filter, 39
- Ceramic filters, 46, 212
- Channel encoder, 390
- Characteristic impedance, 409, 410–412
- Chebyshev response, 39
- Check bits, 397
- Chromatic dispersion, 677–679
- Circuits
AM, 236–240
amplitude demodulator, 240–244
antenna, 505–509
carrier recovery, 386–388
parallel tuned, 15–17
passive, 1–36
series-tuned, 9–15
synchronously tuned, 24
trunk, 572–573
tuned primary untuned secondary, 26–27
- Circular polarization, 511
- Clapp oscillator, 183–184
- Codecs, 347
- Code division multiple access (CDMA), 650
- Codewords, 391
- Codeword synchronization, 362

- Coding gain, 399
Coherent detection, 256, 379
Coincidence, 354
Collinear broadside array, 532
Colpitt oscillator, 179–183
Common-base amplifier, 151–153
Common-emitter (CE) amplifier, 141–150
Companding, 280
Complementary error function, 365
Compression, 345–348
Compression ratio, 280
Conical horn, 538
Conjugate match, 33
Connector losses, 694–696
Connectors, 696–697
Continuously variable slope delta (CVSD) modulator, 354
Continuous phase frequency shift keying (CPFSK), 381
Contour maps, 478
Control electrode, 598
Conversion loss, 159
Conversion transconductance, 160
Convolution encoding, 400–403
Coordinate systems, 509–510
Costa loop, 386
Crater lamp, 586
Critical frequency, 482–483
Critically coupled, 24
Crossbar switching, 565–569
Cross-point switching, 569
CRT (cathode ray tube), 598–602
Crystal lattice filters, 212
Crystal oscillator, 185–186
Cyclical redundancy codes (CRC), 399
Cylindrical scanning, 578

Data compaction ratios, 593
Data words, 391
Decca navigational system, 498
Decibel, 642, 704–705
Decilog, 705
Declination, 633
De-emphasis network, 324–325
Dellinger fadeouts, 491
Delta function, 68
Delta modulation, 353–354
Demultiplexer, 338
Detectors, 309–318
Deterministic waveforms, 74
Deviation ratio, 292
Dielectric lens, 543–544
Differential encoding 388
Differential phase shift keying (DPSK), 388–390

Differential pulse code modulation (DPCM), 352
Diffraction, 493–494
Digital carrier transmission, 648
asynchronous transmission, 362–364
bit-timing recovery, 372–374
carrier recovery circuits, 386–388
differential phase shift keying (DPSK), 388–390
digital carrier systems, 375–386
error control coding, 390–403
eye diagrams, 374–375
hard and soft decision coders, 390
matched filter, 368–372
optimum terminal filters, 372
overview, 361–362
probability of bit error in baseband transmission, 364–368
synchronization, 362
Digital decimation filter, 355
Digital fax transmission, 592–593
Digital line waveforms, 82–101
Digital phase modulation, 297
Digital switching, 569–572
Diode envelope detector, 240
Diode mixer, 159
Direct distance dialing (DDD), 562
Directive gain, 513
Discone, 536
Dispersion, 673–681
Distributed Bragg reflector (DBR), 689
Distributed feedback (DFB), 688
Dot notation method, 20
Double conversion receiver, 214–216
Double-sideband suppressed carrier (DSBSC), 235–236
Double-sided spectrum, 62
Double spotting, 208
Driven arrays, 530–534
Dry silver paper, 587
Duddell mirror oscilloscope, 586
Duplex transmission, 548
Dynamic impedance, 15

Early-late gate, 372
Echo suppressors, 572
Effective area, 515–516
Effective length 516–518
Electrolytic paper, 587
Electromechanical filters, 46
Electronically tuned receivers (ETRs), 216–218
Electronic CCD scanning, 578–581
Electronic telephone, 551–552
Electroresistance recording system, 587
Electrothermal recording, 587
Elliptical polarization, 511

