UNITED STATES NAVAL

TEST PILOT SCHOOL

AIRBORNE SYSTEMS COURSE

TEXTBOOK

COMMUNICATIONS SYSTEM TEST AND EVALUATION

By

George W. Masters

1 January 1981

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER -	2. GOVT ACCESSION NO. AD-A[illegible]	3. RECIPIENT'S CATALOG NUMBER -
4. TITLE *(and Subtitle)* Communications System Test and Evaluation		5. TYPE OF REPORT & PERIOD COVERED - 6. PERFORMING ORG. REPORT NUMBER -
7. AUTHOR(s) George W. Masters		8. CONTRACT OR GRANT NUMBER(s) -
9. PERFORMING ORGANIZATION NAME AND ADDRESS U. S. Naval Test Pilot School Naval Air Test Center Patuxent River, Maryland 20670		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS -
11. CONTROLLING OFFICE NAME AND ADDRESS Academics Branch U. S. Naval Test Pilot School		12. REPORT DATE 1 January 1981 13. NUMBER OF PAGES 320
14 MONITORING AGENCY NAME & ADDRESS*(If different from Controlling Office)* -		15. SECURITY CLASS. *(of this report)* UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE -

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

-

18. SUPPLEMENTARY NOTES

-

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Communications, Avionics, Digital Systems, Data Transmission, Electronics, Signal Processing, Antennas, Electromagnetic Wave, Propagation, Airborne Systems, Test and Evaluation, Flight Testing

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

Textbook for use in teaching the test and evaluation of communications systems including theory of operation, operating characteristics, and test methodology. Topics include communications theory, electromagnetic wave propagation, communications system hardware and test equipment, and test methods.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

COMMUNICATIONS SYSTEMS
TEST AND EVALUATION

TABLE OF CONTENTS

COMMUNICATION SYSTEM
TEST AND EVALUATION

1.0 Introduction

For the purposes of this text, we will assume a broad definition for the term "communication system". Specifically, we will define a communication link as any arrangement that conveys information from one point to another. An obvious example is the two-way voice communication system commonly used in aircraft. A somewhat less obvious example is the digital data link used in tactical data systems. A still less obvious example is an airborne radar. All of these systems meet our broad definition. We will, however, be concerned primarily with systems that utilize electromagnetic waves as the information carrier, although many of the same principles apply to communication systems using other methods of telemetering information.

The principal elements of a typical electronic communications system are: (1) a carrier generator, which produces energy in a form suitable for electromagnetic propagation; (2) a modulator, which impresses the pertinent information onto the carrier; (3) a power amplifier, which boosts the amplitude of the modulated signal to a suitable level; (4) a transmitting antenna, which acts as an impedance-matching device between the transmitting apparatus and the propagation medium and radiates the electromagnetic energy in the prescribed direction(s); (5) a propagation medium, in which the information carrier travels; (6) a receiving antenna, which acts as an impedance-matching device between the propagation medium

and the receiving apparatus and selectively absorbs electromagnetic energy propagating in the prescribed direction(s); (7) a tuner, which selects the desired signal and rejects unwanted signals and noise; (8) a demodulator, which extracts the pertinent information from the carrier; and (9) an indicator which puts the pertinent information in a form suitable for the user.

The principles of operation and the operating characteristics of the communications elements listed above are the subject of this text. The necessary information will be provided to allow the reader to design experiments (tests) which will reveal the significant characteristics of the communication system being evaluated. Sufficient theoretic background will be provided to allow him to analyze the design of the system and thereby anticipate its likely performance characteristics (strengths and weaknesses). Only in this manner can an adequate test plan be devised. The difficulties in designing an adequate test are due, in large part, to the ever-present funding and time limitations. The tester virtually never has enough time or funding to perform exhaustive testing. For that reason, he must concentrate on those tests which reveal the likely performance weaknesses of the subject system. Only by understanding the system design and its implications can he identify the appropriate areas of concentration.

The first section will present the necessary background theory. The second section will discuss the pertinent characteristics of various

types of communication systems. The third section will present the communication system performance characteristics to be experimentally determined and the methods for their measurement. A separate text will present the non-performance (environmental, electromagnetic compatibility, hardware viability, maintainability, reliability, and crew safety) characteristics to be determined, the general methods for their determination, and the applicable specifications and references.

2.0 Communications Theory

2.1 Basic Electronics

2.1.1 Electrical Voltages and Currents -- All matter can be considered to consist of small particles. Some of these particles are said to "possess positive electrical charge", some to "possess negative charge", and some to be "electrically neutral". In solid matter, only the negatively charged particles (electrons) are relatively free to move. For that reason, almost all electrical phenomena pertinent to communications systems can be explained by considering the state of the electrons in the vicinity.

An electrically charged particle creates an electric field in the surrounding space. That is, it exerts a force on any other electrically charged particle in its vicinity. Voltage is a measure of that force. An electrically charged particle in motion constitutes an electric current. Such a moving charged particle creates a magnetic field in the surrounding space. That is, it exerts a force, in addition to that due to its electric field, on any other moving charged particle in its vicinity. The field of electronics is concerned with the motions of electrons and the fields they create.

An electrical circuit is defined as a closed path in which electrical current flows. The current flowing in a circuit depends upon two factors:

the power sources forcing the current to flow and the circuit elements impeding current flow (impedances).

2.1.2 Passive Linear Circuit Elements -- In an electrical circuit, there is a definitive relationship between the voltage drop across an element (difference between the voltages at the two terminals) and the current that is passing, or has passed, through that element. The relationships between voltage and current define the three basic types of passive, linear circuit element: the resistor, the inductor, and the capacitor.

For a purely resistive element, shown in the diagram of Figure 2.1.2.1 (a), the voltage/current relationship, known as Ohms law, is given by:

$$e(t) = R\, i(t) \qquad [2.1.2.1]$$

where e(t) is the voltage across the resistor, in volts; i(t) is the current through the resistor, in amperes; and R is the resistance, in ohms. Note that the voltage across a resistor is proportional to the current through it.

For a purely inductive element, shown in the diagram of Figure 2.1.2.1 (b), the voltage/current relationship is given by:

$$e(t) = L \frac{d}{dt}\left[i(t)\right] \qquad [2.1.2.2]$$

where e(t) is the voltage, i(t) is the current, and L is the inductance

in henries. Thus, the voltage across an inductor is proportional to the time rate of change of current through it.

For a purely capacitative element, shown in the diagram of Figure 2.1.2.1 (c), the voltage/current relationship is given by:

$$e(t) = \frac{1}{C} \int_0^t i(t)\, dt \qquad [2.1.2.3]$$

where e(t) is the voltage, i(t) is the current, and C is the capacitance in farads. Thus the voltage across a capacitor is proportional to the time integral of the current passing through it. Note that:

$$\int_0^t i(t)\, dt = q(t) \qquad [2.1.2.4]$$

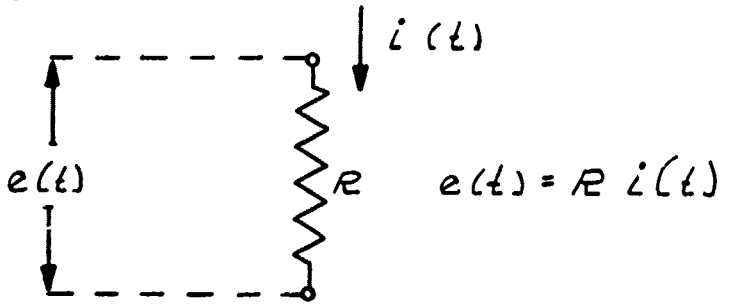
where q(t) is the charge stored in the capacitor, in coulombs.

In addition to the three basic two-terminal circuit elements, there is one other passive, linear circuit element of interest: the transformer. The transformer is a four-terminal device as shown in the diagram of Figure 2.1.2.1 (d). For an ideal transformer, the primary and secondary voltages and currents are related by the equations:

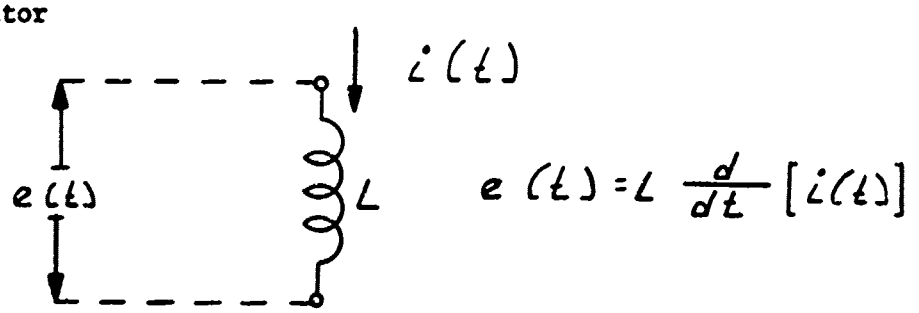
$$\frac{e_2(t)}{e_1(t)} = \frac{N_2}{N_1} \quad \text{and} \quad \frac{i_2(t)}{i_1(t)} = \frac{N_1}{N_2} \qquad [2.1.2.5]$$

where N_1 and N_2 are the numbers of turns in the transformer primary and secondary windings, respectively. In some applications, the transformer is used as an impedance changing device. That is, if an impedance of Z_L

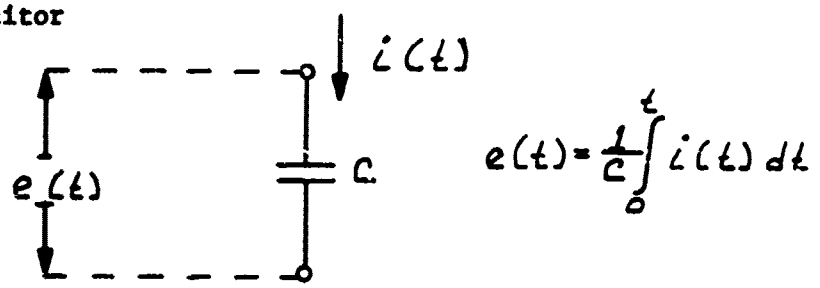
(a) Resistor



(b) Inductor



(c) Capacitor



(d) Transformer

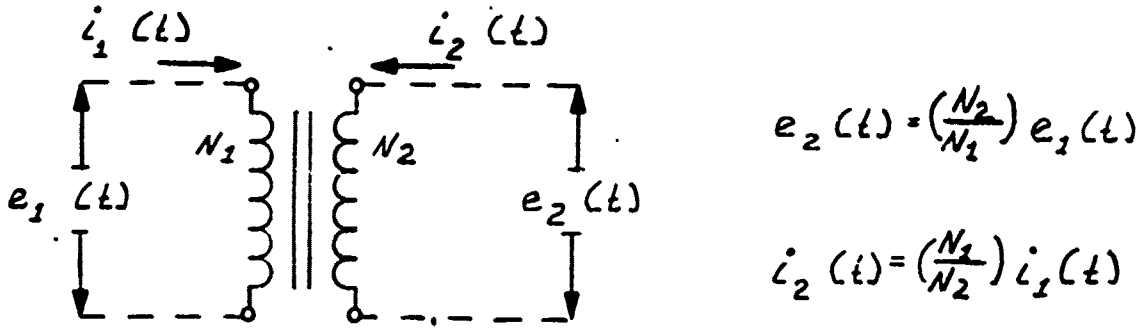


Figure 2.1.2.1 Passive Linear Circuit Elements

is connected to the secondary winding of the ideal transformer in Figure 2.1.2.1 (d), the impedance appearing at the primary terminals is given by:

$$Z_{Pri} = \left(\frac{N_1}{N_2}\right)^2 Z_L \quad (ohms) \qquad [2.1.2.6]$$

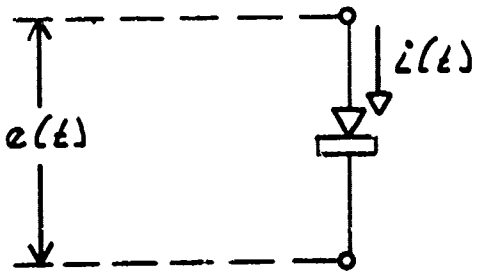
2.1.3 Passive, Nonlinear Circuit Elements -- Although there are numerous nonlinear circuit elements, only one is basic to the understanding of communication system characteristics: the diode or rectifier, shown in Figure 2.1.3 (a). The ideal diode passes current only in one direction. It can be considered a resistive element whose resistance, R_d, is infinite for reversed voltage and zero for forward voltage. The voltage/current relationship of an ideal diode is shown in the figure.

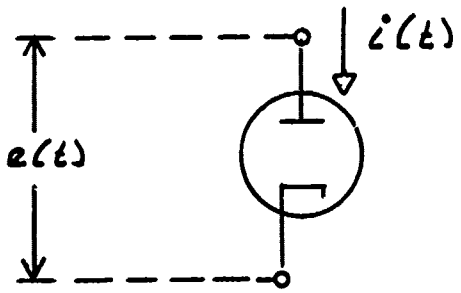
2.1.4 Active Circuit Elements -- An active circuit element is one that adds power (energy) to that of the voltages and currents in a circuit. Of the many active devices, two are of basic interest: the vacuum tube and the transistor. For proper operation, both the vacuum tube and the transistor require that correct biasing currents and voltages be applied to their terminals. The following, simplified discussion assumes proper operating conditions and considers only the signal voltages and currents.

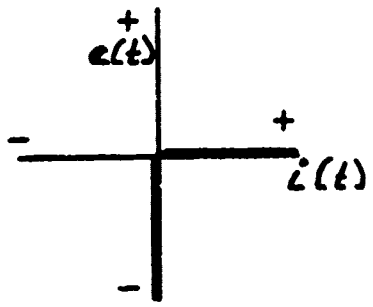
The vacuum tube, shown in Figure 2.1.4.1(a), generally can be considered a voltage amplifying device. That is, for the proper operating conditions:

$$e_p(t) = \mu\, e_g(t) \qquad [2.1.4.1]$$

(a) Diode or Rectifier







$R_D = \infty$ for $e \equiv 0$

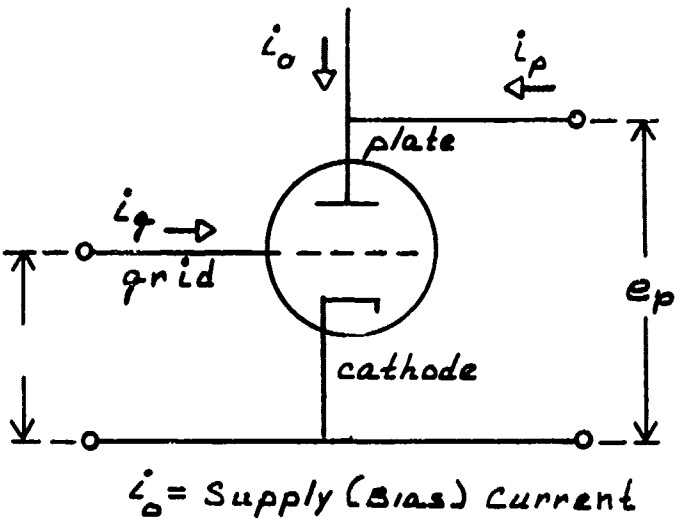
$R_D = 0$ for $e > 0$

$i(t) = 0$ for $e(t) \equiv 0$

$e(t) = 0$ for $i(t) > 0$

Figure 2.1.3 — Passive, Nonlinear Circuit Elements

(a) Vacuum Tube (Triode)



(b) Transistor (NPN)

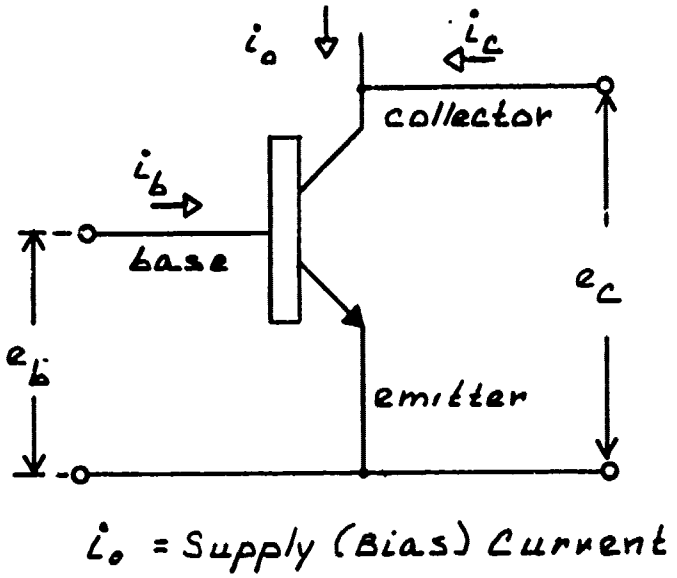


Figure 2.1.4.1-- Active Circuit Elements

where $e_p(t)$ is the output (plate) signal voltage, $e_g(t)$ is the input (grid) signal voltage, and μ is a constant multiplying factor. For most purposes, it can be assumed that the grid current, i_g, is zero. The output current, i_p, depends upon the total impedance of the output circuit.

The transistor, shown in Figure 2.1.4.1 (b), generally can be considered a current amplifying device. That is, for the proper operating conditions:

$$i_c(t) = \beta\, i_b(t) \qquad [2.1.4.2]$$

where $i_c(t)$ is the output (collector) signal current, $i_b(t)$ is the input (base) signal current, and β is a constant multiplying factor. For most purposes, the input (base-to-emitter) junction has the characteristics of a forward-biased diode. The output voltage, e_c, depends upon the output circuit (load) impedance.

2.1.5 Voltage and Current Sources -- In practice, there do not exist ideal voltage or current sources. Often, however, actual circuits have characteristics which approximate those of ideal sources. For that reason, such sources are of practical interest.

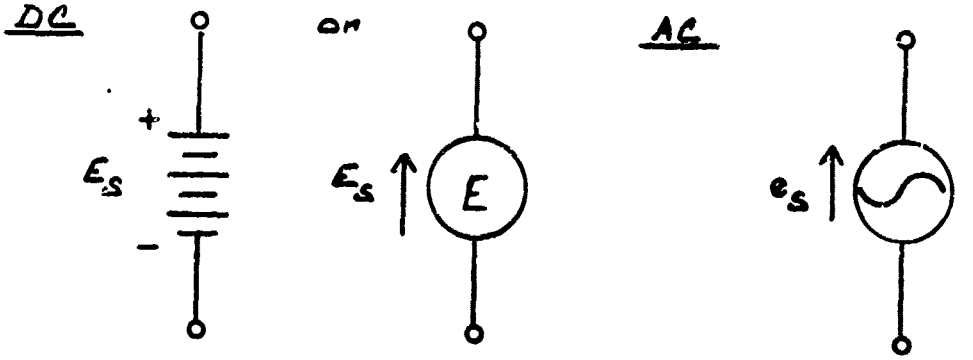
An ideal voltage source is a device that maintains a given voltage across its terminals, regardless of the current drawn from those terminals. A

storage battery generally can be considered a voltage source. For practical purposes, a voltage source can be considered ideal when the external (load) impedance is much larger than the source internal impedance. Note that a voltage source is a generalized short circuit. That is, a short circuit holds the voltage across its terminals at a constant value equal to zero. The symbols commonly used for AC and DC voltage sources are shown in Figure 2.1.5.1 (a).

An ideal current source is a device that maintains a given current flow from its terminals, regardless of the load impedance (voltage) across those terminals. A storage battery with a very large resistance in series generally can be considered a current source. For practical purposes, a current source can be considered ideal when the external (load) impedance is much smaller than the source internal impedance. Note that a current source is a generalized open circuit. That is, an open circuit holds the current from its terminals at a constant value equal to zero. The symbols commonly used for AC and DC current sources as shown in Figure 2.1.5.1 (b).

2.1.6 Direct Current (DC) Circuits -- A DC circuit can be defined as one in which the voltages and currents are constant with time. Under such circumstances, the only passive circuit element of significance is the resistor and the only applicable voltage/current relationship is Ohm's law. A simple, one-loop DC circuit is shown in Figure 2.1.6.1. Also shown in that figure are the various forms of Ohm's law. The impedance of a circuit element is defined as the ratio of the voltage across that

(a) Voltage Source



(b) Current Source

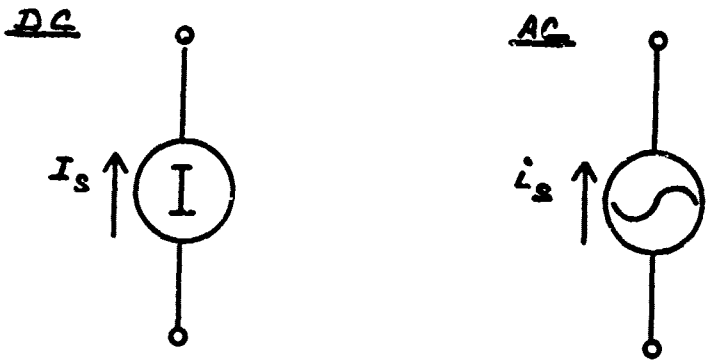
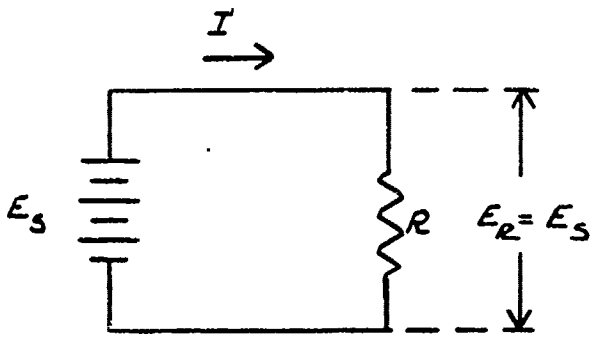


Figure 2.1.5.1 — Voltage and Current Sources



Ohm's Law

$E_S = I_S R$ (volts)

$I_S = \frac{E_S}{R}$ (Amperes)

$R = \frac{E_S}{I_S}$ (Ohms)

Figure 2.1.6.1 -- DC Circuit

element to the current flowing through the element. Thus, for the DC circuit shown in Figure 2.1.6.1, the impedance is:

$$Z = \frac{E_R}{I} = R \quad (\text{ohms}) \qquad [2,1,6.1]$$

2.1.7 Alternating Current (AC) Circuits -- An AC circuit can be defined as one in which the voltages and currents vary with time. When such is the case, all of the circuit elements discussed in Sections 2.1.2 are of significance and the analysis of the circuit becomes much more complex than that for a DC circuit. The principal reason for this complexity is that, whereas a non-time-varying (DC) quantity requires only a magnitude parameter to describe it, a time-varying (AC) quantity requires two parameters: a magnitude parameter (amplitude) and a timing parameter (phase). In order to expedite the handling of the dual nature of AC quantities, complex notation (use of a 2-vector with "real" and "imaginary" parts) is employed. Complex notation is discussed in a later section of this text.