- Elliptic filter. *See* Cauer filter
 End-fire arrays, 532–533
 Envelope, 224
 Equalizer filter, 340, 349, 372
 Equatorial plane, 509
 Equivalent input noise generators, 118–120
 Equivalent input noise resistance, 119
 Equivalent input shot noise current, 119
 Equivalent isotropic radiated power (EIRP), 643
 Error control coding, 390–391
 Excess noise ratio (ENR), 129
 Exponential Fourier series, 63
 Extremely low frequency (ELF) propagation, 498
 Eye diagrams, 374–375
 Facsimile transmission, 576
 Fast Fourier transform, 63, 68–72
 Feedthrough loss, 557–558
 Ferrite-rod antenna, 528–529
 Fiber-optic communications
 connectors and splices, 694–698
 dispersion, 673–681
 light sources for, 682–691
 links, 698–700
 losses in fibers, 668–673
 overview, 654
 photodetectors, 691–694
 principles of light transmission in a fiber, 654–668
 Fiber splicing, 697–698
 Fictitious noise generators, 119
 Field-effect transistor (FET)
 hybrid PI equivalent circuit for a, 155–158
 mixers, 158–163
 noise, 118
 Field strength polar diagram, 522–523
 Filler(s)
 active, 163–164
 antialiasing, 338, 341
 band-pass, 36, 38
 band stop, 36–38
 Cauer, 39
 ceramic, 46, 212
 crystal lattice, 42, 212
 design, 36
 digital decimation, 355
 electromechanical, 46
 equalizer, 340, 349, 372
 high pass, 36, 37
 LC, 39–40
 low pass, 36
 matched, 368–371
 mechanical, 212
 optimum terminal, 372
 passive, 36–39
 piezoelectric crystal, 40–44
 RC low pass, 164–165
 reconstruction, 350
 surface acoustic wave, 44–46
 switched capacitor, 165–169
 syllabic, 354
 transfer, 35
 First link, 121
 Flat-bed scanning, 578
 Flat-topped samples, 340
 Flicker noise, 116–117
 Folded elements, 526–527
 Fold-over value, 69, 71
 Footprints, 624
 Forward error correction (FEC), 390, 394, 397, 405
 Forward transit time, 138, 140
 Foster-Seeley discriminator 310–313
 Fourier, Joseph, 54
 Fourier coefficients, 55
 Fourier transforms
 energy signals and, 65–68
 fast Fourier transform (fft) computer programs, 68–71
 inverse, 71–72
 Fourier trigonometric series, 56
 Four-wire terminating sets, 558
 Four-wire transmission, 560–561
 Frame sync-bit, 350
 Frame synchronization, 362
 Frame-synchronizing signal, 350
 Framing codewords, 374
 Frequency
 lower side, 230, 263
 upper side, 230, 263
 Frequency deviation constant, 283
 Frequency division multiple access (FDMA), 641, 649
 Frequency division multiplexing, 273–278
 Frequency modulation (FM), 283–285
 broadcast receivers, 527–528
 stereo receivers, 328–330
 transmitters, 305–309
 Frequency selectivity, 10, 12
 Frequency shift keying, 380
 Frequency spectrum for sinusoidal AM, 228–287
 Frequency synthesizers
 applications of, 195
 overview, 193–195
 and the phase locked loop, 193–194
 and prescaling, 194–195
 Fresnel losses, 695

- Friis's formula, 127
Fundamental frequency, 53, 589
- Gain, 685
Gain guided laser, 686
Geometric ray tracing, 655
Geostationary orbit, 623, 626
Global Positioning Satellite (GPS), 622
Golay block code, 399
Graded index fiber, 659, 702
Granular noise, 353–354
Grating reflectors, 46
Grounded vertical antennas, 524
Ground reflections, 523–524
Ground wave, 493–495
Group delay time, 417
Group velocity, 416–417, 457
Gyrofrequency, 489–490
- Half-wave dipole, 520–522
Hamming block code, 399
Hard-decision decoding, 390
Hartley oscillator, 184–185
HDTV (high-definition television), 594
Height-gain wave, 494
Helical antennas, 536
High-frequency radio systems, 493–494
High-frequency transformers, 22
High-pass filters, 36
Hog horn, 542
Horizontal polarization, 510
Horns, 538
Hybrid-PI equivalent circuit
 for the BJT, 136–139
 for an FET, 178–182
- Ideal low-pass spectrum, 99
Idle noise, 280
Illumination efficiency, 540
Image, 206
Image rejection, 206
Impulse waveform, 354
Inclination, 622
Inclined highly elliptical orbit, 622
Independent sideband (ISB), 273
Index-guided laser diode (ILD), 688
Index of cooperation (IOC), 583
Index profile, 659
Infrared absorption, 688
In-phase component, 384
Input backoff, 645–646
Insertion loss, 1
- Instantaneous PAM, 340
Integrate and dump circuit, 370
Integrated circuit (IC) balanced mixer, 161
Integrated-circuit receivers, 218–221
Integrated services digital networks (ISDNs), 364
Interdigital transducers (IDTs), 45
Interference fading, 492
Intermediate frequency (IF), 198
Intermodal dispersion, 674
Intermodulation distortion, 640
International Telegraph and Telephone Consultative Committee (CCITT), 276, 640
Interrupted continuous wave (ICW), 376
Intersymbol interference, 96
Inverse fast Fourier transform (ifft), 71
Inverse Fourier transforms, 71
Ionosscatter 491
Ionosonde, 489
Ionospheric propagation, 482–493
Ion resonance (IR) absorption, 668
Isotropic radiator, 512
- Johnson noise. *See* Thermal noise
- Lanes, 498
Large radius bending, 671
Lasers, 586, 682
Laser threshold, 686
L-attenuator, 6–9
LC oscillators, 178–179
LC filters, 39–40
Leakage flux, 35
Leaky modes, 669–671
Left-handed polarization, 638
Level encoding, 85
Light emitting diodes (LEDs), 684
Light gain, 686
Light pipe, 654
Linear arrays, 530–531
Linear correlation coefficient, 297
Linear distortion, 96
Linear polarization, 510
Line codes
 alternate mark inversion (AMI), 88
 dc wander, 88
 differential encoding, 92, 93, 388
 high density bipolar *HDBn*, 91–92
 Manchester, 88
 overview, 85–86
 polar NRZ-L, 88
 unipolar NRZ-L, 86
- Lines of position, 496

- Line spectra, 74
 Line waveform, 82
 Liquid crystal displays, 602
 Lissajous method, 226
 Litzendraht wire, 19
 Loading, 415, 556
 Logarithmic scales, 706
 Log periodic antennas, 537
 Long-wire antennas, 529
 Look angles, 628
 Loop antenna, 527
 Loop lengths, 554
 Loss budget, 700
 Lossless lines, 419
 Lower sideband (LSB), 263
 Lower side frequency (LSF), 230, 263
 Low frequency (flicker) noise, 116
 Low-frequency propagation, 495–498
 Low-frequency transformers, 34–36
 Low-pass equivalent noise voltages, 131
 Low-pass filters, 36

 Magnetic storm, 491
 Makeup codes, 592
 Marconi antennas, 524
M-ary encoding, 95
 Master, 498
 Matched filter, 368–371
 Matched loads, 420
 Material dispersion, 677–679
 Matthead, 136, 365, 446–450
 Maximally flat time delay response, 39
 Maximum likelihood decision, 403
 Maximum usable frequency (MUF), 484–486
 Mechanical filter, 212
 Meridian plane, 510
 Microbending, 671
 Microcap, 136
 Microstrip transmission lines, 443–446
 Microwave antennas, 538–545
 Microwave systems, 471–473
 Mid-rise quantization, 341
 Mid-tread quantization, 341
 Miller input admittance, 143
 Minimum shift keying (MSK), 381
 Mixers
 BJT, 160
 definition of, 158
 diode, 159
 FET, 161
 IC balanced, 161–163
 Mixer stage, 125
 Mode stripping, 671