Fortunately, the analysis of an AC circuit is somewhat simplified when the AC currents and voltages are sinusoidal in nature. This is the case, of course, when only sinusoidal voltages and currents have been applied to the circuit. Even when the time-varying voltages and currents are not sinusoids, however, a sinusoidal analysis still applies if the actual waveforms are considered to be composed of superimposed sinusoidal components. This approach is possible when the circuit elements are linear and,

therefore, the principle of linear superposition applies. The circuit analysis procedure then requires that the applied voltages and currents be broken down into their Fourier components, the resulting individual sinusoidal responses determined, and the actual response obtained by superimposing (adding) the individual responses. (An alternative procedure involves utilizing Fourier or Laplace Transforms).

2.1.8 Sinusoidal Waveforms -- A typical sinusoidal waveform is shown in Figure 2.1.8.1. As indicated in that figure, the waveform shown can be represented as a sine function or as a cosine function with 90° phase shift.

If two sinusoids of the same frequency are added, it is always possible to represent the resulting waveform as a single sinusoid with an appropriate amplitude and phase angle. Conversely, it is always possible to represent any sinusoid of arbitrary amplitude and phase as a sum of two sinusoids of appropriate amplitude and zero phase. For example, if:

$$e(t) = E_1 \cos(\omega t + \phi_1) + E_2 \sin(\omega t + \phi_2) \qquad [2.1.8.1]$$

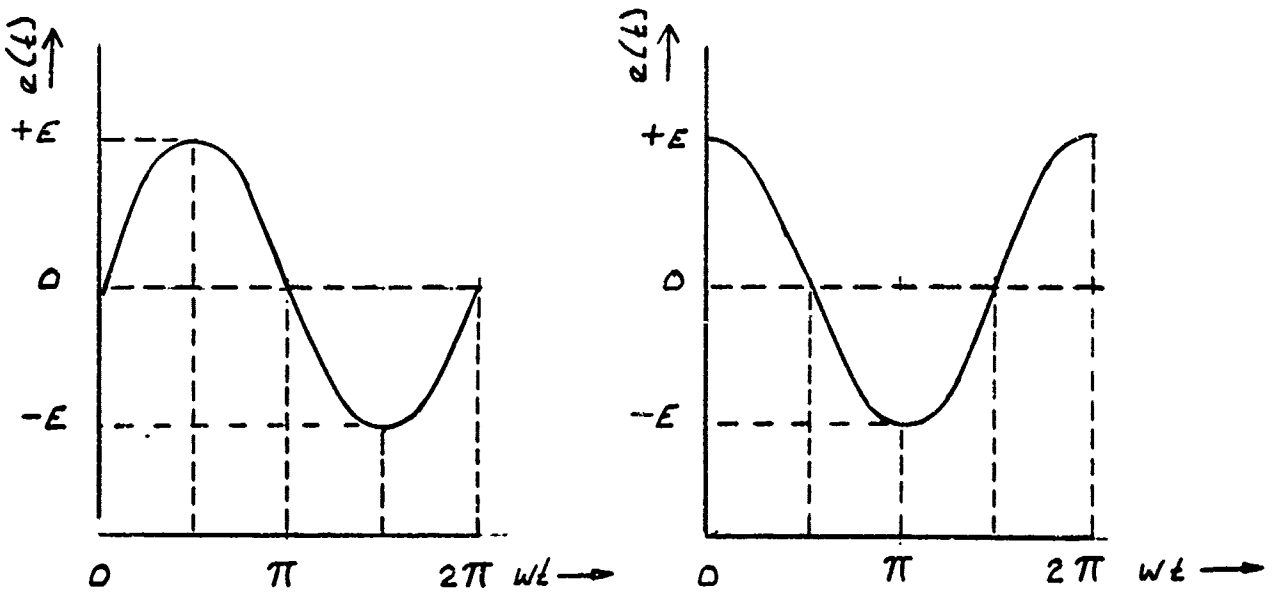
Then:

$$e(t) = E \cos(\omega t + \phi) \qquad [2.1.8.2]$$

where:

$$E = \left[E_1^2 + E_2^2 + 2 E_1 E_2 \sin(\phi_2 - \phi_1)\right]^{1/2}$$

$$\phi = \tan^{-1}\left[\frac{E_1 \sin\phi_1 - E_2 \cos\phi_2}{E_1 \cos\phi_1 + E_2 \sin\phi_2}\right] \qquad [2.1.8.3]$$



$e(t) = E \sin(wt)$

$e(t) = E \cos(wt - \frac{\pi}{2})$

$e(t) = E \cos(wt)$

$e(t) = E \sin(wt + \frac{\pi}{2})$

Figure 2.1.8.1 -- Sinusoidal Waveforms

or, given:

$$e(t) = E\ Cos(\omega t + \phi) \qquad [2.1.8.4]$$

then:

$$e(t) = E_1\ Cos(\omega t) + E_2\ Sin(\omega t) \qquad [2.1.8.5]$$

where:

$$E_1 = E_3\ Cos\phi$$
$$E_2 = -E_3\ Sin\phi \qquad [2.1.8.6]$$

Thus, utilizing Fourier analysis and synthesis, it is always possible to represent an arbitrary time function with arbitrary phase angle as the sum of simple sinusoidal terms of zero relative phase.

2.1.9 Root-Mean-Square Voltages and Currents -- The root mean square value of a time-varying function, f(t), is defined by the equation:

$$f_{rms} = \left[\overline{f^2(t)}\right]^{1/2} \qquad [2.1.9.1]$$

where the time average $\overline{f^2(t)}$ is defined as:

$$\overline{f^2(t)} = \frac{1}{T}\int_0^T f^2(t)\,dt \qquad [2.1.9.2]$$

and T = $1/\omega$ is the period of the function. Thus, to derive the RMS value of a periodic time function:

(1) Square the time function.

(2) Average the square over one period.

(3) Take the square root of the average.

Note that the RMS value of a time function is a constant.

The RMS value of a sinusoid of amplitude E is derived in the following manner.

$$f(t) = E\,Sin(\omega t) \qquad [2.1.9.3]$$

$$f^2(t) = E^2 Sin^2(\omega t) = E^2\left[1/2 - 1/2\,Cos(2\omega t)\right] \qquad [2.1.9.4]$$

$$\overline{f^2(t)} = \frac{1}{T}\int_0^T E^2\left[\frac{1}{2} - \frac{1}{2}\,Cos(2\omega t)\right]dt = \frac{1}{2}E^2 \qquad [2.1.9.5]$$

$$f_{rms} = \left[\overline{f^2(t)}\right]^{1/2} = \frac{E}{\sqrt{2}} = 0.707\,E \qquad [2.1.9.6]$$

The RMS value of the sum of two sinusoids in phase quadrature (90° out of phase) is equal to the root-sum-square (RSS) of their individual RMS values. To obtain the RSS of two such values:

(1) Square the individual RMS values

(2) Sum the squares

(3) Take the square root of the sum.

Thus if:

$$e(t) = E_1 \cos(\omega t) + E_2 \sin(\omega t) \qquad [2.1.9.7]$$

then:

$$e_{RSS} = \left[\tfrac{1}{2}\left(E_1^2 + E_2^2\right)\right]^{1/2} \qquad [2.1.9.8]$$

2.1.10 Ohm's Law for AC Circuits -- Ohm's law is given by the equation:

$$e(t) = Z\, i(t) \qquad [2.1.10.1]$$

where e(t) is the voltage across a circuit component, i(t) is the current flowing through that component, and Z is the impedance of that component. Accordingly, the impedance is defined by the equation:

$$Z = \frac{e(t)}{i(t)} \qquad [2,1,10,2]$$

In Figure 2.1.10.1 are shown simple, one-loop circuits with a sinusoidal current flowing through an impedance. The current is a sinusoid given by:

$$i(t) = I \cos(\omega t) \qquad [2,1,10,3]$$

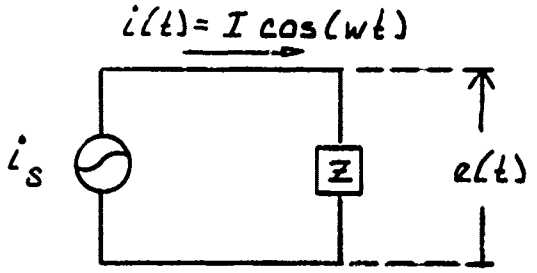
In accordance with the material presented in Section 2.1.2, the voltages across the three types of impedance are those shown in the Figure. For the resistor, the impedance is then:

Ohm's Law

$$e(t) = i(t)\, Z$$

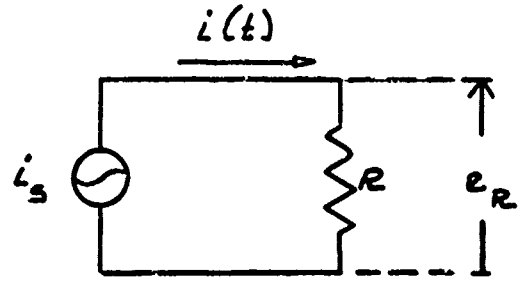
$$i(t) = \frac{e(t)}{Z}$$

$$Z = \frac{e(t)}{i(t)}$$



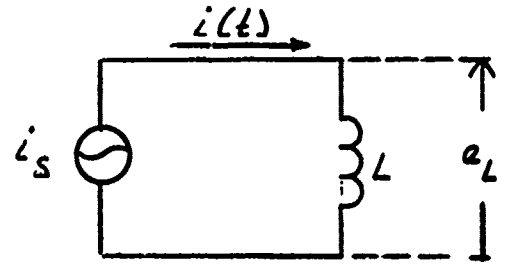
Resistance

$$e_R = R\, i(t) = RI \cos(\omega t)$$



Inductance

$$e_L = L \frac{di}{dt} = -\omega L I \sin(\omega t)$$



Capacitance

$$e_C = \frac{1}{C} \int_0^t i(t) = \frac{I}{\omega C} \sin(\omega t)$$

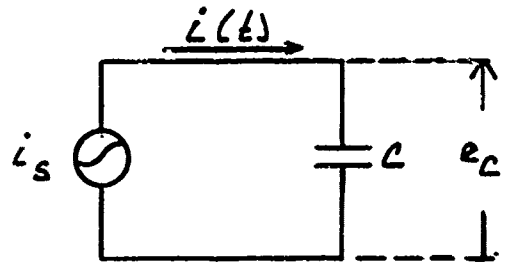


Figure 2.1.10.1 -- AC Impedances

$$Z_R = \frac{RI\,\cos(\omega t)}{I\,\cos(\omega t)} = R \qquad [2.1.10.4]$$

This result agrees with that obtained in Section 2.1.6. For the inductor, the impedance is:

$$Z_L = \frac{-\omega L I\,\sin(\omega t)}{I\,\cos(\omega t)} = -\omega L\,\tan(\omega t) \qquad [2.1.10.5]$$

Obviously, there is a problem since Tan (ωt) is time dependent and goes to infinity periodically. The problem is that, unlike the resistor, the inductor shifted the voltage 90° out of phase with the current as well as affecting the amplitude of the sinusoid. Thus, if Ohm's Law is to be universally applicable, inductive and capacitive impedances must be defined in such a way as to account for both amplitude and phase effects. The method by which that is accomplished is the complex impedance and the resulting mathematical convention is called complex notation.

As stated in Section 2.1.8, an arbitrary time-varying function can be represented, by Fourier analysis, as the sum of a series of sine and cosine terms. Using complex notation, these two components, 90° out of phase, are represented as the real and imaginary parts of $e^{j\omega t}$ via the Euler relationship. That is:

$$\cos(\omega t) = \mathrm{Re}\{e^{j\omega t}\} = \mathrm{Re}\{\cos(\omega t) + j\sin(\omega t)\}$$

$$\sin(\omega t) = \mathrm{Im}\{e^{j\omega t}\} = \mathrm{Im}\{\cos(\omega t) + j\sin(\omega t)\} \qquad [2.1.10.6]$$

Thus, under this convention, multiplying a sinusoid by j, ($\sqrt{-1}$), shifts its phase by 90°. As a result, the three basic types of impedance can now be defined by the following formulas.

For a resistance, $Z_R = R$

For an inductance, $Z_L = j\omega L$ [2.1.10.7]

For a capacitance, $Z_C = -\frac{j}{\omega C}$

Using complex notation, Ohm's law can now be applied to the circuits of Figure 2.1.10.1, taking as the current:

$$i(t) = I\left[\cos(\omega t) + j\sin(\omega t)\right] \qquad [2.1.10.8]$$

with the understanding that only the real part of the result will have meaning. For the resistor:

$$e_R(t) = IR\left[\cos(\omega t) + j\sin(\omega t)\right]$$

$$Z_R = Re\left\{e_R(t)/i(t)\right\} = R \qquad [2.1.10.9]$$

For the inductor:

$$e_L(t) = \omega L I\left[-\sin(\omega t) + j\cos(\omega t)\right]$$

$$Z_L = Re\left\{e_L(t)/i(t)\right\} = j\omega L \qquad [2.1.10.10]$$

For the capacitor:

$$e_c(t) = \frac{I}{\omega c}\left[\sin(\omega t) - j\cos(\omega t)\right]$$

$$Z_c = Re\left\{e_c(t)/i(t)\right\} = -\frac{j}{\omega c} \qquad [2.1.10.11]$$

It can be seen that Equations [2.1.10.9] through [2.1.10.11] are in agreement with Equations [2.1.10.7]. Thus, utilizing complex notation, Ohm's law applies to both AC and DC circuits.

2.1.11 Electrical Power -- The instantaneous electrical power supplied to a circuit component is given by the equation(1):

$$p(t) = e(t)\, i(t) \quad (\text{watts}) \qquad [2.1.11.1]$$

The average power dissipated, (usually as heat), during a time interval T is given by:

$$P = \overline{e(t)\, i(t)} = \frac{1}{T}\int_0^T e(t)\, i(t)\, dt \qquad [2.1.11.2]$$

Using Equation [2.1.11.2], the average power dissipated by the three basic types of impedance can be determined. Substituting into Equation [2.1.11.2] the expressions derived for e(t) and i(t) derived in Figure 2.1.10.1, we have, for the instantaneous and average powers:

(1) It is an artifact of the complex notation that instantaneous real power in complex notation is given by:

$$p(t) = Re\left\{e(t)\, i^*(t)\right\}$$

where i*(t) is the complex conjugate of i(t). The complex conjugate of i(t) is obtained from i(t) by replacing all j's by (-j)'s.

(1) For the resistor,

$$p(t) = e(t)\, i(t) = RI\cos^2(\omega t)$$

$$P = RI^2\,\overline{\cos^2(\omega t)} = \frac{1}{2}RI^2 \qquad [2.1.11.3]$$

$$P = R\left(\frac{I}{\sqrt{2}}\right)^2 = R\,I_{rms}^2$$

(2) For the inductor,

$$p(t) = -\omega L I^2 \sin(\omega t)\cos(\omega t)$$

$$P = -\omega L I^2\,\overline{\sin(\omega t)\cos(\omega t)} = 0 \qquad [2.1.11.4]$$

(3) For the capacitor,

$$p(t) = \frac{I}{\omega C}\sin(\omega t)\cos(\omega t)$$

$$P = \frac{I}{\omega C}\,\overline{\sin(\omega t)\cos(\omega t)} = 0 \qquad [2.1.11.5]$$

From Equations [2.1.11.3] through [2.1.11.5], it is evident that real power is dissipated only in the resistive portion of an impedance.

The power dissipated in an arbitrary circuit depends upon the phase angle between the voltage and current. Given the voltage and current:

$$e(t) = E\cos(\omega t)$$

$$i(t) = I\cos(\omega t + \phi) \qquad [2.1.11.6]$$

The instantaneous power supplied to the circuit is:

$$p(t) = EI\cos(\omega t)\cos(\omega t + \phi) \qquad [2.1.11.7]$$

or, using a trigonometric identity,

$$p(t) = EI\left[\cos^2(\omega t)\cos(\phi) - \sin(\omega t)\cos(\omega t)\sin(\phi)\right]$$

[2.1.11.8]

The average power dissipated in the circuit is, then:

$$P = EI\left[\overline{\cos^2(\omega t)}\cos(\phi) - \overline{\sin(\omega t)\cos(\omega t)}\sin(\phi)\right]$$

$$P = EI\cos(\phi)$$

$$P = \left(\frac{E}{\sqrt{2}}\right)\left(\frac{I}{\sqrt{2}}\right)\cos(\phi) = E_{rms}\, I_{rms}\cos(\phi)$$

[2.1.11.9]

Thus, the average power dissipated in a circuit containing non-resistive (reactive) components is equal to the product of the RMS voltage and current, multiplied by the cosine of the phase angle between voltage and current. The cosine of the phase angle is called the power factor.

2.1.12 Kirchoff's Laws -- Kirchoff's voltage and current laws are the basis for all circuit analysis. Using these principles, the currents and voltages of even the most complex circuits can be determined.

Kirchoff's voltage law states:

> "The sum of the voltages across all components around a closed circuit (loop) is zero."

Consider the circuit shown in Figure 2.1.12.1. From Kirchoff's voltage law, we have:

$$E_{S1} - E_{R1} - E_{R2} - E_{S2} - E_{R3} = 0 \qquad [2.1.12.1]$$

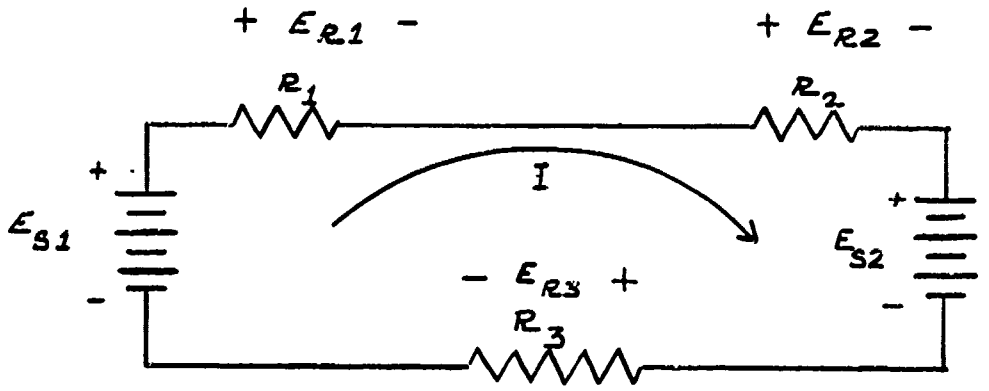


Figure 2.1.12.1 -- Series Resistance Circuit

From Ohm's Law, we have:

$$E_{R1} = I R_1 \; ; \; E_{R2} = I R_2 \; ; \; E_{R3} = I R_3 \qquad [2.1.12.2]$$

Combining Equations [2.1.12.1] and [2.1.12.2], we can solve for the current in the circuit (I).

$$I = \frac{E_{S1} - E_{S2}}{R_1 + R_2 + R_3} \qquad [2.1.12.3]$$

Kirchoff's current law states:

"The sum of the currents flowing into (and out of) a circuit junction (node) is zero."

Consider the circuit shown in Figure 2.1.12.2. From Kirchoff's current law, at Node 1, we have:

$$I_S - I_{R1} - I_{R2} = 0 \qquad [2.1.12.4]$$

From Kirchoff's voltage law, we have:

$$E_{R1} - E_{R2} = 0 \qquad [2.1.12.5]$$

From Ohm's law, we have:

$$E_{R1} = I_{R1} R_1 \; ; \; E_{R2} = I_{R2} R_2 \qquad [2.1.12.6]$$

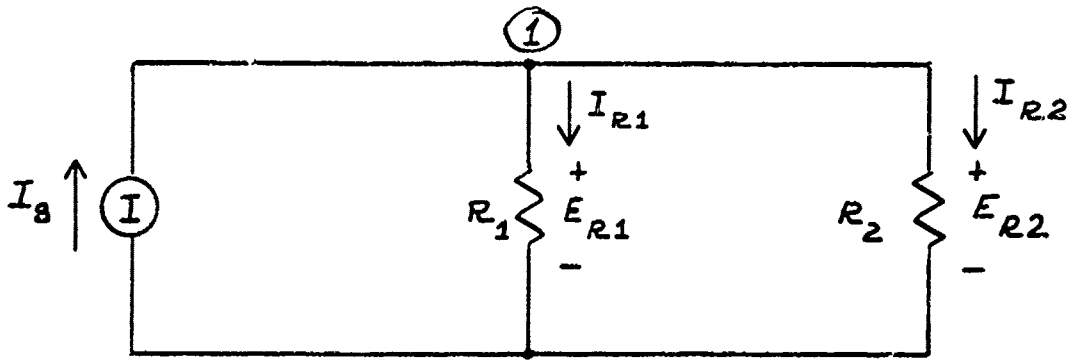


Figure 2.1.12.2 -- Parallel Resistance Circuit

Combining Equations [2.1.12.4] to [2.1.12.6], we can solve for the current in the resistor R1.

$$I_{R1} = \left(\frac{R_2}{R_1+R_2}\right) I_s \qquad [2.1.12.7]$$

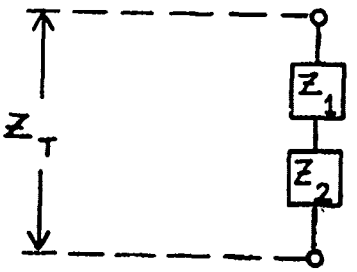
2.1.13 Series and Parallel Circuits -- A consequence of Kirchoff's laws are the rules for combining circuit impedances in series and in parallel.

Combined in series, the impedances add as shown in Figure 2.1.13.1 (a). Note, however, that combining resistors or inductors in series results in a total circuit resistance or inductance equal to their sum, while combining capacitors in series results in a total circuit capacitance smaller than that of either capacitor alone.

Combined in parallel, the reciprocals of the impedances (admittances) add as shown in Figure 2.1.13.1 (b). Note that combining capacitors in parallel results in a total circuit capacitance equal to their sum, while combining resistors or inductors in parallel results in a smaller total circuit resistance or capacitance.

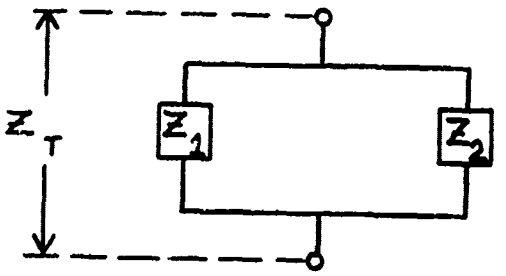
2.1.14 Tuned Circuits -- An electrical circuit that contains both capacitive and inductive components will tend to "ring" when excited. That is, it willoffer a low series impedance to currents or voltages with a frequency close to its resonant frequency. The resonant frequency is determined by the values of inductance and capacitance. In Figure

(a) Series Impedances



$$Z_T = Z_1 + Z_2$$

(b) Parallel Impedances



$$\frac{1}{Z_T} = \frac{1}{Z_1} + \frac{1}{Z_2}$$

$$Z_T = \frac{Z_1 Z_2}{Z_1 + Z_2}$$

Figure 2.1.13.1 -- Series and Parallel Impedances

2.1.14.1 are shown series and parallel tuned circuits containing pure inductances and capacitances.