 Modified Huffman code (MHC), 592
 Modulation, 82
 Modulation depth, 233
 Modulation index, 225–228
 Mogel-Dellinger fadeouts, 491
 Momentum wheel stabilization, 624
 Multifrequency tone dialing (MFTD), 552–554
 Multiple parity, 397
 Mutual inductance, 20–22

 Napier's rules, 629
 Narrow band-pass noise, 130–131
 Negative impedance amplifiers, 558
 Neper, 412, 706
 Network synchronization, 362
 Neutralization, 150–151
 Noise
 amplifier input noise in terms of F, 124
 in AM systems, 252–257
 avalanche, 117
 bipolar transistor, 117
 burst, 117
 equivalent input noise generators and comparison
 of BJTs and FETs, 118–120
 field-effect transistor, 118
 figure, 123
 in FM systems, 320–324
 granular, 353
 idle, 280
 low frequency (flicker), 116–117
 measurement, 117
 narrow band-pass noise, 130–131
 noise factor, 122–126
 noise factor and equivalent input noise
 generators, 126
 noise factor of a lossy network, 127
 noise factor of amplifiers in cascade, 124–126
 overview, 105
 partition, 116
 quantization, 341
 S/N ratio of a tandem connection, 120–122
 shot, 116
 signal-to-noise (S/N) ratio, 120
 temperature, 128–129
 thermal, 105–115
 white, 110
 Nonradiating, 463
 Nonresonance, 529–530
 Nonresonant antennas, 505
 Non-return-to-zero (NRZ) pulse, 84
 Normalized frequency (cutoff parameter), 665
 Norton amplifier, 188
 Notch antennas, 545

- NTSC (National Television System Committee of the United States), 593
Numerical aperture (NA), 658
Nyquist frequency, 339
Nyquist pulses, 97
Nyquist sampling, 69, 355
- Omega system, 498
On-off keying (OOK), 376
Open-circuit loads, 420
Optimum working frequency (OWF), 486
Orbits, 622–623
Oscillators
 amplification with positive feedback, 173–175
 Clapp, 183–184
 Colpitt, 179–183
 crystal, 185–186
 frequency synthesizers, 193–195
 Hartley, 184–185
 LC , 178–179
 overview, 173
 Pierce, 185
 RC phase shift, 175–178
 stability, 191
 voltage-controlled, 186–191
Overcoupled, 24
- Padders, 202
PAL (Phase Alternate Line), 593
Paraboloidal reflector, 538–542
Parallel-tuned circuit, 15–17
Parasitic directors, 535
Parasitic reflectors, 534
Parity encoding, 394–397
Partition, 116
Passive circuits
 attenuator pads, 1–9
 capacitive tap, 29–32
 high-frequency transformers, 22–27
 low-frequency transformers, 34–36
 maximum power transfer and impedance matching, 32–33
 mutual inductance, 20–22
 parallel-tuned circuits, 15–17
 passive filters, 36–46
 self-capacitance of a coil, 17–19
 series-tuned circuit, 9–15
 skin effect, 19–20
 tapped inductor, 27–29
Passive electric network, 1
Passive filters, 36–46
Peak swing, 365
Pels, 576, 577
Perigee, 622
Periodic waveforms
 overview, 53
 trigonometric Fourier series for, 53–55
Phase ambiguity, 388
Phase deviation constant, 293
Phase locked loop (PLL), 193–194
Phase modulation (PM)
 digital, 297
 equivalence between PM and FM, 294–296
 overview, 293–294
 sinusoidal, 296
Phase reversal keying (PRK), 383
Phase-shift coefficient, 412–413
Phase shift keying (PSK), 383
Phase velocity, 409, 415–417, 484
Phasing, 583
Phosphor, 598
Photographic recording, 586–587
Pi-attenuator, 5–6
Pierce oscillator, 185
Piezoelectric crystal filters, 40–44
Pilot carrier SSB, 273
pin photodiode, 691
pipn structure, 692
Pixels, 577
Plane reflector arrays, 536
Plane triangle, 629
Plasma frequency, 482–484
pn photodiode, 691
Polarization, 510–512
Polarization discrimination, 637
Polarization fading, 492
Polarization loss (PL), 643
Polar mount, 632
Polar orbiting satellite, 622
Popcorn noise. *See* Burst noise
Power gain, 153–154, 513–515
Power signals, 74–76
Power spectral density function, 74–75
Predictor block, 352
Pre-emphasis network, 325
Prescaling, 194–195
Primary line constants, 408–409
Private automatic branch exchange (PABX), 573
Private branch exchange (PBX), 573
Probability of bit error, 364
Product detectors, 263
Propagation coefficient, 412–415
Pseudoternary code, 90
PSpice, 136
Pulse amplitude modulation (PAM), 336–340