In Figure 2.1.14.1 (a), the current flowing through the series L-C circuit is assumed to be:

$$i(t) = I \cos(\omega t) \qquad [2.1.14.1]$$

The voltage across the circuit is, then:

$$e(t) = -\left(\omega L - \frac{1}{\omega C}\right) I \sin(\omega t) \qquad [2.1.14.2]$$

and the (complex) impedance of the circuit is:

$$Z_{LC} = j\left(\omega L - \frac{1}{\omega C}\right) \qquad [2.1.14.3]$$

From Equation [2.1.14.3], it can be seen that the impedance of an ideal series-tuned circuit is zero at the (resonant) frequency, for which

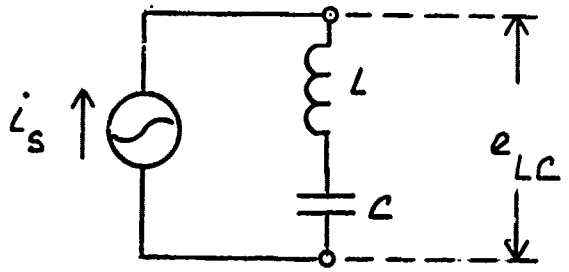
$$\omega L = \frac{1}{\omega C}$$

or

$$\omega_R = \frac{1}{\sqrt{LC}} \qquad [2.1.14.4]$$

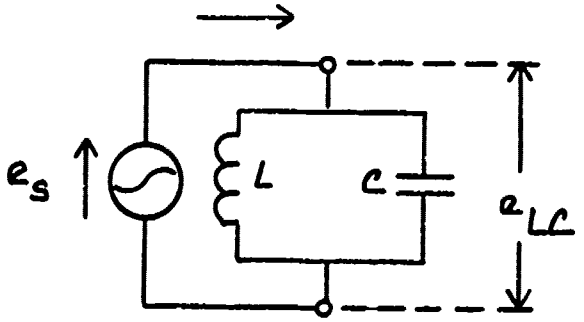
A series-tuned circuit is often employed as a frequency "trap". That is, a filter that "shorts out" signals of a prescribed frequency.

(a) Series - Tuned Circuit



$$i_s = I \cos(\omega t)$$

(b) Parallel - Tuned Circuit



$$e_s = E \cos(\omega t)$$

Figure 2.1.14.1 -- Tuned Circuits

In Figure 2.1.14.1 (b), the voltage applied across the parallel L-C circuit is assumed to be:

$$e(t) = E\,Cos(\omega t) \qquad [2.1.14.5]$$

The current through the circuit is, then:

$$i(t) = \left(\frac{1}{\omega L} - \omega C\right) E\,Sin(\omega t) \qquad [2.1.14.6]$$

and the impedance of the circuit is:

$$Z_{LC} = j\left(\frac{1}{\frac{1}{\omega L} - \omega C}\right) \qquad [2.1.14.7]$$

From Equation [2.1.14.7], it can be seen that the impedance of an ideal parallel-tuned circuit is infinite at the resonant frequency, given by:

$$\omega_R = \frac{1}{\sqrt{LC}} \qquad [2.1.14.8]$$

A parallel tuned circuit is often employed as a band-pass filter. That is, a filter that passes only signals of a prescribed frequency.

2.1.15 Series R-L-C Circuits -- In Figure 2.1.15.1 is shown a series R-L-C circuit. This circuit is assumed to be excited with a current given by:

$$i(t) = I\,Cos(\omega t) \qquad [2.1.15.1]$$

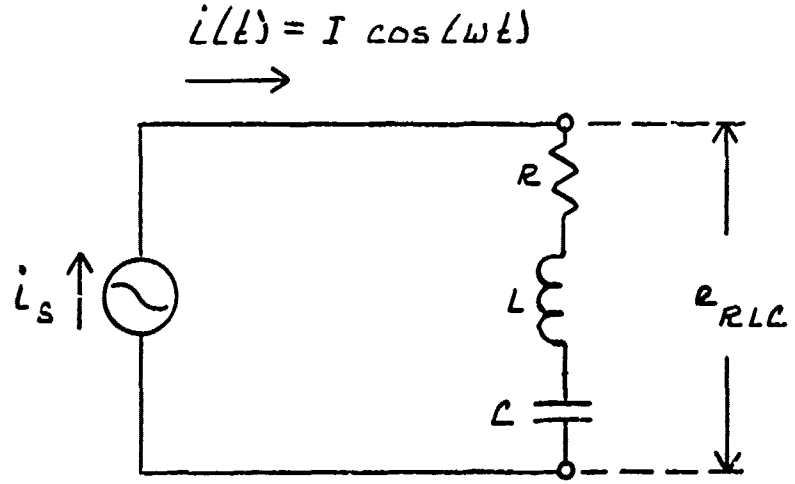


Figure 2.1.15.1 -- Series R-L-C Circuit

Applying Kirchoff's voltage law, the total voltage across the R-L-C combination is:

$$e_{RLC} = IR\,Cos(\omega t) + I\left(\omega L - \frac{1}{\omega C}\right) Sin(\omega t) \qquad [2.1.15.2]$$

or, employing Equation [2.1.8.2],

$$e_{RLC} = I\sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}\; Cos(\omega t + \phi) \qquad [2.1.15.3]$$

where:

$$\phi = Tan^{-1}\left[\frac{1-\omega^2 LC}{\omega RC}\right] \qquad [2.1.15.4]$$

At the resonant frequency, $\omega_R = \frac{1}{\sqrt{LC}}$, the reactive portions of the circuit impedance cancel; the phase angle, ϕ, goes to zero; and the voltage becomes, simply:

$$e_{RLC} = IR\,Cos(\omega t) \qquad [2.1.15.5]$$

Thus, at its resonant frequency, the series R-L-C impedance is:

$$Z_{RLC} = \frac{e(t)}{i(t)} = R \qquad [2.1.15.6]$$

This result is typical of tuned circuits, (such as antennas), operating at their resonant frequencies. The reactive part of the impedance

vanishes, the phase shift between voltage and current goes to zero, and the real power transfer to the circuit is at a maximum.

2.1.16 Impedance Matching -- In Figure 2.1.16.1 is shown a simple, one-loop circuit involving a power source and a load. A D.C. circuit is chosen in order to simplify the following analysis. The subject of the analysis is the selection of a load resistance, R_L, so as to "optimize performance" for a given source resistance, R_S.

For a source voltage equal to E_S, the current in the load is:

$$I_L = \frac{E_S}{R_S + R_L} \qquad [2.1.16.1]$$

The load voltage is, then:

$$E_L = \frac{R_L E_S}{R_S + R_L} \qquad [2.1.16.2]$$

and the power dissipated in the load is:

$$P_L = \frac{R_L E_S^2}{(R_S + R_L)^2} \qquad [2.1.16.3]$$

The desired "performance", not yet defined, determines the optimum value for R_L. If the required performance entails maximizing the load voltage,

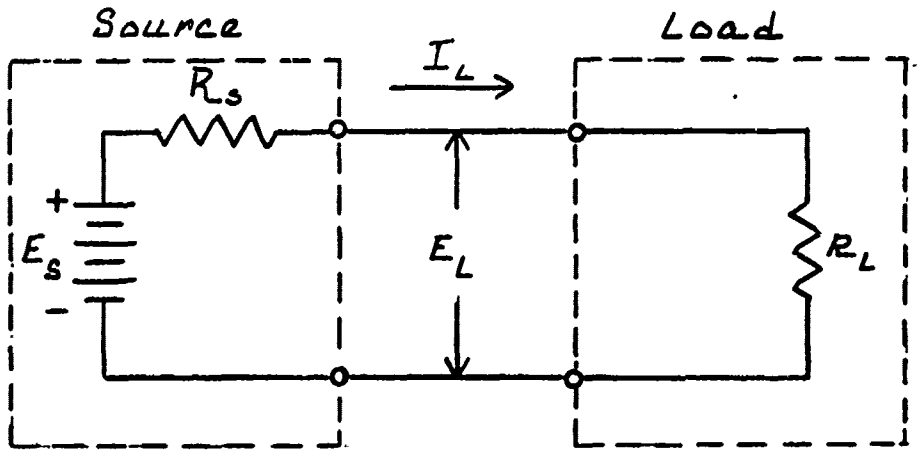


Figure 2.1.16.1 -- Source/Load Matching

Equation [2.1.16.2] indicates that the optimum value for R_L is infinity. If maximum load current is required, Equation [2.1.16.1] indicates that the optimum value for R_L is zero. In many applications, however, the requirement is not to maximize voltage or current, but to maximize power transfer from the source to the load. In order to determine the value of R_L required to maximize P_L in Equation [2.1.16.3], the calculus of variations requires that R_L satisfy the equation:

$$\frac{\partial}{\partial R_L}[P_L] = 0 \qquad [2.1.16.4]$$

Performing the partial differentiation and solving the resulting equation yields the result:

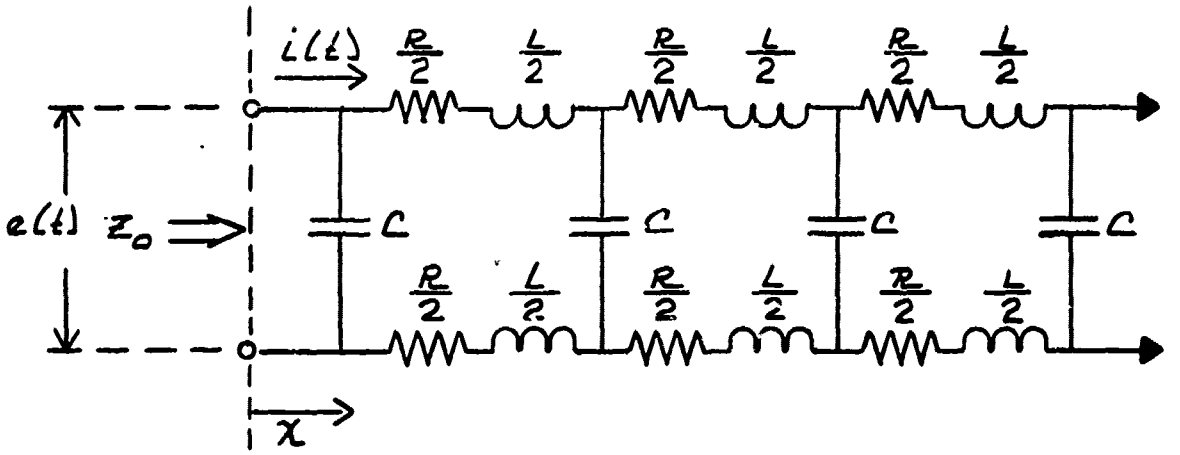
$$(R_L)_{Max\ P} = R_S \qquad [2.1.16.5]$$

Thus, for maximum power transfer from the source to the load, the load impedance must be matched to the source impedance. In order to maximize the radiant output from a transmitting station, for example, the impedances of the transmitter, transmission line, antenna, and the wave-propagating medium must all be equal.

2.1.17 Distributed-Parameter Circuits -- In all of the circuits we have considered thus far, the circuit impedances (resistances, inductances, and capacitances) have been considered to be in discrete components connected together by perfect conductors and with no time-lag from point

to point. For many situztions, that assumption is valid. As the frequency of the signals increases, however, that assumption becomes increasingly poor. Specifically, when the wavelength of the signals approaches the physical size of the circuitry, the propagation time lags approach the period of the waveforms and interactions take place between the various parts of the circuit. Under those circumstances, the discrete-parameter assumption breaks down and one must utilize distributed-parameter methods which are concerned with traveling wave phenomena.

2.1.18 Transmission Lines -- An example of a situation in which a distributed-parameter analysis is necessary is that of a transmission line at high frequencies. The simplest type of transmission line is a pair of wires. At relatively low frequencies, the voltages and currents at each end of the line are not only in phase with those at the other end of the line, but are essentially identical to them. At sufficiently high frequencies (short wavelengths), however, the inherent internal inductances and capacitances of the line and wave propagation times become significant. One way of analyzing the operation of the line is to employ what is known as a lumped-parameter model. The inherent resistance, inductance, and capacitance of the line are "lumped" into many, small, discrete components as shown in Figure 2.1.18.1. It is then possible to extend discrete-parameter analysis techniques to derive equations which adequately describe the characteristics of the line.



R = Resistance per Unit Length
L = Inductance per Unit Length
C = Capacitance per Unit Length
Z_o = Characteristic Impedance

Figure 2.1.18.1 -- Lumped - Parameter Model of Transmission Line

These distributed parameter equations incorporate the effects of the resistance, inductance, and capacitance, per unit length, of the line, and are partial differential equations of the form shown in Equation [2.1.18.1].

$$\frac{\partial^2 i}{\partial x^2} - (\gamma/\omega_s)\frac{\partial^2 i}{\partial t^2} = 0 \qquad [2.1.18.1]$$

where x is distance along the line, i is the current in the line at point x and time t, γ is a propagation constant incorporating the parameters of the line, and ω_s is the signal frequency. It can be shown that Equation [2.1.18.1] is separable into two independent equations, one involving only the distance x and the other involving only time. The two equations are:

$$\frac{\partial^2 i}{\partial x^2} - \gamma^2 i = 0 \quad \text{and} \quad \frac{\partial^2 i}{\partial t^2} - \omega_s i = 0 \qquad [2.1.18.2]$$

where:

γ = Propagation Constant = $\alpha + j\beta$

α = Attenuation Constant = $R/2Z_o$

β = Phase Constant = $\omega_s\sqrt{LC}$

Z_o = Line Characteristic Impedance = $\sqrt{L/C}$

The general solution to Equations [2.1.18.2] is of the form:

$$i(x,t) = I_1 e^{-\alpha x}\sin(\omega_s t - \beta x) + I_2 e^{+\alpha x}\sin(\omega_s t + \beta x + \phi) \qquad [2.1.18.3]$$

This solution represents two traveling waves, one traveling to the right and one traveling to the left (reflected). The waves have wavelength λ and are exponentially decaying (with distance traveled) with decay constant α. The velocity of propagation v_p and wavelength λ are given by:

$$v_p = \omega_s/\beta \qquad \text{and} \qquad \lambda = 2\pi/\beta \qquad [2.1.18.4]$$

The relative phase angle between the incident and reflected waves, ϕ, depends upon the phase constant, β, and the length of the transmission line. The amplitude, I_2, of the reflected wave depends upon the reflection coefficient at the end of the line.

2.1.19 Reflected Traveling Waves -- A traveling wave is always partially reflected by any discontinuity in the characteristic impedance of the medium in which it is propagating. If, for example, the characteristic impedance of a transmission line, Z_o, (defined in Section 2.1.18), changes at some point in the line, a wave traveling down the line will be partially reflected back toward the source. The ratio of the amplitude of the reflected wave to that of the incident wave is the reflection coefficient, ρ. That is:

$$\rho = E_r/E_i \qquad [2.1.19.1]$$

where E_r and E_i are the voltage amplitudes of the reflected and incident waves, respectively. The reflection coefficient depends only upon the characteristic impedances of the line. Specifically,

$$\rho = \frac{Z_{02} - Z_{01}}{Z_{02} + Z_{01}} \qquad [2.1.19.2]$$

where Z_{01} and Z_{02} are the characteristic impedances of the line before and after the discontinuity, respectively.

2.1.20 Wave Interference -- Due to the nature of electromagnetic waves, two waves of the same frequency can combine in such a way as to add (constructive interference) or subtract (destructive interference), depending upon their relative phase. The incident and reflected waves traveling in a transmission line are of the same frequency and will, therefore, interfere constructively or destructively, depending upon the phase angle, ϕ, the phase constant, β, and the distance along the line, x. If the resulting amplitude is plotted as a function of the distance x, it will exhibit a form similar to that shown in Figure 2.1.20.1, known as a "standing wave". In general, the standing wave "waveform" is periodic, but non-sinusoidal. The maximum and minimum values of the amplitudes exhibited by the standing wave have special significance. Specifically, the ratio of the maximum to the minimum values shown in Figure 2.1.20.1, known as the voltage standing wave ratio, VSWR, is related to the magnitude of the reflection coefficient (of the discontinuity that produced the reflected wave) by the relationship:

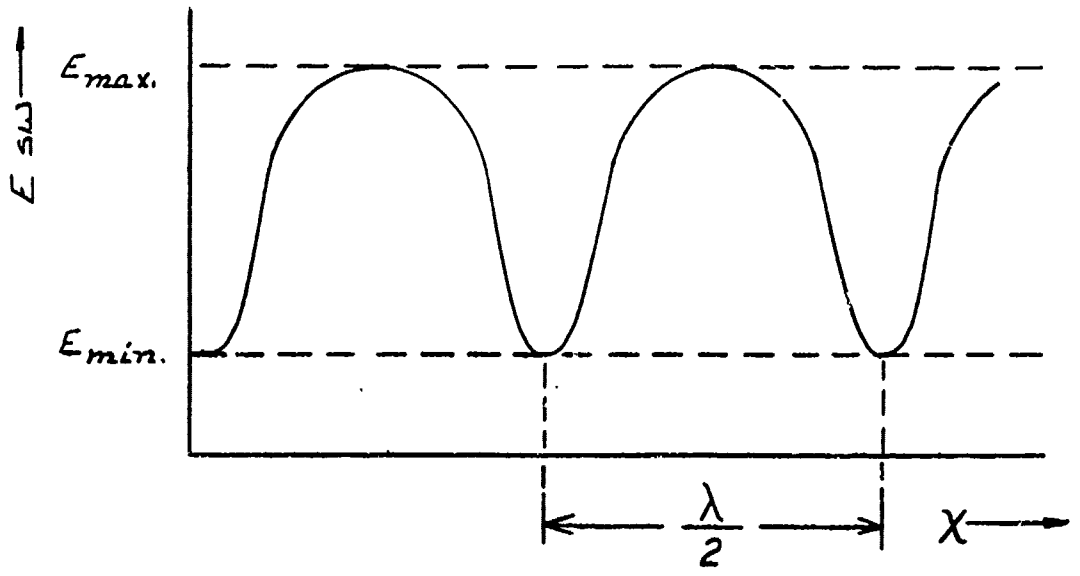


Figure 2.1.20.1 -- Standing Wave Voltage Amplitude

$$|\rho| = \frac{VSWR - 1}{VSWR + 1} \qquad [2.1.20.1]$$

where:

$$VSWR = \frac{E_{Max}}{E_{Min}} \qquad [2.1.20.2]$$

Also, as previously indicated:

$$\rho = \frac{Z_{o2} - Z_{o1}}{Z_{o2} + Z_{o1}}$$

Thus, the VSWR can be measured in order to determine the degree of impedance mismatch in a transmission system.

2.1.21 The Decibel -- In communications theory, the range of many variables is very large. For that reason, many quantities are expressed in decibels which are based on logarithms. Specifically, the amplitudes of signals (volts, amperes, etc.) are expressed in decibels defined by:

$$E(dB) = 20 \log_{10}(E/E_{Ref}) \qquad [2.1.21.1]$$

where E_{Ref} is equal to unity unless otherwise stated. Quantities involving power, (the square of amplitude), are expressed decibels defined by:

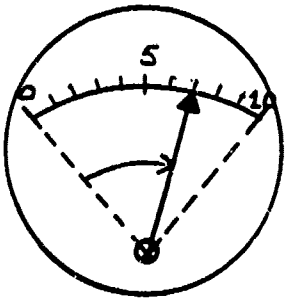
$$P(dB) = 10 \log_{10}(P/P_{Ref}) \qquad [2.1.21.2]$$

where P_{Ref} is equal to unity unless otherwise stated. In this text, we will almost always be concerned with the "power decibel" defined by Equation [2.1.21.2].

2.2 Digital Systems

2.2.1 The Nature of Digital Systems -- Due to the numerous and varied forms in which digital techniques are applied, the essential characteristic of a digital system is difficult to recognize. The one characteristic which identifies a system as digital is the use of a symbolic code to convey information. (An analog system employs the magnitude of an analog quantity for that purpose.) As an example, consider the D'Arsonval meter depicted in Figure 2.2.1.1. This instrument utilizes a current-carrying coil, placed in a magnetic field and restrained by a spring, to measure electrical current. The deflection of a pointer attached to the coil is proportional to the current. Clearly, the basic D'Arsonval movement is an analog device. The pointer deflection is the analog of the current. Even when graduations are put on the face of the meter, the instrument is still analog. The marks merely aid in estimating the pointer deflection. They quantize or discretize the deflection, but they do not digitize it. Thus, a discrete or quantized-data system is not necessarily digital, although all digital systems manipulate discrete quantities. When numerals (symbolic code characters) are added, however, the meter becomes digital. The scale with numerals is , in fact, an analog-to-digital converter.

A significant characteristic of a "symbolic code" is that its meaning depends upon prior definition. One might infer the coil current directly from the angular deflection of the pointer (given information about the



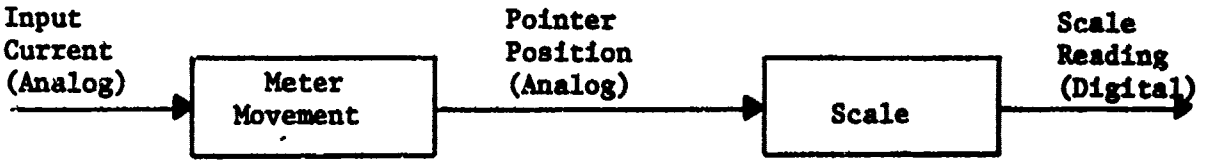


Figure 2.2.1.1 -- D'Arsonval Meter

meter's construction). But one would require prior definition of the meaning of the character "5", for example. The character "5" is a symbolic code, in the decimal system, for a count equal to the number of fingers on a human hand. Another example of a symbolic code is the character "101" which, in binary arithmetic, has the same meaning as that assigned to the character "5" above. Another essential characteristic of a digital system is the use of a pre-determined set of rules to manipulate the symbolic codes. (Actually, the rules are part of the definitions for the symbols). An example of these rules is the operation specified by the (decimal) symbols:

"5" "+" "5" "=" "10" [2.2.1.2]

Another example is the set of operations hard-wired into a digital computer arithmetic logic unit.

In recent times, the application of digital techniques has proliferated to the extent that there remains scarcely an engineering discipline not rapidly being "digitized." The reason for this proliferation is a reversal of the normal relationship between the need for a technology and the development of that technology. Digital technology has become so advanced, and offers such advantages in performance, that designers are now actively seeking applications for digital technology, rather than seeking a technology to satisfy their applications. The result is not always a superior, or even acceptable, design. A digital system has important weaknesses as well as strengths.

2.2.2 Advantages of Digital Systems -- The principal advantages of digital systems are listed below.

Flexibility

Accuracy

Noise Rejection

Long-Term Stability

Simplicity of Binary Devices

Physical Characteristics

The most important advantage is flexibility. The operation of a system controlled by digital software, (rather than hardware), can be extensively altered, in the field, merely by inputting new instructions (software). Furthermore, when a general purpose digital computer is employed, that computer can be used for many different systems.

The advantage in accuracy derives not from an inherent accuracy of digital systems, but from the fact that the precision can be improved, (at the expense of other system characteristics), at the discretion of the designer. That is, the number of digits in a computer word can be increased until the desired precision is obtained. Similarly, the number of iterations in a successive-approximation process or terms in a power series expansion can be adjusted. In most cases, an improvement in accuracy will be accompanied by a penalty in memory required or computing time. Generally, the accuracy of a digital system is determined by the analog-to-digital (A/D) and digital-to-analog (D/A) converters utilized at the input and output (I/O) ports. The excellent noise rejection characteristics of

most digital systems are primarily the result of using binary numbers. When, due to noise, an analog voltage is in error by ten percent, the parameter value represented by that voltage is in error by ten percent. Such is not the case in a binary digital system. Only two digit values are possible, zero and one. Any error in system signal levels, due to noise or otherwise, will be totally rejected until that error reaches a magnitude sufficient to make the machine mistake a zero for a one, or vice versa. Thus, theoretically, the errors can approach fifty percent of the signal without effect. On the other hand, digital systems are more susceptible to short-term errors than are analog systems. A voltage "spike" that would be ignored by a slow-moving analog system can cause a miscount in a high-bit-rate digital system. The susceptibility of digital systems to such "glitches" can be alleviated, in part, by the use of error-correcting codes and redundant computations.

Long term drift is a severely limiting characteristic of the devices used for analog time integration. In contrast, the output of a digital integrator (pulse counter) remains constant as long as the input remains below the threshold for a "one". For that reason, the long-term stability of digital systems incorporating integrators is excellent.

An inherent advantage of binary digital systems is the simplicity of devices which need only recognize two states (e.g. zero and one). This simplicity is the key to economics in the manufacture of large-scale

digital systems. It is also responsible for the impressive economies in the size, weight, and power requirements of modern digital computers.

2.2.3 Disadvantages of Digital Systems -- The disadvantages of digital systems are as impressive as the advantages. The principal disadvantages are listed below.

Susceptibility to Gross Errors

Difficult Software Validation

Sampled-Data Effects

Quantization Effects

Analog/Digital Conversions

Slow Integration

Large Number of Elements

The most subtle disadvantage, and perhaps the most important for man-rated systems, is the propensity of digital systems to make gross errors. The errors in question are not errors in computer words used to represent the magnitudes of quantities (such errors are generally not catastrophic); the troublesome errors are errors in words used as addresses in memory or as instructions. For example, a simple one-bit error in an address word, if not corrected, could result in the use of the output from a pitch-axis gyro, in a digital flight control system, in place of the yaw-axis gyro output. The same error in an analog flight control system would probably require the use of a wrench or a soldering iron! Another type of gross error commonly encountered in digital systems is that known as a "jump the track" error, in which the system begins executing an entirely unwanted

portion of the digital program as a result of misreading a bit in an instruction word. For example, the flight control system of the Space Shuttle Orbiter might suddenly apply the atmospheric reentry autopilot equations while in the terminal phase of landing. Of course, provisions are made to detect and correct such "jump the track" errors.

From an economic standpoint, perhaps the most important disadvantage of digital systems is the difficulty in test and evaluation of the associated software. This process, called software validation and verification, is greatly complicated by the very flexibility of the digital system. In a complex system, there are so many alternate paths and branches in the execution of the software program that only another digital computer can examine them all; and, even then, the time and cost may be prohibitive. In the development of today's digital systems, more money is spent in software V and V than in the original software design.

One of the more serious disadvantages of digital systems is the sampled-data nature of the signals involved. The fact that the signals are not continuously updated is not usually a problem. Techniques of interpolation and extrapolation suffice. The most important sampled-data effect is frequency aliasing or folding, the generation of spurious signals at frequencies above and below the frequency of the original signal. This process can, for example, fold noise or other unwanted signals down into the bandwidth of a data processing system even though the original signal

frequencies were too high in frequency to be of concern. Frequency aliasing is discussed in detail in Section 2.5.6 of this text.

Quantization effects (discontinuities in the available values of the signal quantities) are not usually a serious problem only because the bit-size of the system can usually be made small enough to prevent the sudden changes from seriously affecting system performance. The principal effect of an insufficiently small quantum size is the destabilization of an otherwise stable closed-loop system.

Also related to quantization effects are the inherent problems encountered with digital integrators. Although all physical processes are, in fact, quantized, the quantum sizes are so infinitesimal as to be imperceptible. Therefore, the integration of physical quantities is essentially a continuous process. Digital integration is, at best, an approximation. In order to minimize the "jumps" inherent in digital integration, a very small integration interval must be used, sometimes orders of magnitude smaller than the smallest periods present in the signals being processed. For that reason, it is sometimes impossible to achieve "real-time" digital integration in, for example, a process-control system.

The representative signals of most mechanisms are analog. Therefore, D/A and A/D converters are necessary as interface units between the digital components and the physical system. It was previously noted that these devices are usually the limiting factors in system accuracy. They also

are the greatest contributers to the needs for space, weight, and power. Finally, these converters are usually the components which introduce the most troublesome quantization effects. In an analog system, there is, of course, no need for A/D and D/A converters.

For a required system speed-of-response, digital signal processing requires much more hardware frequency-bandwidth capability than does analog processing. The bandwidth requirements for digital signal processing are discussed in Section 2.5.5 of this text.

It was previously noted that the inherent simplicity of binary devices make for economy of production. The resulting standardization also creates high reliability (large mean-time-between-failures). The very limited information-carrying capacity of a single binary device makes it necessary, however, to incorporate large numbers of them in a practical system. Thus, although the MTBF of a single element is large, the MTBF for the entire system can be a problem.

2.2.4 Binary Codes -- Binary codes are characterized by the designation of only two possible states; e.g. one and zero, plus and minus, on and off, high and low, yes and no, etc. All information can be imparted utilizing only two such states. Mathematically, binary codes thus employ a radix of two. Decimal codes employ a radix of ten. For example, in the decimal system:

$$1306_{10} \equiv 1 \times 10^3 + 3 \times 10^2 + 0 \times 10^1 + 6 \times 10^0 \qquad [2.2.4.1]$$

and, in the binary system:

$$1011_2 \equiv 1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 = 11_{10} \qquad [2.2.4.2]$$

Values less than unity are constructed in the same manner in the decimal and binary systems. For example, in the decimal system:

$$1.52_{10} \equiv 1 \times 10^0 + 5 \times 10^{-1} + 2 \times 10^{-2} \qquad [2.2.4.3]$$

or

$$1.52_{10} \equiv 1 + 0.5 + .02$$

And, in the binary system:

$$1.01_2 \equiv 1 \times 2^0 + 0 \times 2^{-1} + 1 \times 2^{-2} \qquad [2.2.4.4]$$

or

$$1.01_2 \equiv 1 + 0 + 1/4 = 1.25_{10}$$

The "decimal point" in binary arithmetic is, of course, called a binary point.

To convert a binary number to its decimal equivalent:

(1) Convert each binary digit (bit) to its decimal equivalent.

(2) Sum the decimal equivalents.

For example:

$$1011_2 \equiv 1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 \qquad [2.2.4.5]$$

$$1011_2 \equiv 8 + 0 + 2 + 1$$

$$1011_2 \equiv 11_{10}$$

To convert a decimal number to its binary equivalent:

(1) From the decimal number, subtract the decimal equivalent of the largest-value-possible binary digit.

(2) Add that binary digit to the binary number being determined.

(3) Repeat steps (1) and (2) as required to reduce the decimal number to zero.

For example, to convert 11_{10} to binary:

11 (decimal number)
- 8 => 1000
3
- 2 => 0010
1
- 1 => 0001
0 1011 (binary equivalent)

An examination of the table of binary numbers presented in Figure 2.2.4.1 reveals that, in the straight binary system, changes often occur in more than one digit as the number changes by a single count. Due to hardware tolerances, it is impossible to ensure that two or more such transitions

DEC. No.	Cyclic Binary Code (Gray Code)				Straight Binary Code			
0	0	0	0	0	0	0	0	0
1	0	0	0	1	0	0	0	1
2	0	0	1	1	0	0	1	0
3	0	0	1	0	0	0	1	1
4	0	1	1	0	0	1	0	0
5	0	1	1	1	0	1	0	1
6	0	1	0	1	0	1	1	0
7	0	1	0	0	0	1	1	1
8	1	1	0	0	1	0	0	0

Figure 2.2.4.1 -- Cyclic Binary Codes

will occur simultaneously in practice. Depending upon the order in which the transitions occur, large transition errors can arise. There are codes, such as the Gray code shown in the figure, that allow only one bit to change at a time, thereby limiting transition errors to a single count. Such codes, called cyclic codes, are generally employed in digital systems for operations, such as analog-to-digital conversions, where ambiguous transitions are a problem.

It is possible to construct codes which provide the means for detecting errors in reading them. The brute force approach is simple redundancy. The value to be encoded is incorporated twice into the code word and the two read-in values are compared. Only identical errors in reading the two sets of bits would result in a failure to detect an error. A simpler and more common error-detecting provision is the addition of a "parity bit" to the extreme right end of the digital word. Parity in a binary word is the state of being "odd" or "even"; that is, containing an odd or an even quantity of "ones". If the extra "parity" bit is set equal to zero when the number of "ones" in a word is otherwise even, and equal to one otherwise, the resulting word, including the parity bit, will always have even parity. If a simple bit in the word is misread, a parity check wll reveal the error. An even number of such errors will escape detection, but multiple errors are relatively unlikely.

The sign of a binary number is designated by attaching a "sign" bit to the left end of the code word. The sign bit is a "one" if, and only if,

the designated quantity is negative. In order to expedite the handling of negative numbers, the symbol is generally "complemented" in addition to having the sign bit affixed. Subtraction of a number is then performed by addition of its complement. There are two types of binary word complements in common use. The so-called "ones complement", is obtained by replacing all zeros in the original word by ones in the complement, all ones by zeros, and attaching the sign bit. Thus, for example:

$$5_{10} = \overbrace{0}^{\text{Sign Bit}} | 1\ 0\ 1_2$$

$$5^c_{10} = 1 | 0\ 1\ 0_2 \qquad [2.2.4.6]$$

The twos complement is obtained by: replacing all zeros and ones in the original word by ones and zeros in the complement, adding one to the least significant bit of the result, and attaching the sign bit. Thus, for example:

$$5_{10} = \overbrace{0}^{\text{Sign Bit}} | 1\ 0\ 1_2$$

$$5^c_{10} = 1 | 0\ 1\ 1_2 \qquad [2.2.4.7]$$

In Section 2.2.7 of this text, it is demonstrated that adding the twos complement of a binary number is the equivalent of subtracting that number.

The formula for obtaining the complement, N^c, of a number, N, in a number system of radix R is:

$$N^c = R^n - N \qquad [2.2.4.8]$$

where n is the number of digits in N, including the sign bit.

2.2.5 Other Digital Codes -- There are a number of digital codes, other than binary, that are commonly encountered. One such code is binary-coded decimal (BCD). The BCD code provides a natural interface between the decimal system, to which humans are accustomed, and the binary system, in which almost all digital computers operate. In BCD, the decimal digits are individually expressed in four-digit binary. Thus, for example:

$$312_{10} \equiv 0011/0001/0010_{BCD} \qquad [2.2.5.1]$$

Another non-binary code is octal, which has a radix of eight. (The octal symbols and their decimal equivalents are shown in Figure 2.2.5.1 (a).) The octal code is often employed in digital systems as a convenience in expressing numbers which would be unwieldy if expressed directly in binary. For example, the binary equivalent of 312_{10} is given by:

$$312_{10} \equiv 100111000_2 \qquad [2.2.5.2]$$

The octal equivalent is given by:

$$312_{10} \equiv 470_8 \qquad [2.2.5.3]$$

Thus, the octal number is nearly as compact as the decimal equivalent. The convenience of the octal code is a result of the ease of conversion

(a) Octal Code Symbols

0	1	2	3	4	5	6	7	Octal
0	1	2	3	4	5	6	7	Decimal

(b) Hexadecimal Code Symbols

0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	Hexadecimal
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Decimal

Figure 2.2.5.1 Octal and Hexadecimal Codes

between binary and octal. In order to convert a binary number to octal:

(1) Mark off the digits in the binary word into groups of three to either side of the binary point.

(2) Convert each group of three bits independently into octal. For example, the procedure for converting the binary number 11010.10001_2 (26.531_{10}) to its octal equivalent, 32.42_8 is shown in Figure 2.2.5.2 (a). To convert octal to binary, convert each octal digit independently into binary. Another attractive feature of the octal code is that it makes full use of the binary digits. That is, it takes exactly three binary digits (a count of seven) to express the values of the digits in an octal word.

A digital code frequently employed as the language of communication at the machine/human interface is hexadecimal, which has a radix of sixteen. (The hexadecimal symbols and their decimal equivalents are shown in Figure 2.2.5.1 (b)). The advantages of hexadecimal are exactly those of octal, but with provision for more unique symbols (16). Note the economy of four bits in expressing hexadecimal words. Conversions between binary and hexadecimal are performed in a manner similar to those for octal. To convert a binary number to hexadecimal:

(1) Mark off the digits in the binary word into groups of four to either side of the binary point.

(2) Convert each group of four bits independently into hexadecimal. For example, the procedure for converting the binary number 10001111.101_2 (143.625_{10}) to its hexadecimal equivalent, $8F.A_{16}$, is shown in Figure

(a) Conversion of 11010.100010_2 to 32.42_8

0	1	1	0	1	0	1	0	0	0	1	0	(Binary)
		3			2			4			2	(Octal)

(b) Conversion of 10001111.1010_2 to 8 F.A_{16}

1	0	0	0	1	1	1	1	1	0	1	0	(Binary)
			8				F				A	(Hexadecimal)

Figure 2.2.5.2 Binary-to-Octal and Binary-to-Hexadecimal Conversions

2.2.5.2 (b). To convert hexadecimal, convert each hex digit independently into binary.

2.2.6 General Radix Conversion Procedure -- The general procedure for converting a digital word from one number system (radix) to another can be considered in two parts: the procedure for integer numbers and the procedure for non-integer numbers.

The procedure for converting integer numbers is as follows:

(1) Divide the old number by the new radix. (All operations must be performed in the old number system.)

(2) Record the remainder from Step (1) as the least significant digit (LSD) in the new number.

(3) Divide the quotient from Step (1) by the new radix.

(4) Record the remainder of Step (3) as the next LSD in the new number.

(5) Repeat Steps (3) and (4) until the quotient is zero (no integral quotient possible).

(6) Record the final remainder as the most significant digit (MSD) of the new number.

For example, the procedure for converting the decimal number 25_{10} to its binary equivalent, 11001_2 is shown in Figure 2.2.6.1 (a).

The procedure for converting non-integer numbers is as follows.

(1) Convert integer portion of old number using procedure for integer numbers.

(a) Conversion of 25_{10} to 11001_2

12	6	3	1	0
2 ⌊25	2 ⌊12	2 ⌊6	2 ⌊3	2 ⌊1
-24	-12	-6	-2	-0
1	0	0	1	1
↑ LSD	↑	↑	↑	↑ MSD

(b) Conversion of 0.625_{10} to 0.101_2

0.625	0.250	0.500
X 2	X 2	X 2
1.250	0.500	1.000
↑ MSD	↑	↑ LSD

Figure 2.2.6.1 Binary-to-Decimal Conversions

(2) Multiply less-than-integer portion of old number by new radix.

(3) Record integer portion of product in Step (2), (the overflow), as the most significant digit in the new number.

(4) Multiply less-than-integer portion of the product in Step (3) by the new radix.

(5) Record the integer portion of the product in Step (4) as the next MSD in the new number.

(6) Repeat Steps (4) and (5) until sufficient significant figures in the new number have been obtained.

(7) Record the final overflow as the least significant digit in the new number. For example, the procedure for converting the decimal number 0.625_{10} to its binary equivalent, 0.101_2 is shown in Figure 2.2.6.1 (b).

2.2.7 Binary Arithmetic -- The rules (set theory) governing the arithmetic manipulation of binary numbers are identical to those governing arithmetic operations with decimal numbers. Addition and subtraction are performed exactly as with decimal numbers, keeping in mind the fact that each column has twice the value of its neighbor to the right. Thus, when you carry to, or borrow from, the next column, you carry or borrow two units rather than ten as in the decimal system. Examples of binary addition and subtraction are shown in Figure 2.2.7.1 (a) and (b). An example of subtraction by addition of the twos complement is shown in Figure 2.2.7.1 (c). Note that, after the addition, the overflow from the sign bit is discarded. Examples of binary multiplication and division are shown in Figure 2.2.7.2. Note that, in multiplication, you repetitively add either the multiplicand

(a) Addition of 9_{10} and 3_{10}

Decimal System	Binary System
9	1001
+ 3	+ 11
= 12	= 1100

(b) Subtraction of 3_{10} from 9_{10}

Decimal System	Binary System
9	1001
- 3	- 11
= 6	= 110

(c) Subtraction of 3_{10} from 9_{10} by twos Complement Addition

+ 3 = 0	0011
- 3 = 1.	1101 (3^C)

Decimal System	Binary System
9	0 | 1001
- 3	+ 1 | 1101 (3^C)
= 6	0 | 0110

Figure 2.2.7.1 -- Binary Addition and Subtraction

(a) Multiplication of 12_{10} and 13_{10}

Decimal System	Binary System
12	1100
X 13	X 1101
36	1100
12	0000
= 156	1100
	1100
	= 10011100

(b) Division of 156_{10} by 13_{10}

Decimal System	Binary System
12	1100
13 ⟌ 156	1101 ⟌ 10011100
- 13	- 1101
26	1101
- 26	1101
0	000

Figure 2.2.7.2 -- Binary Multiplication and Division

itself or zero. Thus, multiplication in a computer requires only shift and add operations. Division requires only shift, complement, and add operations. Shift registers are discussed in a later section of this text.

2.2.8 Digital Computer Functions -- The principal functions of the digital computer are those listed below.

(1) Storage and Retrieval of Data

(2) Execution of Programmed Sequences

(3) Numerical Calculation

(4) Logical Manipulation

(5) Timing and Counting

The first item, storage and retrieval of data, is self-explanatory. Many digital computer programs do nothing more than accept data, store it, retrieve it, and output it on command. The execution of programmed sequences is another non-computational task frequently performed by digital computers. This task requires only the storage of an instruction sequence and a clocking or counting operation. Digital numerical operations include addition, subtraction, multiplication, division, exponentiation, integration, differentiation, and function generation. It is important to note that many logical (non-numerical) operations are executed in digital computers. The binary format is well suited to handle any logical manipulation that requires a (two-state) yes or no conclusion. All logical requirements can be formulated as a sequence of yes-or-no decisions.

Digital computers incorporate "clocks"; that is, devices which output a precisely-timed series of pulses. The clock output allows timing, sequencing, and counting operations.

All of these operations entail the storage, interpretation, and manipulation of computer words. A computer word is a symbol usually consisting of ones and zeros (binary) that conveys one or more of three types of information. The three possible types of information represented by a computer word are:

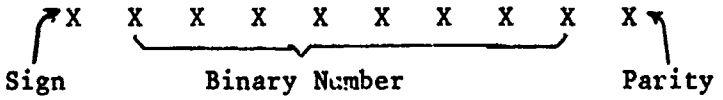
(1) a numerical quantity

(2) an instruction code

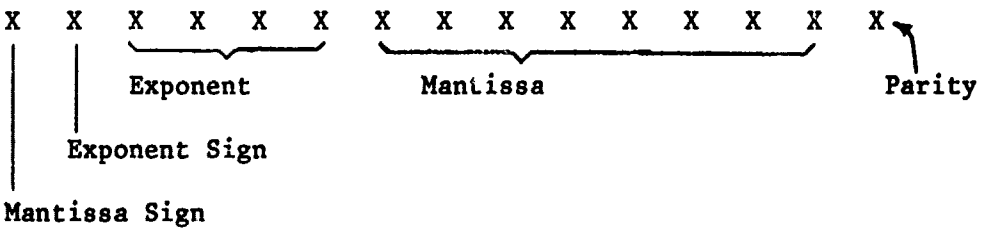
(3) a storage address

2.2.9 Computer Word Format -- Three digital computer word formats are shown in Figure 2.2.9.1. In Figures 2.2.9.1 (a) and (b) are shown, respectively, the fixed-point and floating-point formats used for representing numerical quantities. The fixed-point format has three parts: a sign bit, a parity bit, and several bits representing the magnitude of the quantity in straight binary form. The fixed-point format has the advantage of simplicity in use and manipulation. It has the disadvantage of imposing a fixed limit on the magnitude of the quantity represented. The floating-point format represents the magnitude of the quantity in so-called scientific notation. (Scientific notation employs one binary number representing the mantissa of the quantity and a separate binary number representing powers of two.) The floating-point format thus consists

(a) Fixed - Point Format



(b) Floating- Point Format



(c) Instruction Format

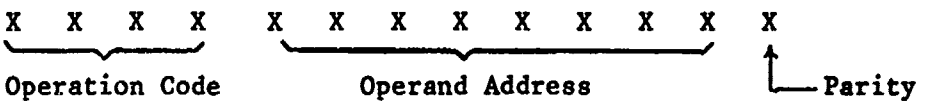


Figure 2.2.9.1 -- Digital Computer Word Formats

of two sign bits (one for mantissa and one for exponent), a parity bit, several bits for the mantissa, and several bits for the exponent. The number of bits required for the magnitude representation depends upon the magnitude and precision of the quantity to be represented.

A typical format for a computer instruction word is shown in Figure 2.2.9.1 (c). It is common practice to specify, in one word, both the operation to be performed and the address of the quantity on which it is to be performed.

An important parameter associated with a digital computer is the length of (number of bits in) its computer word. The required length of the word depends upon the following factors.

(1) Magnitude of Quantity Represented

(2) Precision (No. of Places) Required

(3) Round-Off Bits Required

(4) Sign Bits Required

(5) Parity Bits Required

The only factor not previously discussed is item (3): Round-Off Bits. If it is desired to employ the usual decimal system round-off criterion, (A remainder equal to or greater than one-half increases the least significant digit by one), then two binary digits are required in the computer word to provide the information for the round-off decision. (The value of one bit has exactly one half the value of the next higher bit).

As an example of required computer word length, consider the required length of the fixed-point word required to represent a quantity with magnitude up to 1000_{10}. Such a word would allow a system precision of 0.1% (one part in one thousand). In order to equal (or exceed) a count of 1000_{10}, a digital word requires ten bits (giving a count of 1023_{10}.) In addition, the word requires a sign bit and a parity bit, and two round-off bits, for a total of fourteen. Thus,a 0.1% fixed-point system requires a word length of fourteen bits.

Consider, now, the word-length requirement for the same precision in floating point. The floating-point word would still require ten bits (a count of 1023) to express the mantissa to one part in a thousand. In addition, it will require four bits (a count of 15) to express the exponent (10), two sign bits (one for the mantissa, one for the exponent), two bits for mantissa round-off, and one bit for parity, for a total of nineteen bits. Thus, it takes five more bits to achieve 0.1% precision in floating-point than in fixed-point arithmetic.