Pulse code modulation (PCM)
 compression, 345–348
 delta modulation, 353–354
 differential, 352
 quantization, 341–344
 receiver, 348–352
 sigma-delta A/D conversion, 354–356

Pulse frequency modulation (PFM), 356–357

Pulse position modulation (PPM), 357–358

Pulse(s)
 functional notation for, 84–85
 non-return-to-Zero (NRZ) pulse, 84
 return-to-zero (RZ) pulse, 84

Pulse shaping, 96–101

Pulse time modulation (PTM), 357

Pulse train, 59–62

Pulse width modulation (PWM), 358–359

Pyramidal horn, 538

Q-factor, 10–12, 16

Q-function, 366

Quadrantal spherical angle, 629

Quadrature component, 131

Quadrature phase shift keying (QPSK), 384

Quantization, 341–344

Quarter-wave transformer, 427

Quasi-steady-state analysis, 297

Quaternary encoding, 83

Radiation fields, 510

Radiation resistance, 507

Radio-frequency transmission lines, 443

Radio horizon, 476–478

Radio-wave propagation
 extremely low frequency, 498–503
 in free space, 468–473
 ionospheric, 482–493
 overview of, 468
 surface waves, 493–495
 tropospheric, 473–482

Raised cosine response, 97

Random waveform, 74

Raster scanning, 576, 599

Ratio detector, 314–318

Rayleigh scattering losses, 668

RC phase shift oscillator, 175–178

RC low-pass filters, 164–165

Receivers
 adjacent channel selectivity, 210–212
 AM, 247–252
 and automatic gain control, 212–214
 discrete component AM, 247–248
 double-conversion, 214–216

electrically tuned, 216
 facsimile, 581–588
 image rejection and, 206–208
 integrated-circuit, 218–221
 sensitivity and gain, 205–206
 spurious responses and, 208–210
 superheterodyne, 198–200
 television, 608–612
 tracking, 201–205
 tuning range of, 200–201

Reciprocity theorem, 507

Reconstruction filter, 350

Rectangular waves, 56–59

Redundant receiver, 639

Reed-Solomon (RS) block code, 399

Reflectionless match, 33

Refraction, 655

Repetition encoding, 394

Resonant antennas, 505

Return loss, 559

Return-to-zero (RZ) pulse, 84

Rhombic antennas, 530

Right-handed polarization, 511

Roll, pitch, yaw (RPY) axes, 624

Roll-off factor, 98

Rotation of the line of apsides, 622–623

Round-Travis detector, 310

Run length coding, 592

Sample-and-hold technique, 340

Sampled data system, 167

Sampling function, 61

Sampling theorem, 338

Satellite communications
 antenna look angles, 627–635
 attitude control, 624–626
 digital carrier transmission, 648–649
 downlink power budget calculations, 646–647
 frequency plans and polarization, 637–638
 geostationary orbit, 623
 Kepler's first law, 620–621
 Kepler's second law, 621
 Kepler's third law, 622
 limits of visibility, 635–637
 multiple access methods, 649–650
 orbits, 622–623
 overall link budget calculations, 647–648
 overview, 620
 power systems, 624
 satellite station keeping, 626–627
 transponder, 638–642
 uplink power budget calculations, 642–646

Saturation flux density, 644

- Saturation point, 640
Sawtooth waveform, 59
Scanning, 577–581
Scanning receiver, 201
Scan spot, 581
Scatter propagation, 491
SECAM (Sequential Coding and Memory) systems, 593–594
Secant law, 494–496
Second harmonic, 53
Sectoral horns, 538
Selective fading, 492
Selectivity, 210–212
Self-capacitance, 17–19
Self-clocking synchronization, 362, 372
Semiconductor laser diodes, 685
Sensitivity, 205–206
Series-tuned circuits, 9–16
Short-circuit loads, 420
Shot, 116
Sidebands, 232–233
Sideband splatter, 235, 377
Sidetone, 549
Sigma-delta A/D conversion, 354–356
Signal-to-noise (S/N) ratio, 120, 278–280
Sine function, 61
Single-mode fiber, 667
Single-sideband (SSB) modulation
 balanced modulators, 264–266
 companding, 280
 description, 262
 generation, 267–271
 modified systems, 273–278
 principles of, 262–264
 reception, 271–272
 signal-to-noise ratio for, 278–280
Sinusoidal FM, 285–287
 average power in, 291–292
 frequency spectrum for, 228–231
 measurement of modulation index for, 293
Sinusoidal phase modulation (PM), 296
Sinusoidal waveforms, 51–53
Skin effect, 19–20
Slope overload, 353
Slope polarity switch, 354
Slot, 544–545
Slotted-line measurements, 421–424
Smith chart, 428–438
Soft-decision decoding, 390
Solid-state high-power amplifiers (SSHPAs), 640
Source coding, 82
Spillover, 538
Spin stabilization, 624
Spread spectrum, 650
Spread spectrum multiple access, 650
Spurious, responses, 208–210
Squaring loop, 386
SQUID, 503
Stability, 150–151
 amplitude, 192–193
 frequency, 191–192
 linearity, 193
 of oscillators, 191
 self-starting, 192
Stagger tuning, 211
Standard AM, 223
Standing waves, 417–419, 459–460
Star quad, 442
Step-by-step switching, 564–565
Step index profile, 659
Stripe laser, 686
Strowger stepping switch, 564
Subrefraction, 479
Subscriber line interface circuit (SLIC), 549
Subsidiary communications authorization (SCA), 309
Sudden ionospheric disturbances (SIDs), 496
Superheterodyne receivers, 198–200
Superrefraction, 478
Surface acoustic wave filters, 44–46
Surface acoustic waves (SAWs), 44–46
Surface waves, 493–495
Swing, 371
Switched capacitor filters, 165–169
Syllabic filter, 354
Symbol period, 83
Symbols, 82
Symbol synchronization, 362
Synchronization, 362, 582–583
Synchronous carrier signal, 263
Synchronous detection, 379
Synchronously tuned circuits, 24
Synchronous transmission, 362
T1 system, 350
Tapped inductor, 27–29
T-attenuator, 2–5
Tchebycheff response. *See* Chebyshev response
Teledeltos paper, 587
Telephone lines, 440–442
Telephone systems
 public network, 561–573
 wire, 548–561
Television
 cameras, 595–598
 displays, 598–602
 high-definition (HDTV), 594, 614–618