As an example of the word-length requirements for instructions, consider the length required for a repertoire of one hundred instruction codes. The unique designation of (at least) one hundred instructions requires a word length of seven bits (a count of 128). In addition, a parity bit is required, making a total of eight bits for an instruction set of one hundred.

As an example of word-length requirements for words used as storage (or input/output port) addresses, consider the requirement for an eight thousand (8K) word (address) memory. The unique designation of (at least) eight thousand addresses requires a word length of thirteen bits. In addition, a parity bit is required, making a total of fourteen bits.

2.2.10 Digital Computer Architecture -- In Figure 2.2.10.1 is shown the general architecture of a typical digital computer. The "heart" of the computer is the central processor unit (CPU), where numerical and logical manipulations are normally performed. The control unit retrieves instructions and data stored in memory, decodes the instructions, and causes them to be executed, with the retrieved data, by issuing a sequence of commands to the arithmetic logic unit (ALU), the memory, and the Input/Output (I/O) ports.

The arithmetic logic unit executes the basic numerical and logical manipulations. Some ALU's are limited to addition and subtraction but most can also perform multiplication and division without detailed instructions. It should be noted that the operations performed by any digital computer can be of two kinds: hard-wired and software programmed. The basic operations; addition, subtraction and, sometimes, multiplication and division; are hard-wired into the computer hardware. Other, special-purpose instructions can also be hard-wired in. A computer devoted to special hard-wired operations, with little or no general purpose capability,

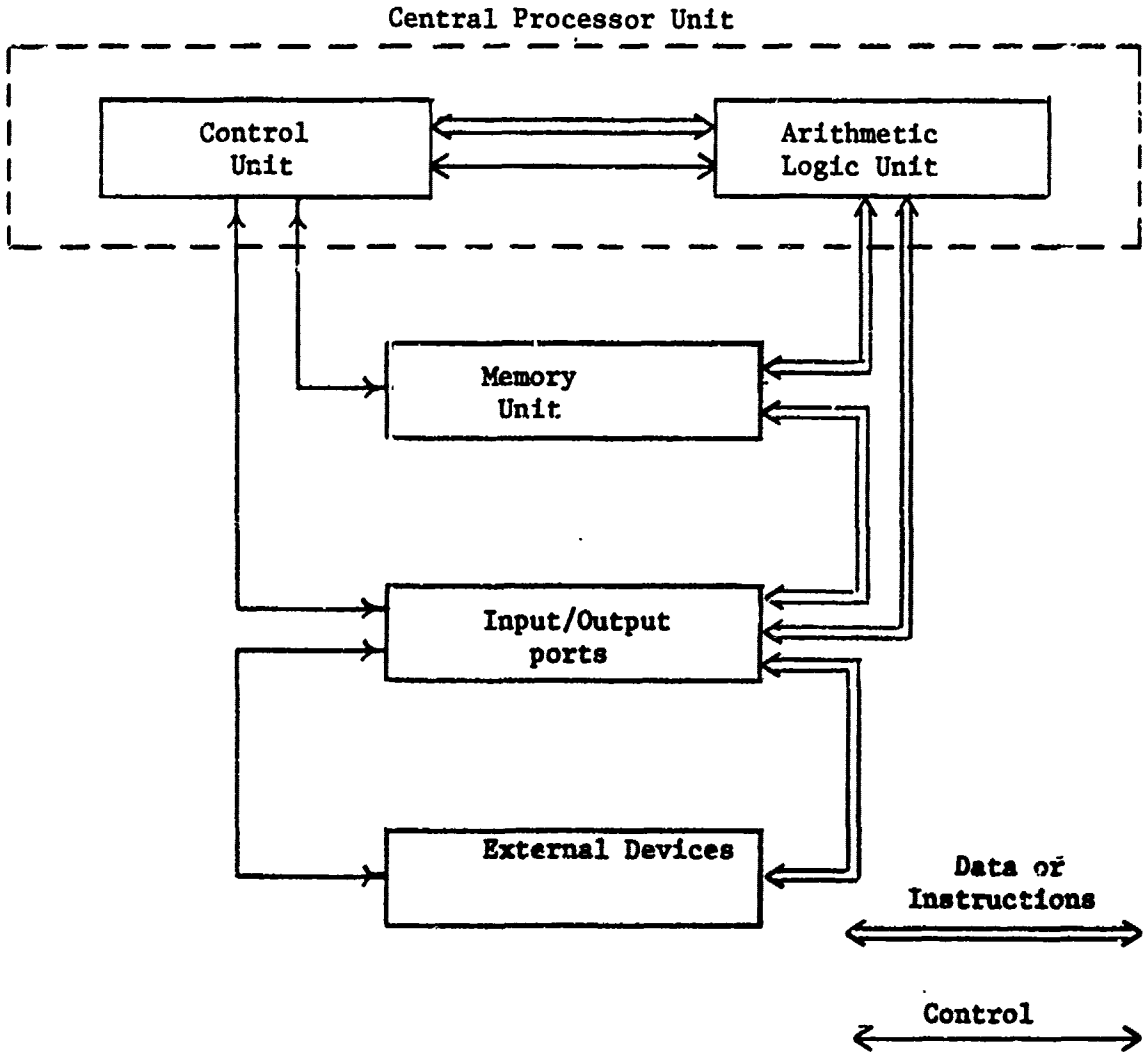


Figure 2.2.10.1 -- Architecture of Typical Digital Computer

is called a special purpose computer. Generally, special purpose computers are faster, smaller, and cheaper than general purpose machines. They are, of course, limited to the tasks hard-wired into the CPU.

The memory unit receives, stores, and outputs digital words as required by the control unit. The stored information can be one of several kinds, with information of different types and for different purposes stored in portions of the memory reserved for that purpose. One portion of memory may be reserved for temporary storage of information being inputted from a peripheral device, another for storage of information being outputted, and a third for storage of information being manipulated by the CPU (working storage). Other sections of memory are used to store the program software and data for the program being executed. Still other portions of memory are used for storing special purpose computational routines, (such as the power series expansion for computing sin (x)), or tabulated values of functions, (such as sin (x) versus x).

Currently available memory units make various provisions for the reading (inputting) and writing (outputting) of information. The most common types of memory are:

(1) Read and Write

(2) Read Only (ROM)

(3) Destruct Read-Out (DRO)

(4) Non-Destruct Read-Out (NDRO)

(5) Serial Access (SAM)

(6) Random Access (RAM)

(7) Addressable

A read-and-write memory, as the name implies, is any memory that allows both the inputting and outputting of data during computation. A read-only memory is one that requires special provision or apparatus to input information; and, thus allows only outputting data during computation. A destruct-read-out memory is one, such as magnetic core, for which the information is lost from memory by the act of reading it out. A non-destruct-read-out memory is one for which the information is not lost from memory in reading it out. A serial-access memory is one, such as magnetic tape, that requires that the entire memory be scanned until the desired item is found. A random-access memory is one, such as magnetic core, that allows any desired item to be examined independently. An addressable memory is one that allows manipulation of a stored word without first outputting it to the CPU and then returning it to memory.

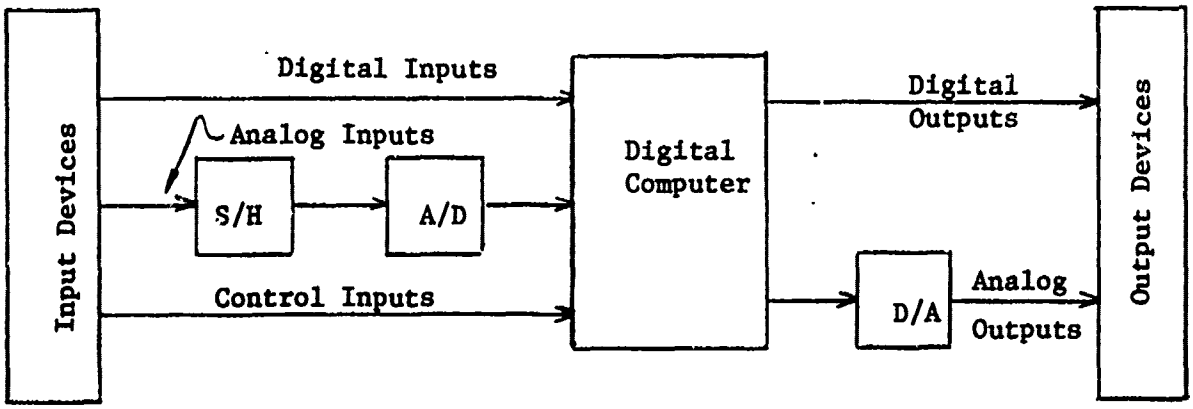
The computer input/output ports are data registers with provision for inputting or outputting the stored word as directed by the control unit or external device. The word may be transferred either serially (one bit at a time, in sequence) or in parallel (all bits at the same time). Serial data handling is slower but requires only one data line. Sometimes provision is made to translate information from one form to another; for example, from keyboard character symbols to machine language binary code. The external devices are any devices which interface with the

computer; such as other computers, data lines, keyboards, external memory, display consoles, and recorders.

2.2.11 Digital System Architecture -- The block diagram of a generic digital system is shown in Figure 2.2.11.1. The digital computer, (described in Sections 2.2.8 to 2.2.10 of this text), is, of course, the principal performer in a digital system. There are, however, other components essential to the functioning of the system; namely, the input/output (I/O) devices and the interface units required to "translate" the signals exchanged between the computer and the external devices.

The input devices are of two kinds: those that provide control inputs and those that provide data inputs. Typical control input devices are control switches, program tape readers, and other computers. Typical data input devices are keyboards (human operator inputs), data tape readers, sensors, and other computers.

Output devices are also of two kinds: those that display information to the system operator, and those that utilize computer outputs to perform system functions. Some computer outputs are discrete, designating the occurrence of events. Others are continuous, representing time-varying quantities. Typical output devices are displays, recorders, actuators, and other computers.



S/H = Sample - and - Hold Circuit

A/D = Analog - to - Digital Converter

D/A = Digital - to - Analog Converter

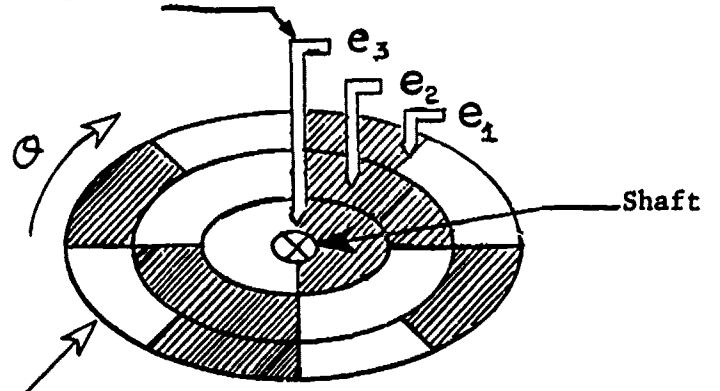
Figure 2.2.11.1 -- Generic Digital System

Interface units include A/D and D/A converters, coder/decoders, and signal conditioners. Analog-to-digital converters are required when analog inputs are included, since the digital computer accepts only digitally-coded signals. Similarly, digital-to-analog converters are required when analog outputsare required.

2.2.12 Digital Computer Hardware -- An electromechanical analog-to-digital converter is illustrated in Figure 2.2.12.1 (a). The device shown is a shaft position encoder, consisting of a shaft-driven disk with three brushes riding on segmented tracks. The tracks are divided into alternately conducting and non-conducting segments arranged in a binary pattern. When a brush contacts a conducting segment (shaded) the corresponding output voltage, e, is V. When the brush contacts a non-conducting segment, the output voltage is zero. Thus, the input to the converter is analog shaft position. The output consists of three voltages, (e_1, e_2, and e_3), which represent a three-bit, parallel-output, binary digital representation of the input. As the shaft angle increases in the direction indicated, the output voltages will assume the values shown in the graphs of Figure 2.2.12.1 (b). Therein, it can be seen that e_1 represents the least-significant-bit and e_3 the most-significant-bit of a three-bit straight binary code.

An electrical digital-to-analog converter is illustrated in Figure 2.2.12.2 (a). The device shown is a weighted-current-summing resistive network with input voltages e_1, e_2, and e_3, (such as the output voltages from the encoder shown in Figure 2.2.12.1), representing the bits in a

(a) Shaft Position Encoder Brushes



θ = Shaft Position

e = Output Voltage

Disk with conducting and non-conducting segments

(b) Output Voltage Pattern

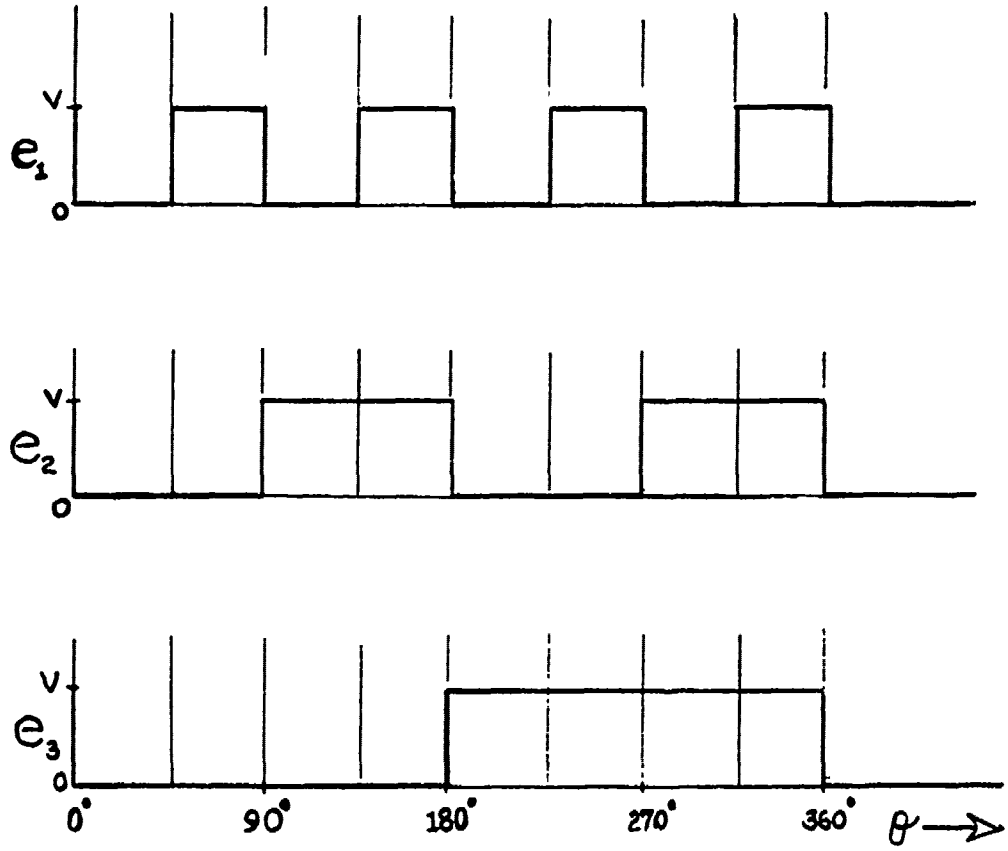
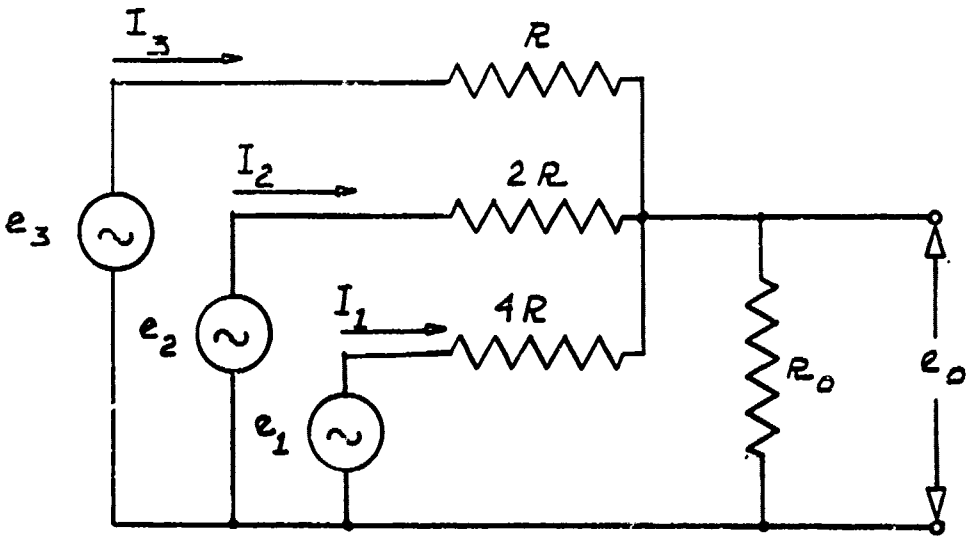


Figure 2.2.12.1 -- Analog-to-Digital Converter

(a) Weighted - Current Summing Circuit



For $R_o \gg R$:

$$e_o = \frac{1}{7}[e_1 + 2e_2 + 4e_3]$$

(b) Weighted - Current Sum

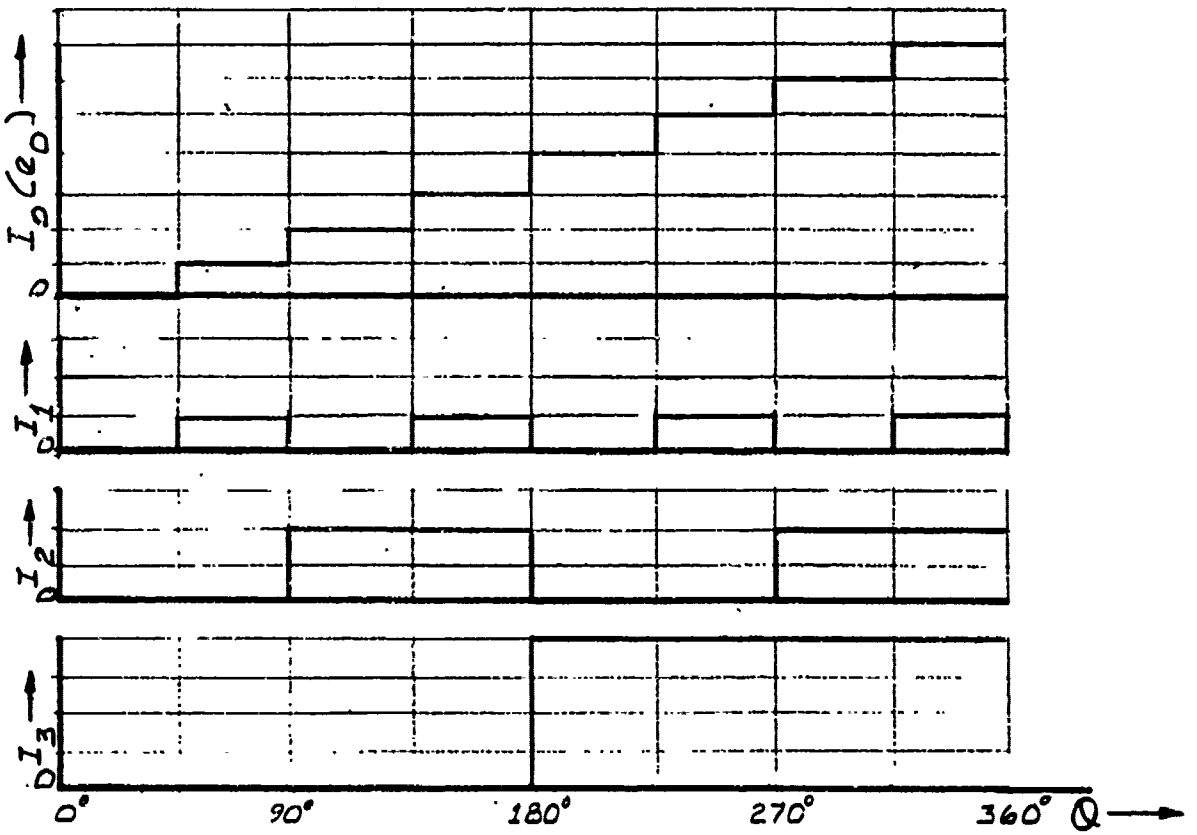


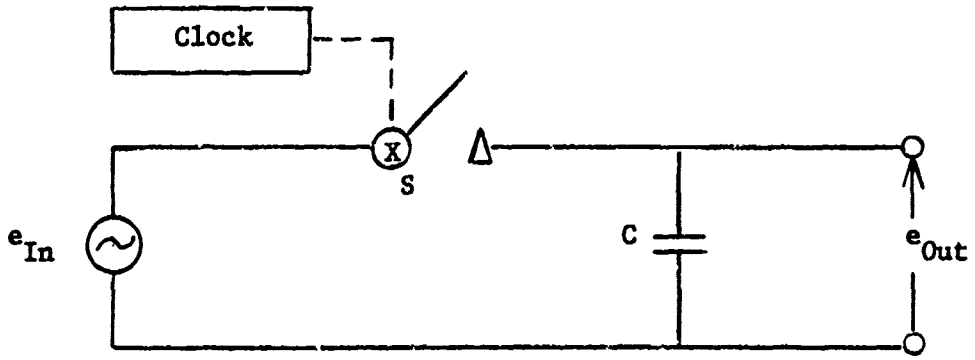
Figure 2.2.12.1 -- Digital-to-Analog Converter

straight-binary-coded word. Because of the binary weighting of the series resistors, the total current flowing in the summing resistor, R_o, is a (quantized) reconstruction of the original analog quantity represented by the three voltages. Plots of the individual and total currents versus the analog angle are shown in Figure 2.2.12.2 (b).

The input to an analog-to-digital converter must be held constant during the conversion process. If the input signal is continuously changing, it must be "frozen" by a device called a sample-and-hold circuit. A simple sample-and-hold circuit is shown in Figure 2.2.12.3 (a). The switch, s, driven by the clock, periodically closes momentarily. During switch closure, the capacitor, c, charges to the then-current value of the input voltage, e_{in}. The output voltage, e_{out}, (voltage on the capacitor) remains at the last sampled value of e_{in} during the period between switch closures. Thus e_{out} and e_{in} will have a time history similar to that shown in Figure 1.1.12.3 (b).