- Television (*Contd.*)
 overview, 593–594
 receivers, 608–611
 signals 606–608
 transmitters, 612–614
- Terminal set (TS), 549
- Terminating codes, 592
- Thermal, 105–116
- Third harmonic, 53
- Third-order intermodulation products, 161
 3dB bandwidth, 13–15, 24–25
- Threshold level, 324
- Threshold margin, 324
- Time division multiple access (TDMA), 649
- Time division multiplexed (TMD) signal, 338
- Time-domain reflectometry (TDR), 438–440
- Time mode switching, 569
- Timing jitter, 375
- Total internal reflection, 655
- Touch Tone signaling, 550–551
- Tracking, 201
- Traffic, 567
- Transfer characteristic, 297
- Transfer xerography, 587
- Transistor modulators, 301
- Transmission bridges, 556
- Transmission lines
 characteristic impedance, 410–412
 as circuit elements, 424–428
 lossless Lines at radio frequencies, 419–420
 Mathcad in calculations, 446–450
 microstrip, 443–446
 overview, 407–408
 phase and group velocities, 415–417
 phase velocity and line wavelength, 409–410
 primary line constants, 408–409
 propagation coefficient, 412–415
 radio-frequency Lines, 443
 slotted-line measurements at radio frequencies, 421–424
 Smith chart and, 428–438
 standing waves, 417–419
 telephone lines and cables, 440–442
 time-domain reflectometry, 438–440
 voltage standing-wave ratio, 420–421
- Transmission path loss (TPL), 469
- Transponders, 638–642
- Transverse electric (TE) mode, 454
- Transverse-electromagnetic (TEM) mode, 407, 453, 468, 510, 661, 709–710
- Transverse magnetic (TM) mode, 464–466, 662
- Trapezoidal method, 225
- Traveling ionospheric disturbances (TIDs), 491
- Traveling wave tubes (TWTs), 640
- Tree diagram, 400
- Trimmers, 202
- Triple repetition, 394
- Troposcopic propagation, 473–482
- Tropospheric scatter propagation, 479
- Trunk circuits, 572–573
- Tuned primary untuned Secondary circuits, 26–27
- Tuning, 10
- Tuning range, 200–201
- Turnstile antennas, 533–534
- Twin cable, 442
- Two-wire repeaters, 558
- Two-wire transmission, 558–560
- Ultraviolet absorption, 668
- Unconditionally stable, 150
- Undercoupled, 24
- Unipolar nonreturn to zero (UPNNRZ), 680
- Unipole antenna, 525
- Uniqueness, 68
- Unity gain transition frequency, 140–141
- Untuned primary tuned secondary circuit, 26–27
- Upper sideband (USB), 263
- Upper-side frequency (USF), 230, 263
- Valid station stop, 217
- Varactor diode modulators, 297–301
- Variable reactor, 187
- Vertical antennas, 523–526
- Vertical polarization, 510
- Very low-frequency propagation, 495–498
- VHF/UHF
 antennas, 536–537
 radio systems 479–482
- Video bandwidth, 606
- Virtual channels, 364
- Virtual height, 486–489
- Voltage-controlled oscillator (VCO), 173–191
- Voltage magnification factor, 11
- Voltage standing wave (VSW), 417
- Voltage standing wave ratio (VSWR), 420–421
- Voltage transfer function, 25–26
- Waveforms
 bandwidth requirements for analog information signals, 76–77
 digital line, 82–101
 energy signals and Fourier transforms, 66–68
 exponential Fourier series, 62–64
 fast Fourier transform, 68–71
 formulas for Fourier coefficients, 64–66

- general periodic, 53
- general properties of periodic, 62
- intersymbol interference, 96
- inverse fast Fourier transform, 71–72
- line codes and, 85–94
- M*-ary encoding, 95–96
- notation for pulses, 84–85
- power signals, 74–76
- pulse shaping, 96–101
- pulse train, 59–62
- rectangular, 56–59
- sawtooth, 59
- signal filtering, 73–74
- sinusoidal, 51–53
- spectrum for the trigonometric
 - Fourier series, 55–56
- symbols, bunits, bits, bauds, 92–93
- trigonometric Fourier series for periodic, 53–55
- Waveform synthesis, 58
- Waveguide dispersion, 679–681
- Waveguides
 - other modes, 464–467
 - overview, 453
 - rectangular, 453–454
- Waveguide terminations, 460–462
- Wave impedance, 459
- Wavelength division multiplexed (WDM), 673
- Wavelengths, toward generator, 431
- Wavelengths toward load, 428, 431
- Wavetrap, 14, 17
- Weaver, D.K., 269
- White noise, 110
- Yagi–Uda array, 535–536
- Zoned lenses, 544