The basic building block of a binary digital computer is the bistable element or "flip-flop". A bistable device has two stable states and can, therefore, store a binary digit. Examples of bistable elements are "toggle" switches, relays, electronic multivibrators, charge-coupled semiconductors, and magnetic cores, wires, tapes, drums, disks, and other devices such as magnetic "bubble" memory arrays. Connected in such a way as to be free-running (continuously alternating between the two states), the multivibrator is used as the "clock" in a computer. Arranged

(a) Sample - and - Hold Circuit



(b) S and H Circuit Output

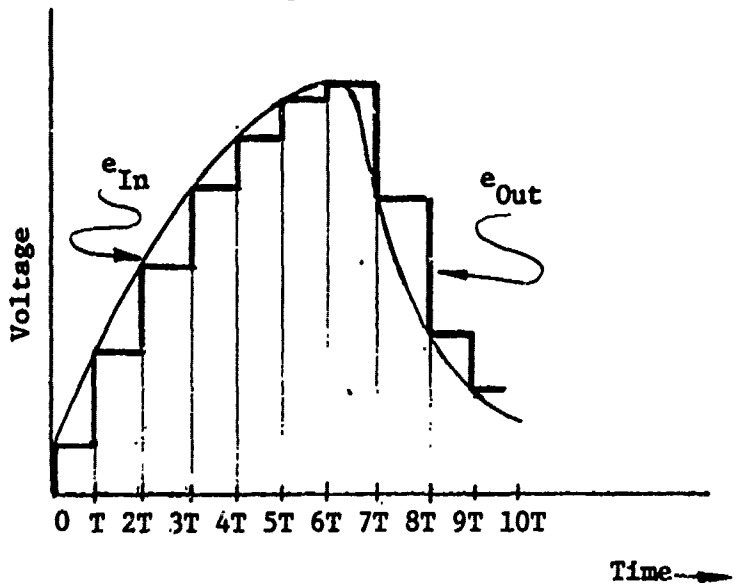


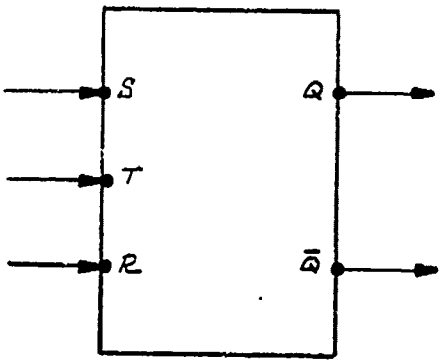
Figure 2.2.12.3 -- Sample - and - Hold Device

in a string to store the bits in a "word", a number of bistable elements comprise a "register". Registers are the devices used to store and manipulate computer words in the arithmetic logic unit. Arrays of registers constitute memories.

A common representation for a bistable element is shown in Figure 2.2.12.4. This type of flip-flop has three inputs. A pulse at the "set" input, s, sets the ouput, Q, to a "one". A pulse at the "reset" input, R, sets the output, Q, to a "zero." A pulse at the "transfer" input. T, reverses the state of the output. The complementary output, $\overline{Q}$, is always at a state opposite that of the output, Q.

In addition to flip-flops (bistable elements), there are several other basic devices employed in digital computers. These are the AND Gate, the OR Gate, the Exclusive OR (EOR) Gate, the NOT Gate or inverter, the NAND gate, and the NOR Gate. The symbol and notation for the AND Gate are shown in Figure 2.2.12.5 (a). The relationship between its inputs and output, known as a "truth table" is shown in Figure 2.2.12.5 (b). The output is a one if, and only if, both inputs are ones. A simple diode AND circuit is shown in Figure 2.2.12.5 (c).

The symbol and notation for the OR Gate are shown in Figure 2.2.12.6 (a). Its truth table is shown in Figure 2.2.12.6 (b). The output is a one if either input is a one. A simple diode OR Gate is shown in Figure 2.2.12.6 (c).



S = Set Input

T = Transfer Input

R = Reset Input

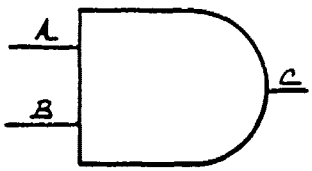
Q = Output

$\bar{Q}$ = Complementary Output

Figure 2.2.12.4 -- Bistable Element

(a) AND Gate Symbol and Notation

Symbol



Notation

C = A•B

(b) AND Gate Truth Table

A	B	C
0	0	0
0	1	0
1	0	0
1	1	1

(c) AND Gate Diode Circuit

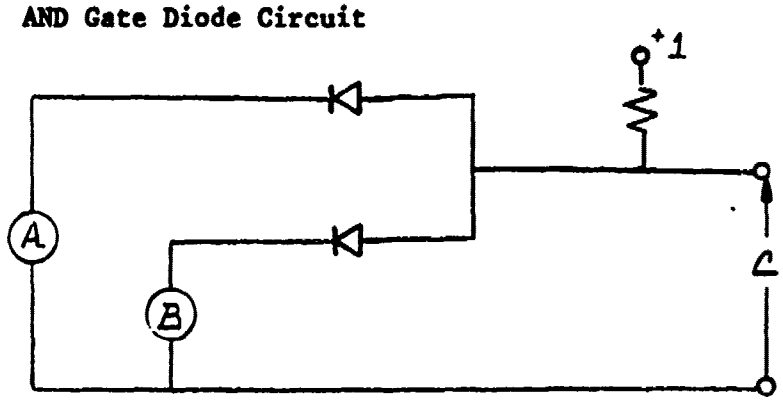
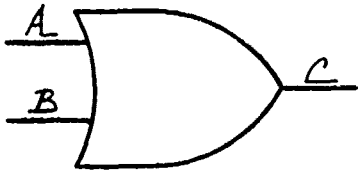


Figure 2.2.12.5 -- AND Gate

(a) OR Gate Symbol and Notation

Symbol



Notation

C = A + B

(B) OR Gate Truth Table

A	B	C
0	0	0
0	1	1
1	0	1
1	1	1

(c) OR Gate Diode Circuit

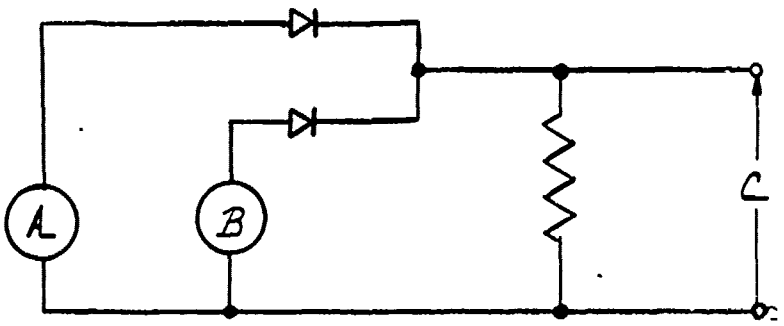


Figure 2.2.12.6 -- OR Gate

The symbol and notation for the Exclusive OR Gate are shown in Figure 2.2.12.7 (a). Its truth table is shown in Figure 2.2.12.7 (b). The output is a one if either, but not both, input is a one.

The symbol and notation for the NOT Gate, or inverter, is shown in Figure 2.2.12.8 (a). Its truth table is shown in Figure 2.2.12.8 (b). The output is a one if, and only if, the input is a zero.

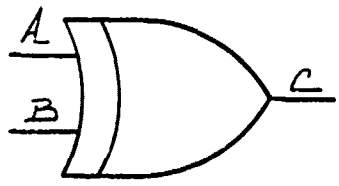
The symbol and notation for the NAND Gate are shown in Figure 2.2.12.9 (a). Its truth table is shown in Figure 2.2.12.9 (b). The output is a zero if, and only if, both inputs are ones. Note that the NAND Gate is an AND Gate followed by an inverter.

The symbol and notation for a NOR Gate is shown in Figure 2.2.12.10 (a). Its truth table is shown in Figure 2.2.12.10 (b). The output is a one if, and only if, both inputs are zeros. Note that a NOR Gate is an OR Gate followed by an inverter.

As previously indicated, a register is a string of bistable elements used to store and/or manipulate a binary word. There are three principal types of interest: the data storage register, the counter, and the shift register. Registers can be serial or parallel input and they can be serial or parallel output. A serial input/serial output register is shown in Figure 2.2.12.11 (a). Note that all digits are fed into the left-hand

(a) Exclusive OR Gate Symbol and Notation

Symbol



Notation

$C = A \oplus B$

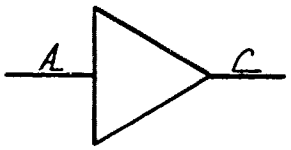
(b) Exclusive OR Gate Truth Table

A	B	C
0	0	0
0	1	1
1	0	1
1	1	0

Figure 2.2.12.7 -- Exclusive OR Gate

(a) Inverter Symbol and Notation

Symbol



Notation

$C = \overline{A}$

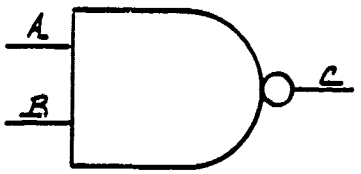
(b) Inverter Truth Table

A	C
0	1
1	0

Figure 2.2.12.8 — Inverter

(a) NAND Gate Symbol and Notation

Symbol



Notation

$C = \overline{A \cdot B}$

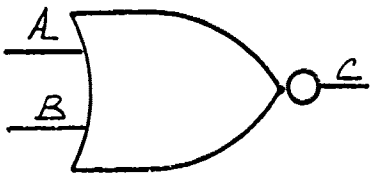
(b) NAND Gate Truth Table

A	B	C
0	0	1
0	1	1
1	0	1
1	1	0

Figure 2.2.12.9 — NAND Gate

(a) NOR Gate Symbol and Notation

Symbol



Notation

$$C = \overline{A + B}$$

(b) NOR Gate Truth Table

A	B	C
0	0	1
0	1	0
1	0	0
1	1	0

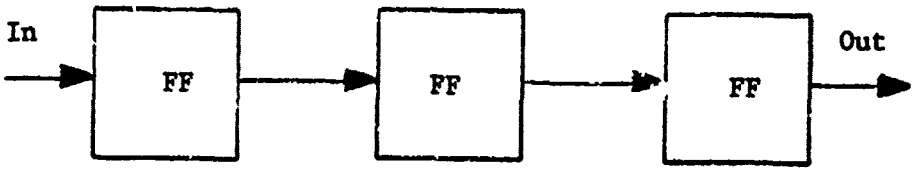
Figure 2.2.12.10 — NOR Gate

flip-flop and all digits are fed out of the right-hand flip-flop, in a serial fashion. Note also that the bits are first in, first (FIFO). A parallel input/parallel output register is shown in Figure 2.2.12.11 (b). All digits are simultaneously inputted and/or outputted, in parallel. Note the interconnecting arrows necessary for word manipulation but not for input/output. A parallel input/serial output register is shown in Figure 2.2.12.11 (c). Such a device would be a series pulse-to-parallel pulse converter.

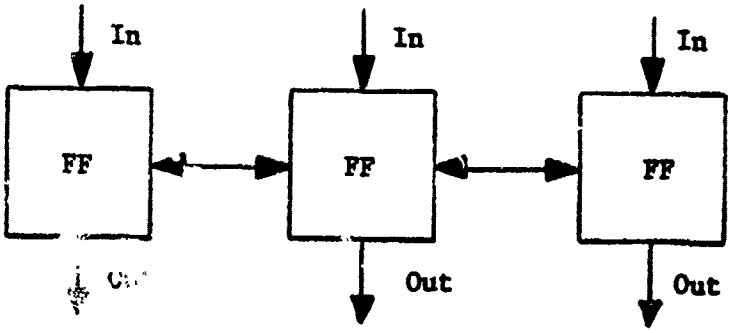
Typical data storage registers are shown in Figure 2.2.12.12. The serial input/serial output register shown in Figure 2.2.12.12 (a) inputs (and outputs) one bit for each pulse applied to the transfer terminal, T. The parallel input/output register shown in Figure 2.2.12.12 (b) inputs (and outputs) all bits simultaneously when a pulse is applied to transfer terminal, T.

A binary counter register is shown in Figure 2.2.12.13. Input pulses applied to the Input terminal result in a binary representation of the total count accumulating in the register. A pulse applied to the Reset terminal resets all bits to zero. The register can be read out in parallel, non-destructively, by applying a pulse to the parallel transfer line. The register can be read out serially, and destructively, by pulsing the serial transfer line with a pulse train.

(a) Serial Input/Serial Output Register



(b) Parallel Input/Parallel Output Register



(c) Parallel Input/Serial Output Register

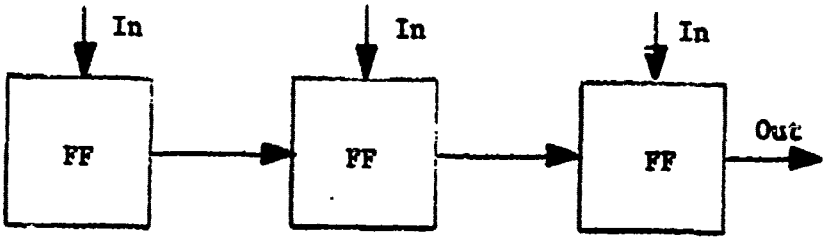
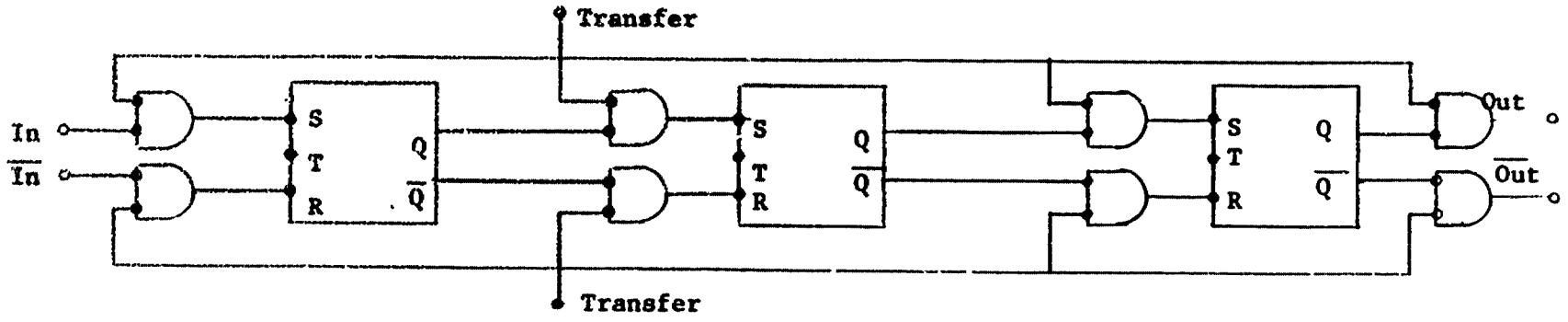


Figure 2.2.12.11 -- Binary Registers

(a) Serial Input/Output Register



(b) Parallel Input/Output Register

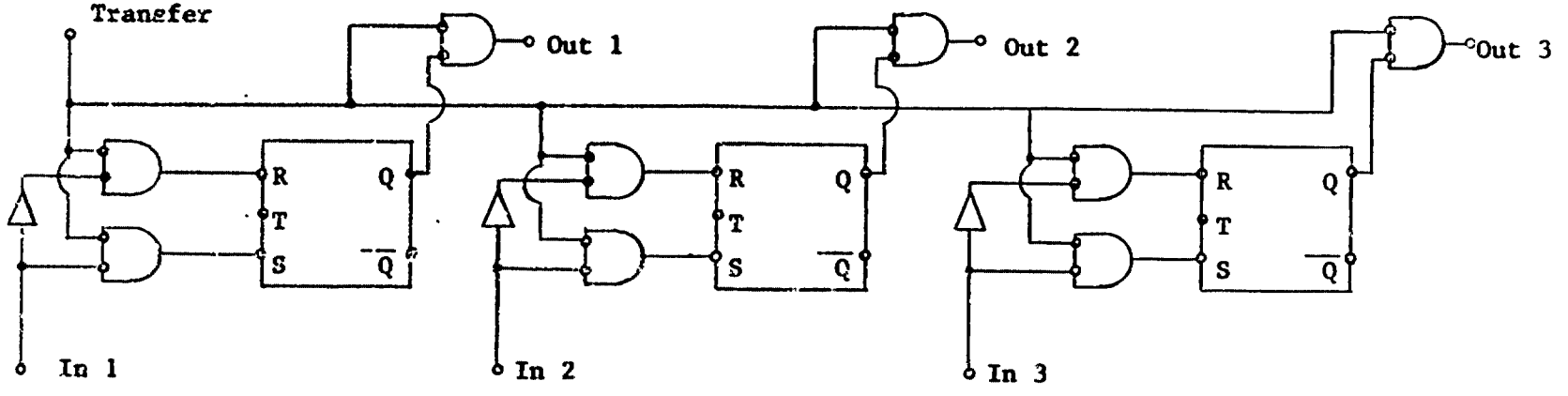


Figure 2.2.12.12 -- Data Storage Registers

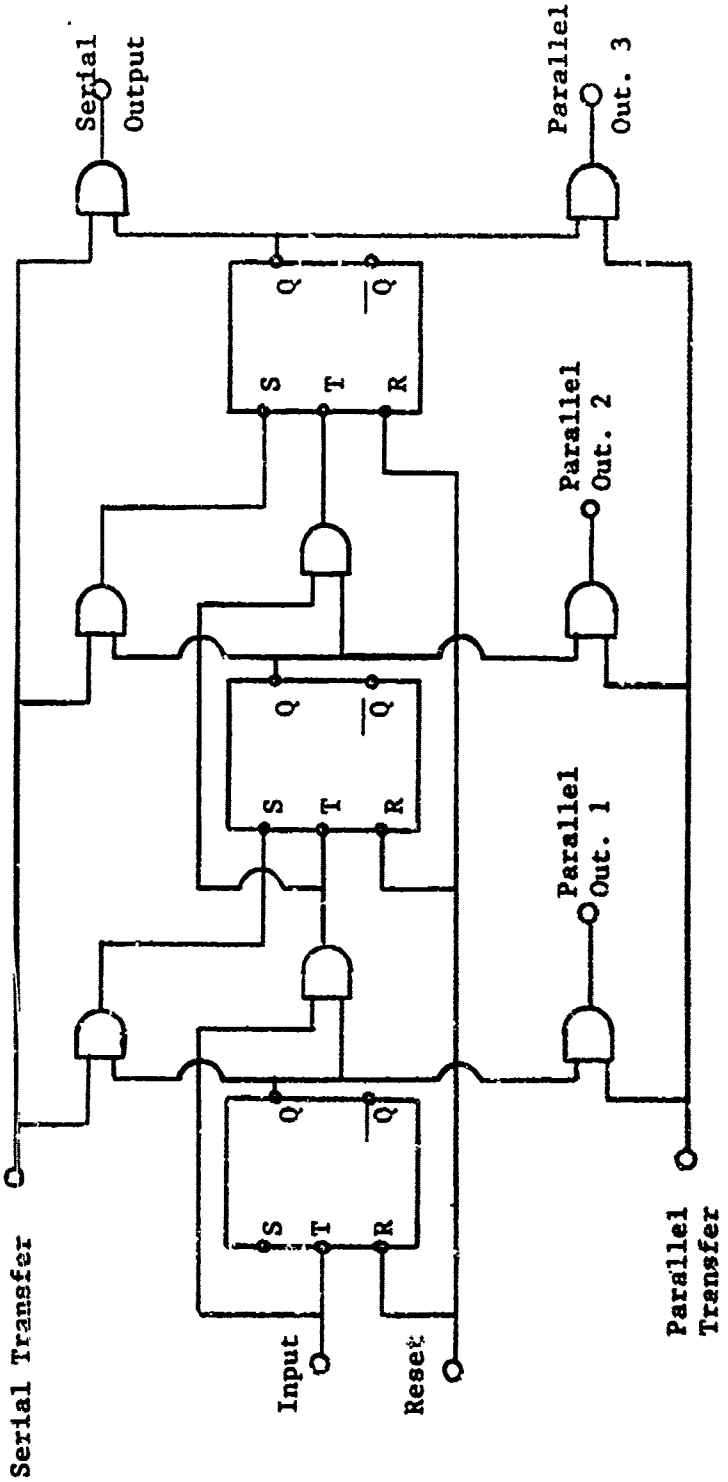


Figure 2.2.12.13 -- Binary Counter

The block diagram of a shift register is shown in Figure 2.2.12.14. Upon application of a pulse to the shift line, the contents of each bistable element is transferred to the right. The contents of the right-hand element is transferred out and the bit on the input line is transferred into the left-hand element. Thus, as it is shown, this shift register is serial input/serial output.

In Figure 2.2.12.15 is shown the block diagram of a full adder. A full adder adds two binary bits, including a carry-in bit from a previous operation and a carry-out from the current operation.

2.2.13 Digital Software -- As previously indicated, the operation of a digital computer is controlled by a set of instructions (software) either inputted from a storage (memory) device or by an operator, or both. Almost all of the systems with which we are concerned input the detailed program from storage (tape, punched cards, etc.) with only minimal inputs from a human operator. When inputted, the program is stored in memory (generally magnetic core) internal to the machine, for use during program operation. Essentially, the instructions contained in the software execute the program by setting bistable elements in the computer in the sequential pattern specified in the program.

The instructions in a digital program must, of course, be written in a code (language) understood by the computer. One such language is the "machine language" for that particular computer. Machine language is the most

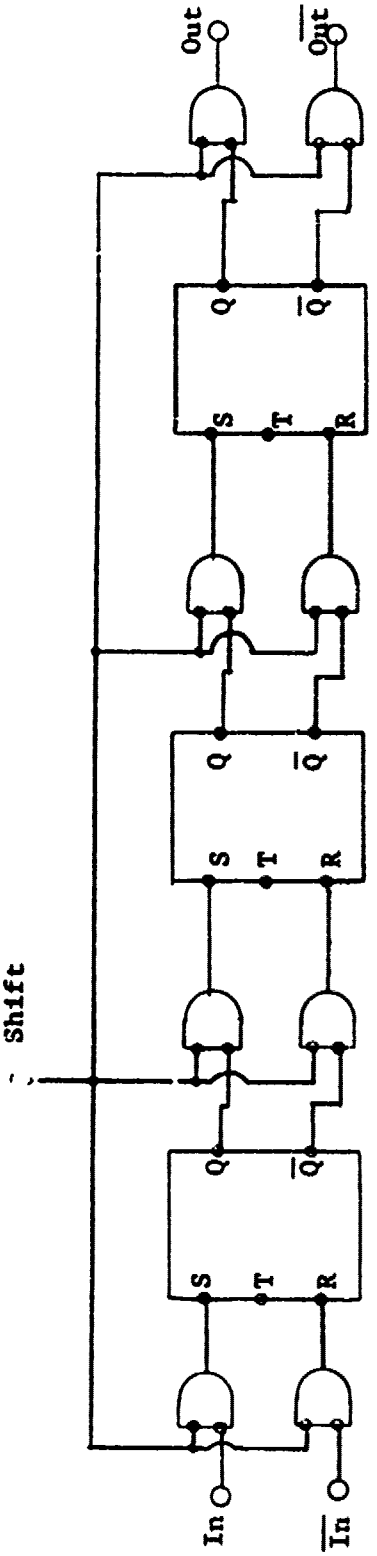


Figure 2.2.12.14 -- Shift Register

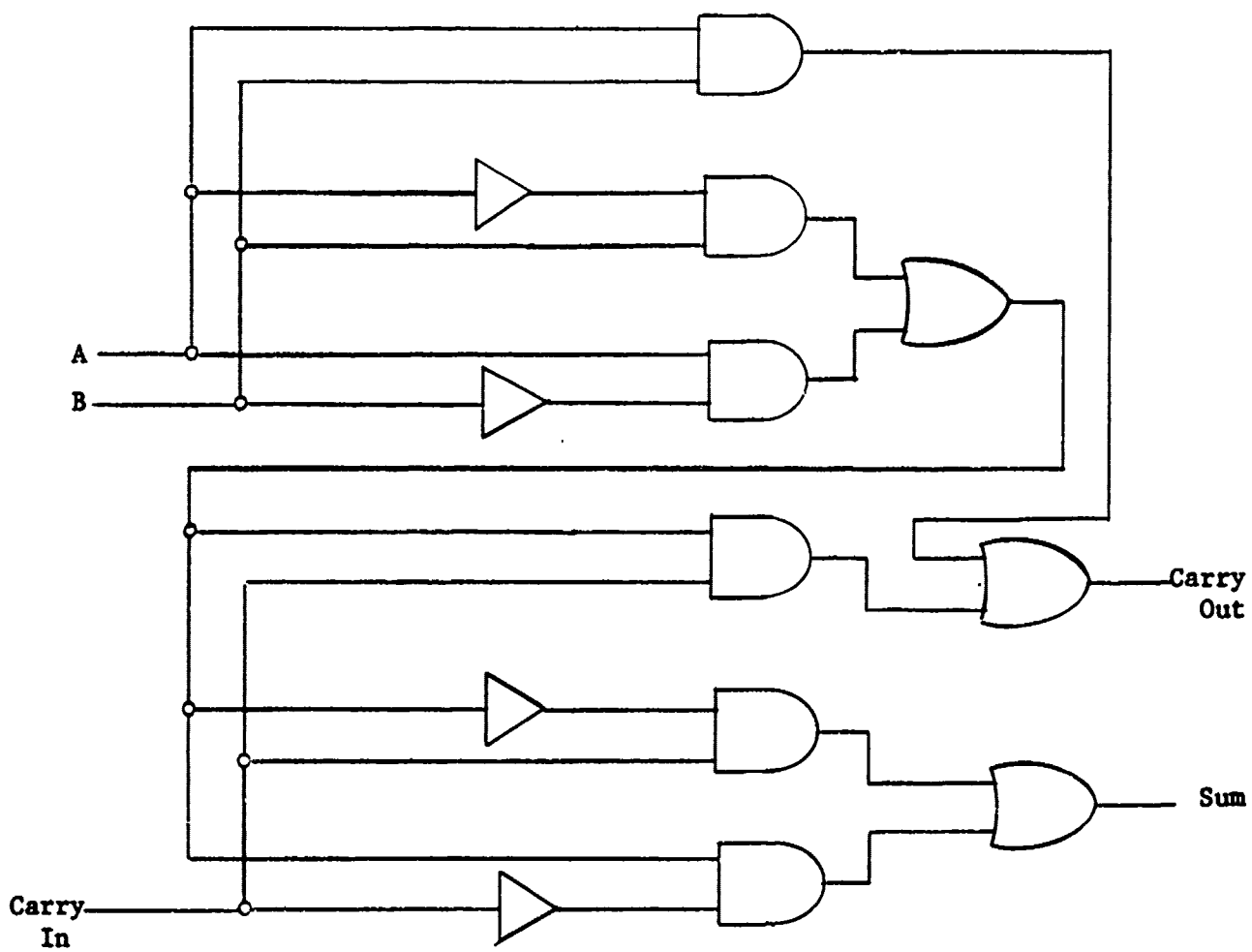


Figure 2.2.12.15 -- Full Adder

detailed (lowest-order) language a programmer can use. Higher-order languages are languages which consolidate commonly encountered groups of machine-language instructions into single instruction codes. Such languages are easier for the programmer to use than machine language but are less efficient for the machine, and require more execution time. In addition, in order to understand a higher-order language, the computer must incorporate an interpreter or compiler to "de-code" the higher-order instructions.

Because software is less tangible and more easily altered than hardware, it is more difficult to generate, control, and test. The problems associated with software generation are not of direct interest here and will not be discussed further. The problems associated with control can be summed in one word: documentation. All software and software changes must immediately be documented and appropriately reviewed before implementation. In addition, it is highly advisable for the buyer to retain control, and ownership, of the software at all stages of development.

The testing of software is commonly termed V and V: validation and verification. Validation is the process of demonstrating that the basic system equations and operational concepts are appropriate for the mission to be accomplished. Validation generally takes place early in software development and, therefore, utilizes a "scientific" language program, and a

general-purpose computer, to examine system behavior. Verification takes place only when the actual program has been written, and utilizes the final program tapes or other media, utilizing either the actual system computer or a general purpose computer programmed to have precisely the operating characteristics of the actual computer to be employed in the system. If a general purpose computer is used for verification, it must be programmed to have the same timing, input/output, word size, underflow, overflow, and other real-time characteristics of the actual computer. Such a computer simulation is called an interpretive computer simulation (ICS).

2.3 Electric and Magnetic Fields

2.3.1 The Bohr Model of the Atom -- In Figure 2.3.1.1 is shown the Bohr model of the (Lithium) atom. While this model should not be taken literally, it does adequately describe the electromagnetic phenomena with which we are here concerned. The Bohr model supposes matter to be composed of three kinds of particles: positively-charged protons, negatively-charged electrons, and neutrally-charged neutrons. The protons and neutrons are located in a small, heavy cluster, called the nucleus, at the center of the atom. The electrons are located in concentric shells around the nucleus, each shell containing a specific number of electrons. It is the nature of these "electrically-charged" particles that they create in the surrounding space, an "electric field." If the charges are in motion, they create, in addition, a "magnetic field." The "field" is a concept used to explain "action at a distance". That is, charged particles exert forces on other charged particles without apparent contact. It is the phenomena associated with these fields with which the discipline of electronics is concerned.

The electrons in shells completely filled with the prescribed number of electrons are tightly bound to the atom. Those in sparsely populated outer shells are loosely bound and can easily move from atom to atom. Materials which have loosely bound electrons are conductors of electrical current. Materials which have few, if any, loosely bound electrons are electrical

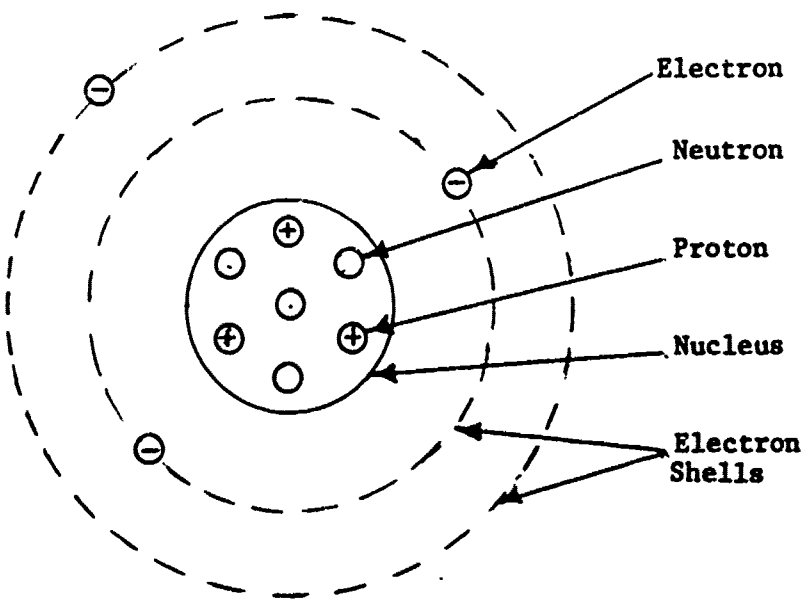


Figure 2.3.1.1 - - The Bohr Model of the Atom

insulators. Materials which have no loosely bound electrons but which have some electrons not too tightly bound are called semiconductors because, under sufficient external influence, electrons can escape from the atom and take part in electrical conduction. Transistors and semiconductor diodes are devices that control these" external influences" in such a way as to cause current to flow only as desired by the designer.

It is possible for a body to have too many, or too few, electrons for the available "holes" in its atomic shells. When that is the case, a net negative or positive charge, (respectively), exists on the body, with a resulting electric field in the space surrounding the body.

2.3.2 The Electric Field -- When a net charge, q, exists on a small body, (such as a particle or incremental portion of a large body), it creates a radial vector field at a point, p, in the vicinity of the body, given by:

$$|\vec{E}| = \mathcal{E}_r = \frac{q}{4\pi\epsilon r^2} \quad \text{(Volts/Meter)} \qquad [2.3.2.1]$$

where $\vec{E}$ is the electric field intensity, q is the charge in coulombs, ϵ is the electric permittivity of the medium in Farads per meter, and r is the radial distance from the particle to the point, p. The direction of the field is radially outward if the charge, q, is positive.

2.3.3 The Magnetic Field -- when a particle with charge q is in motion with velocity v, a magnetic field is created. The magnetic field intensity, $\vec{H}$, at a point p is given by:

$$\vec{H} = \frac{q}{4\pi r^3}(\vec{v} \times \vec{r}) \quad \text{(Amps/Meter)} \qquad [2.3.2.2]$$

where $\vec{v}$ is the velocity vector of the particle in meters per second, $\vec{r}$ is the radius vector from the particle to the point, p, in meters, and $\vec{v} \times \vec{r}$ is the vector or cross product.

2.3.4 Maxwell's Field Equations -- Fundamental to all electronic theory are the equations relating electric and magnetic fields to electric charges and currents. These equations, derived by Maxwell, can be written in integral form as follows:

$$(1) \quad \oint_{\ell} \mathcal{E}_{\ell}\, d\ell = -\mu \iint_{A} \left(\frac{dH_n}{dt}\right) dA \qquad [2.3.4.1]$$

$$(2) \quad \oint_{\ell} H_{\ell}\, d\ell = \sigma \iint_{A} \mathcal{E}_n\, dA + \epsilon \iint_{A} \left(\frac{d\mathcal{E}_n}{dt}\right) dA \qquad [2.3.4.2]$$

$$(3) \quad \epsilon \oint_{A} \mathcal{E}_n\, dA = \iiint_{V} \rho\, dV \qquad [2.3.4.3]$$

$$(4) \quad \iint_{A} H_n\, dA = 0 \qquad [2.3.4.4]$$

where E_1 is the component of $\vec{E}$ in the direction of the integration path, H_n is the normal component of $\vec{H}$, l is path length, A is surface area, μ is the magnetic permeability of the medium in henries per meter, σ is the conductivity of the medium in mhos per meter, and ρ is charge density. The integrations are indicated as being around closed contours.

Maxwell's equations can be stated as follows.

(1) The voltage induced in a closed-path of conductor (e.g. a single turn of transformer winding) is equal to the time rate-of-change of magnetic flux linking (passing through) that closed path.

(2) The line integral of magnetic field intensity around a closed path is equal to the total electric current linking that closed path.

(3) The total electric field flux emanating from a closed volume is equal to the total electric charge contained in that volume.

(4) The total magnetic flux emanating from a closed volume is zero. (Magnetic lines of force follow closed paths).

Employing Maxwell's equations, it is possible to derive the characteristics of the circuit elements (inductance of a coil, capacitance of a capacitor, etc.). It is also possible to predict the propagation characteristics of antennas and wave propagation phenomena.

2.4 Electromagnetic Waves

2.4.1 Schrodinger's Wave Equation -- When Maxwell's field equations, presented in Section 2.3.3, are applied to charge-free space and appropriately combined, there result, in cartesian coordinates, the following equations.

$$\frac{\partial^2 \vec{E}}{\partial x^2} + \frac{\partial^2 \vec{E}}{\partial y^2} + \frac{\partial^2 \vec{E}}{\partial z^2} - \mu\sigma \frac{\partial \vec{E}}{\partial t} - \mu\epsilon \frac{\partial^2 \vec{E}}{\partial t^2} = 0$$
$$\frac{\partial^2 \vec{H}}{\partial x^2} + \frac{\partial^2 \vec{H}}{\partial y^2} + \frac{\partial^2 \vec{H}}{\partial z^2} - \mu\sigma \frac{\partial \vec{H}}{\partial t} - \mu\epsilon \frac{\partial^2 \vec{H}}{\partial t^2} = 0 \qquad [2.4.1.1]$$

where the coefficients are as previously defined. These equations, known as Schrodinger's wave equations, predict that time-varying electric and magnetic fields interact in such a way as to propagate energy through space. That is, the solutions of Schrodinger's equations are traveling waves.

Typical solutions to Schrodinger's wave equations are of the form:

$$\mathcal{E} = \mathcal{E}_0 e^{\mp \alpha s} \cos(\omega t \mp \beta s)$$
$$H = H_0 e^{\mp \alpha s} \cos(\omega t \mp \beta s) \qquad [2.4.1.2]$$

where

$$\alpha = Re\left\{\left[-\mu\epsilon\omega^2 + i\mu\sigma\omega\right]^{1/2}\right\}$$
$$\beta = Im\left\{\left[-\mu\epsilon\omega^2 + i\mu\sigma\omega\right]^{1/2}\right\} \qquad [2.4.1.3]$$

Equations [2.4.1.2] represent a plane wave with initial amplitude E_o (and H_o), traveling in the direction of $\mp s$, with angular frequency w,

exponential decay constant α, and phase constant β. The velocity of propagation of the traveling wave is:

$$v = \omega/\beta \qquad [2.4.1.4]$$

The wavelength of the wave is:

$$\lambda = 2\pi/\beta \qquad [2.4.1.5]$$

It is important to note that, for a single propagating electromagnetic wave, both the electric field, $\vec{E}$, and the magnetic field, $\vec{H}$, exist. They do, in fact, generate each other and always exist in-phase electrically, but in space quadrature; that is, for a wave traveling in the +z direction, the x-component of $\vec{E}$, E_x, generates a y-component of $\vec{H}$, H_y, and vice versa. It is also important to note that, in general, solutions to Schrodinger's equations exist for $\vec{E}$ and $\vec{H}$ components in all possible directions perpendicular to the direction of travel and with all possible phases. If all such components exist, the wave is said to be unpolarized.

The complete solution for a plane wave traveling in a non-conducting medium ($\sigma = 0$), in the +z direction, and plane-polarized in the x direction, (see Figure 2.4.1.1), is:

$$\begin{aligned} \mathcal{E}_x &= \mathcal{E}_{x0}\cos(\omega t - 2\pi z/\lambda) \\ \mathcal{E}_y &= 0 \\ \mathcal{E}_z &= 0 \\ H_x &= 0 \\ H_y &= H_{y0}\cos(\omega t - 2\pi z/\lambda) \\ H_z &= 0 \end{aligned} \qquad [2.4.1.6]$$

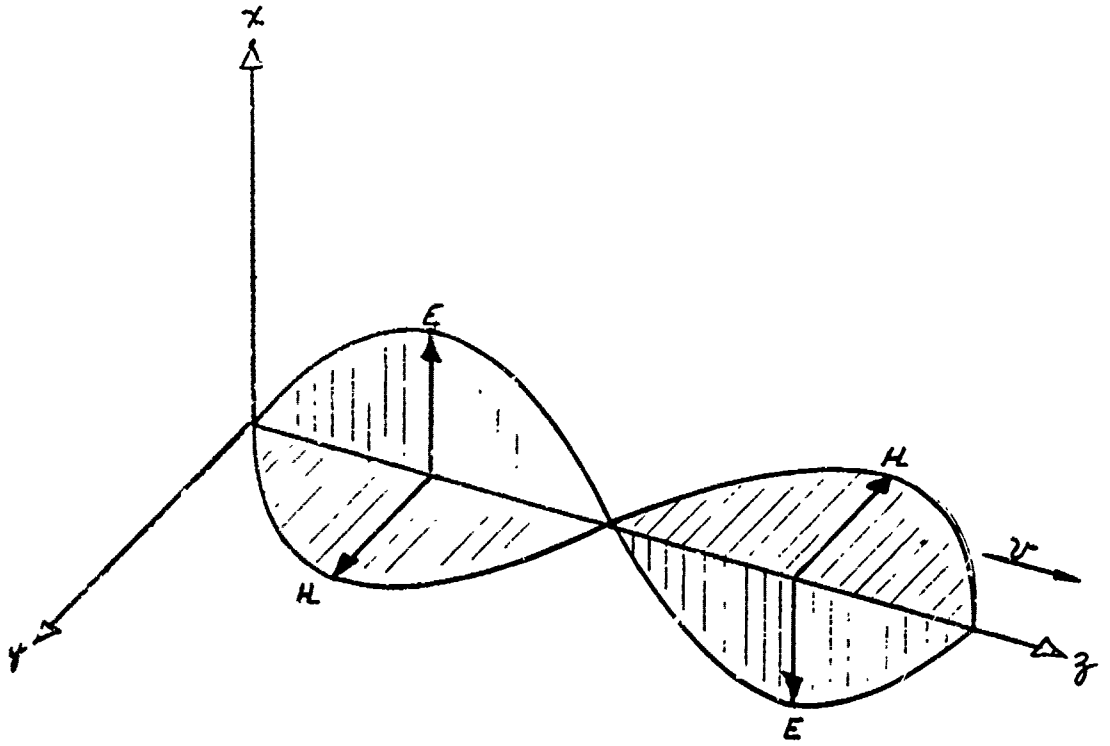


Figure 2.4.1.1 -- Plane Polarized Wave

The amplitude of the magnetic field, H_{yo}, is related to the amplitude of the electric field, E_{xo}, by the relation:

$$H_{yo} = \sqrt{\epsilon/\mu}\ E_{xo} \qquad [2.4.1.7]$$

The decay constant, α, phase constant, β, and velocity of propagation, v, are given by:

$$\alpha = 0$$
$$\beta = \sqrt{\mu\epsilon}\ \omega$$
$$v = 1/\sqrt{\mu\epsilon} \qquad [2.4.1.8]$$

If the medium were conducting, ($\sigma \neq 0$), the E and H amplitudes in Figure 2.4.1.1 would decay exponentially in the +z direction. Thus, the velocity of propagation of an electromagnetic wave is dependent upon the electric and magnetic properties of the medium. The frequency, f, wavelength, λ, and velocity, v, are related by the equation:

$$v = \lambda f \qquad [2.4.1.9]$$

The direction of polarization of an electromagnetic wave always designates the direction of the electric field. (The magnetic field is always understood to be there, and always at right angles in space to the electric field). A plane polarized wave (electric field constrained to a single

plane\ is shown in Figure 2.4.i.2(a), indicating only the electric field vector. As previously indicated, an unpolarized wave has components of electric field vector in all possible directions perpendicular to the direction of propagation and of all possible phases. The most frequently encountered type of polarization is plane polarization, already defined. Another type of polarization frequently encountered is circular polarization. A circularly polarized wave actually consists of two plane polarized waves, of equal amplitude, with their planes of polarization (E vectors) at right angles and also ninety degrees out of phase in time. The resultant E vector, as shown in Figure 2.4.1.2(b) rotates about the direction of propagation, its tip describing a helix as it propagates. An elliptically polarized wave, as shown in Figure 2.4.1.2(c), consists of two plane polarized waves, as for a circularly polarized wave, but with unequal amplitudes.

The phenomena associated with the propagation of electromagnetic waves are determined by the properties of the medium of propagation and by the frequency, or wavelength, of the waves. The reciprocal relationship between frequency and wavelength, ($\lambda = v/f$), is indicated in Figure 2.4.1.3. Also indicated in that figure are the terms commonly used to designate the various parts of the electromagnetic spectrum. The frequency designations, in ascending order are: ultra low, very low, low, medium, high, very high, ultra high, super high, and extremely high. The designated bands are bounded on the low end by zero frequency and on the high end by infrared radiation.

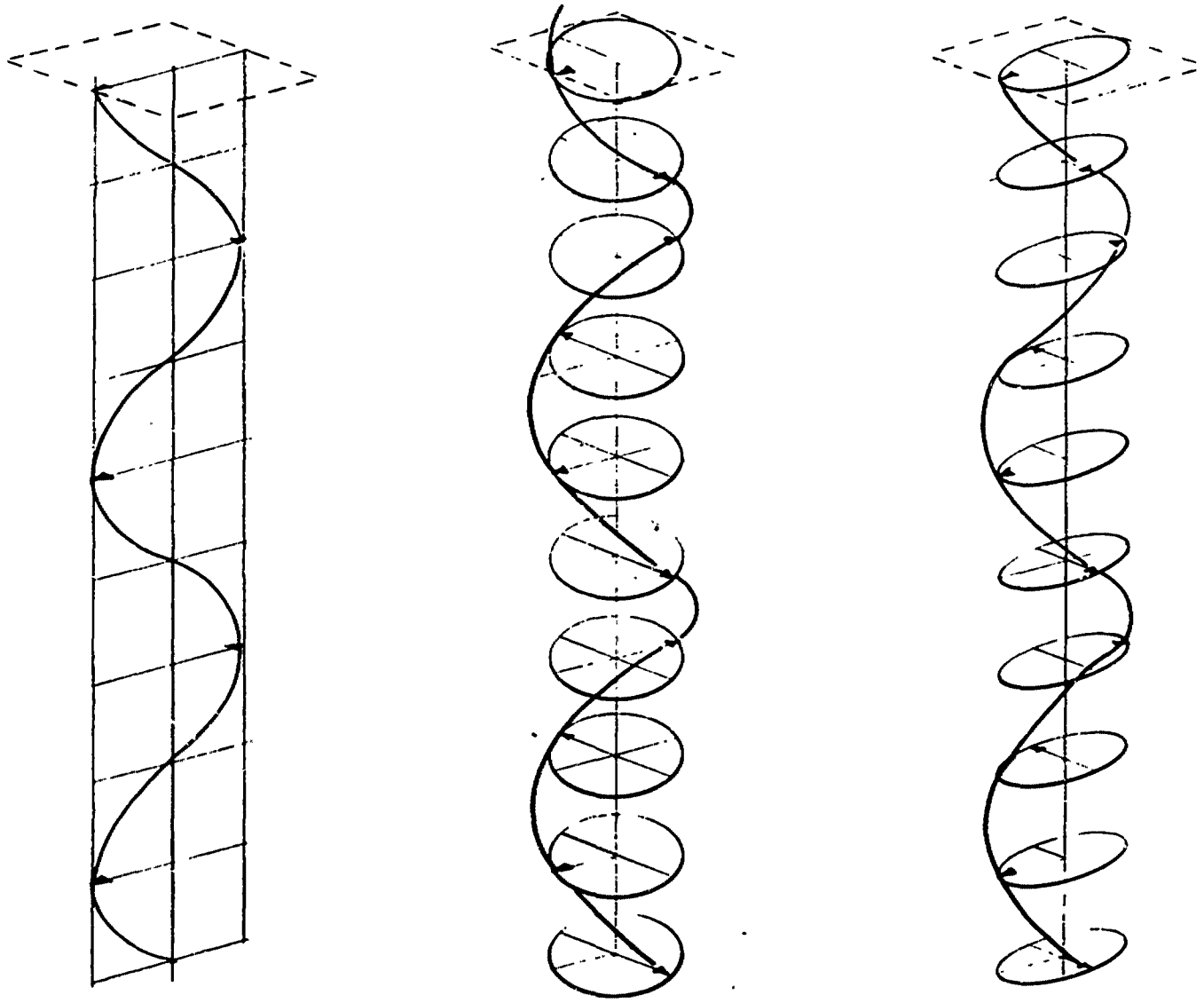


Figure 2.4.1.2 -- Polarized Electromagnetic Waves

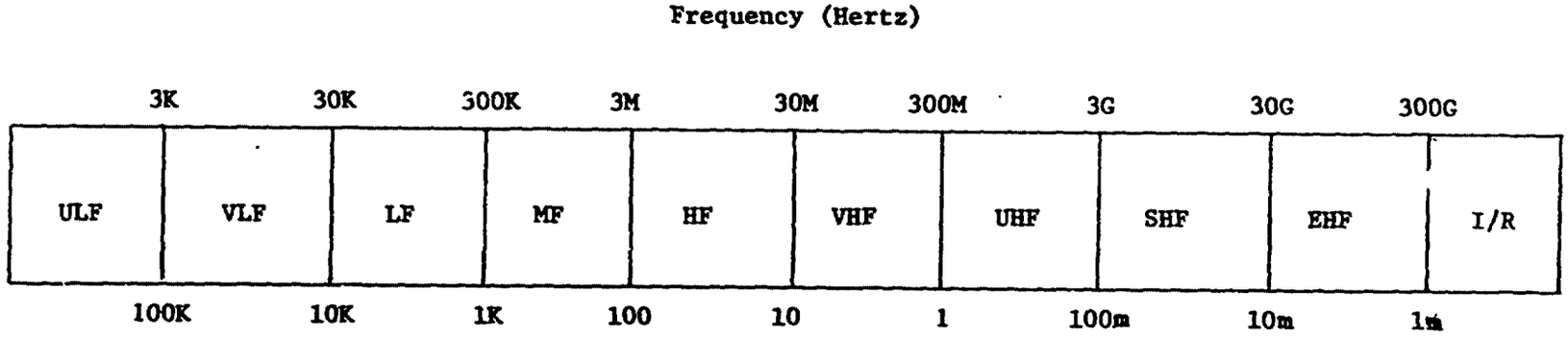


Figure 2.4.1.3 -- Electromagnetic Spectrum Frequency Designations

As previously indicated, the propagation of electromagnetic radiation is a function of the electric and magnetic properties of the medium. The most important of those properties are the electric permittivity, ε, the magnetic permeability, μ, and the electric conductivity, σ. The electric permittivity is a measure of the energy stored in a material by an electric field of intensity, E. The magnetic permeability is a measure of the energy stored in a material by a magnetic field of intensity, H. The electric conductivity is a measure of the electric current produced in a material by a given electric field of intensity, E. For a good conductor, the conductivity, σ, is very high, the electric permittivity, ε, is high, and the magnetic permeability, μ, is variable. For a dielectric, the conductivity is very low, the electric permittivity is variable, and the magnetic permeability is low. Earth, water, and ice, the most frequently encountered materials in nature, exhibit the properties of both conductors and dielectrics, the dominant characteristics depending upon the frequency of the radiation.

2.4.2 Wave Propagation Phenomena -- There are six basic phenomena associated with electromagnetic wave propagation. They are free-space attenuation, absorption, reflection, refraction, diffraction, and interference. The combined effects of these phenomena determine the direction of propagation, the amplitude, and the phase of a propagating wave.

Free space attenuation is a result of the geometric spreading of the wave energy as it propagates outward and is unrelated to energy loss or

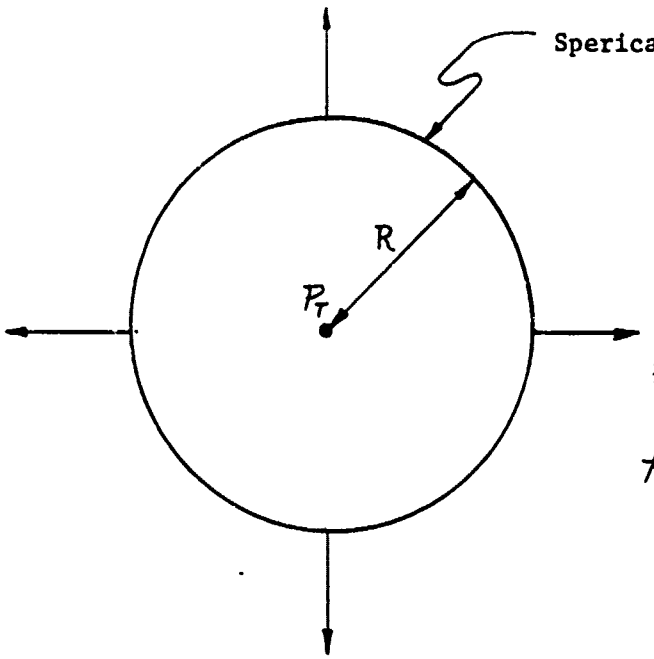
dissipation due to absorption. When there is no provision for directing the energy in a preferred direction, the energy from an electromagnetic wave source spreads spherically outward, as shown in Figure 2.4.2.1(a). If the power at the source is P_T (Watts), the power density at a range R (Meters) is given by:

$$p_R = \frac{P_T}{4\pi R^2} \quad \text{(Watts/Meter}^2\text{)} \qquad [2.4.2.1]$$

That is, the power, P_T, has spread out over a spherical surface area, $4\pi R^2$. Even if the source has directive gain, (that is, directs the energy in a preferred direction), the power density still falls off as the square of the range from the source, as shown in Figure 2.4.2.1(b). This reduction in power density with (the square of) range is always present in free-space wave propagation and, as previously indicated, is independent of any antenna directive gain or attenuation due to energy absorption.

Absorption represents an actual loss (dissipation) of energy to the medium of propagation. Due to absorption, the field intensity of the wave falls off exponentially with distance traveled as indicated in Equation [2.4.1.2]. (Notice that Equation [2.4.1.2] is the equation for a plane wave, rather than a spherical wave, and therefore does not include a $1/R^2$ factor to account for spherical spreading). The exponential decay coefficient, α, depends upon the effective conductivity of the medium and is a function of frequency. Even the atmosphere introduces serious absorption at higher frequencies.

(a) Non-Directive (Isotropic) Source



(b) Directive Source

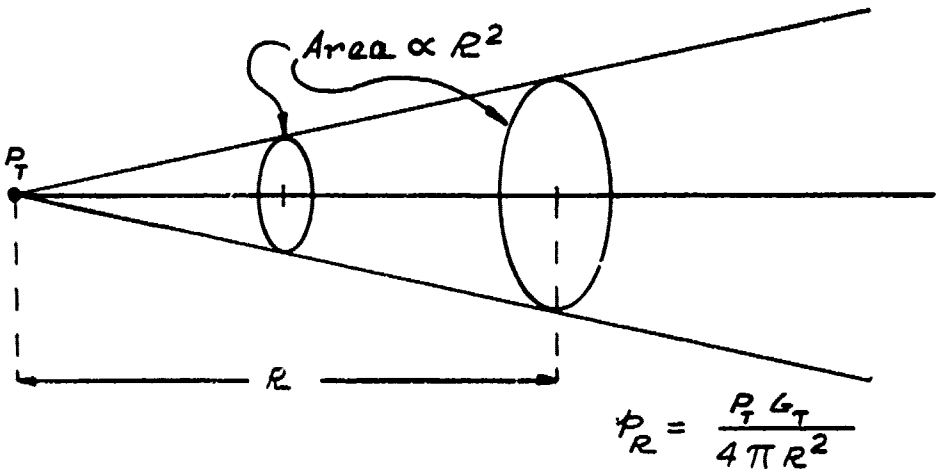


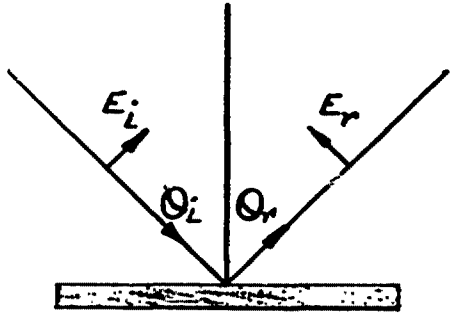
Figure 2.4.2.1 -- Geometric Spreading of Wave Energy

Reflection always occurs when an electromagnetic wave encounters an interface between two media with unlike electrical properties. The degree to which the wave is reflected at the interface depends upon the electrical properties of the media and upon the frequency, polarization, and angle of incidence of the wave.

When an electromagnetic wave is reflected by the interface between a dielectric material and a perfect conductor, as shown in Figure 2.4.2.2, all of the energy in the wave is reflected. That is, the reflection coefficient is unity. The effect of reflection upon the phase of the wave is a function of the polarization of the wave. The phase is unaffected for components of the wave with vertical polarization (E vector constrained to a plane perpendicular to the reflecting surface as shown in Figure 2.4.2.2(a)). The phase is reversed, however, for horizontally-polarized components (E vector constrained to a plane parallel to the reflecting surface), as shown in Figure 2.4.2.2(b). In any case, the propagation path always obeys the basic law of reflection: the angle of reflection, θ_r, is equal to the angle of incidence, θ_i, as indicated in the figure.

When an electromagnetic wave is reflected by the interface between two dielectric materials of different dielectric constants, as shown in Figure 2.4.2.3(a), part of the energy is reflected and part enters the second medium. The reflection coefficient and phase reversal depend upon the angle of incidence as shown in Figures 2.4.2.3(b) and (c), respectively. For grazing incidence, ($\theta i = 90°$), the reflection

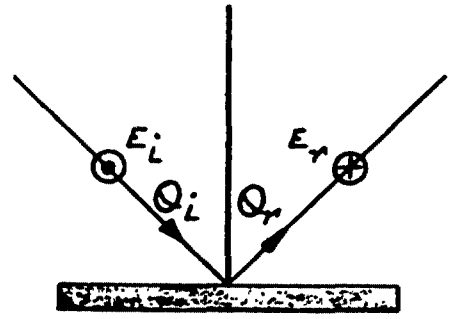
(a) Vertically Polarized Wave



$\theta_r = \theta_i$

No Phase Reversal

(b) Horizontal Polarized Wave



$\theta_r = \theta_i$

Phase Reversal

Figure 2.4.2.2 -- Reflection of an Electromagnetic Wave by the Interface Between a Dielectric and a Perfect Conductor

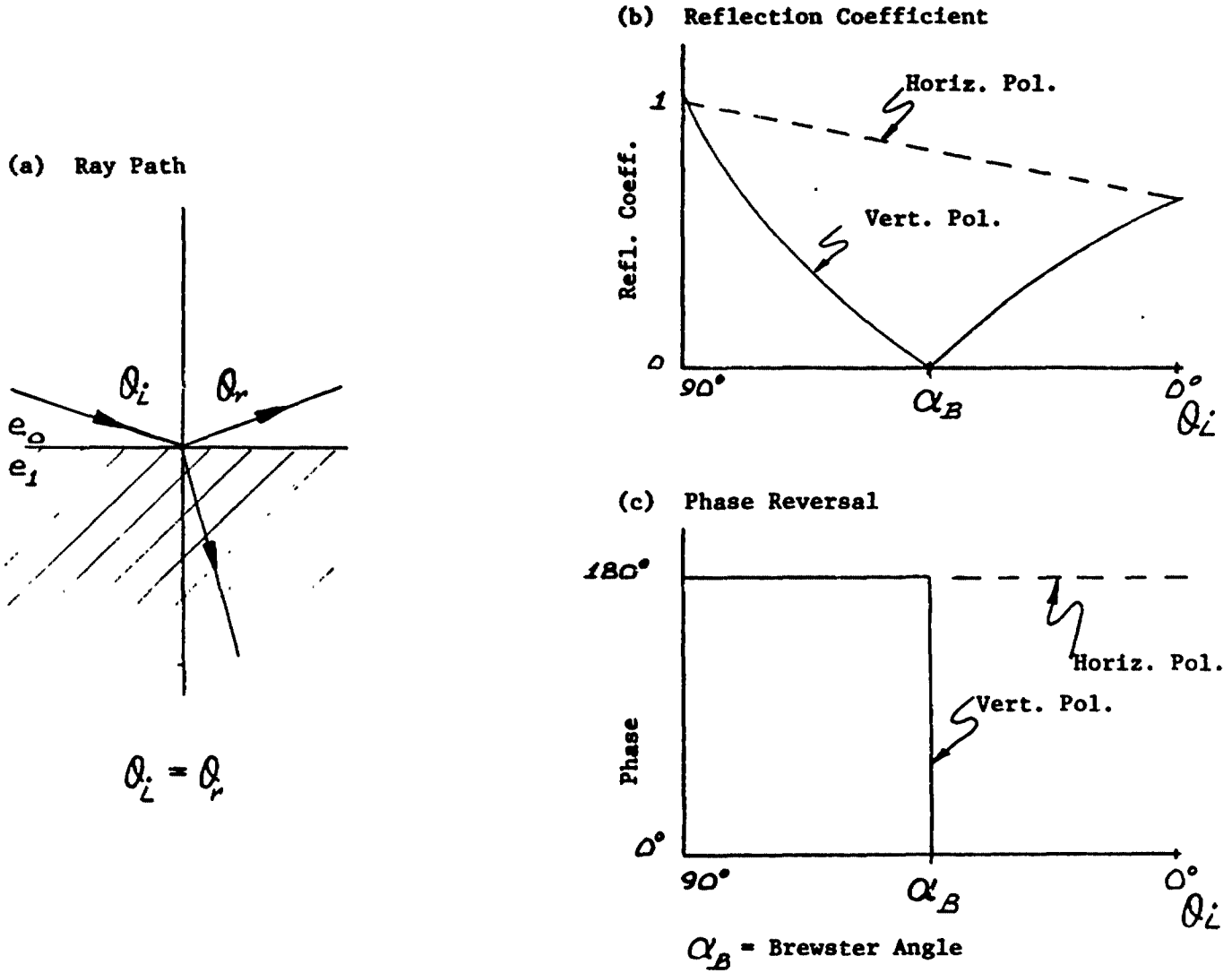
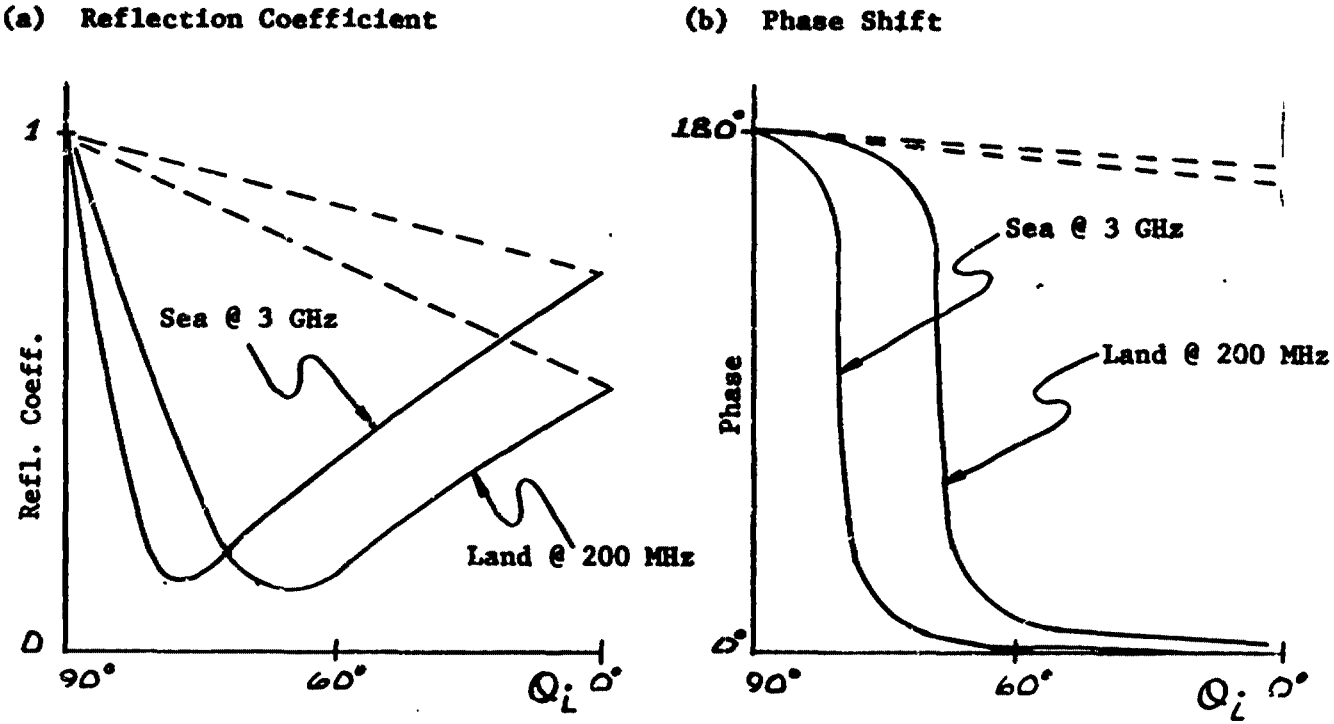


Figure 2.4.2.3 -- Reflection of an Electromagnetic Wave by the Interface Between Two Dielectrics

coefficient is unity. It decreases with decreasing angle of incidence until, at a unique angle, known as the Brewster Angle, it becomes zero. With further decrease of θ_i, it increases monotonically to a value less than unity at normal incidence ($\theta_i = 0$). The reflection coefficient for the horizontally-polarized component of the wave decreases monotonically with decreasing angle of incidence, from a value of unity at $\theta_i = 90°$ to a smaller value, characteristics of the material, at $\theta_i = 0°$.

For angles of incidence greater than the Brewster Angle, the phase of the vertically-polarized component of a wave is reversed upon reflection from the surface of a dielectric. For angles less than the Brewster Angle, the phase is unaffected. The phase of the horizontally-polarized component of a wave is always reversed upon reflection. In any case, of course, the path of the reflected ray obeys the basic law of reflection ($\theta_r = \theta_i$).

When an electromagnetic wave is reflected by the interface between the air and the land or sea, the reflection characteristics are strongly dependent upon the frequency of the radiation. For frequencies below about one megahertz for land and one gigahertz for water, the characteristics of both media approximate those of a good conductor. For frequencies above these stated, their characteristics approximate those of a dielectric. In Figure 2.4.2.4, the reflection coefficient and phase shift characteristics are shown versus angle of incidence. At the frequency illustrated, 200 megahertz for land and 3 gigahertz for sea, the characteristics shown can be seen to resemble those of a dielectric.



Solid Line = Vertically - Polarized Wave
Dashed Line= Horizontally - Polarized Wave

Figure 2.4.2.4 -- Reflection of an Electromagnetic Wave by the Interface Between Air and Land or Sea

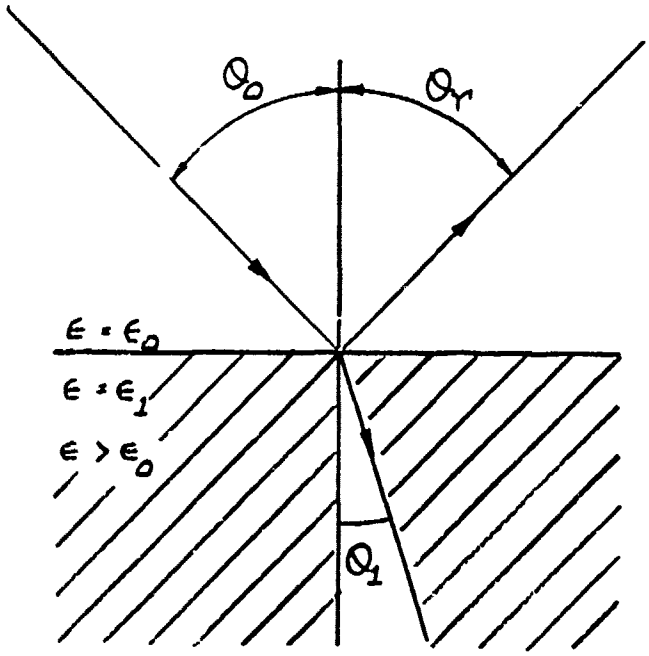
Refraction is the change in direction of propagation of a wave as it crosses the interface between two media. As previously indicated, when a propagating electromagnetic wave encounters an interface between two media with different electrical properties, part of the wave energy is reflected back into the first medium and part enters the second medium. The reflection coefficient depends upon the dielectric coefficients of the media, the frequency of the radiation, and the angle of incidence. The portion of the wave entering the second medium is refracted, as shown in Figure 2.4.2.5. The manner in which the wave is refracted by the interface depends, as does that for reflection, upon the electrical properties of the two media and upon the frequency, polarization, and angle of incidence of the wave. When the second medium is a good conductor, such as a metal, very little energy enters that medium. For that reason, we shall be concerned mainly with refraction of a wave at the interface between two dielectrics, as illustrated in Figure 2.4.2.5. The refracted ray obeys the basic law of refraction, known as Snell's Law, given by:

$$\frac{\sin\theta_1}{\sin\theta_0} = \sqrt{\frac{\epsilon_1}{\epsilon_0}} \qquad [2.4.2.2]$$

For small angles, a first-order approximation is often assumed for Snell's Law. That is, for small angles:

$$\frac{\theta_1}{\theta_0} = \sqrt{\frac{\epsilon_1}{\epsilon_0}} \qquad [2.4.2.3]$$

It should be noted that the electric permittivities, ϵ_0 and ϵ_1, are generally



Snell's Law

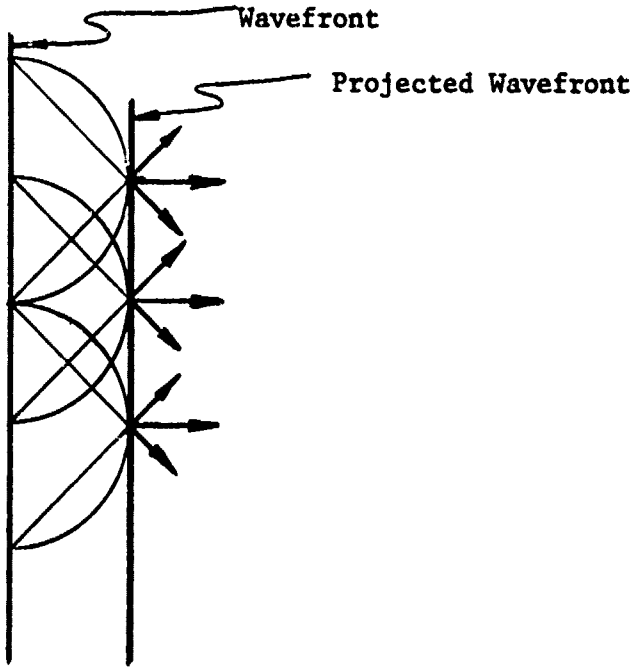
$$\frac{\sin \theta_1}{\sin \theta_0} = \sqrt{\frac{\epsilon_1}{\epsilon_0}}$$

Figure 2.4.2.5 — Refraction of an Electromagnetic Wave by the Interface Between Two Media

functions of frequency. Thus, waves of different frequency will be refracted at different angles. The permittivities can also be a function of polarization; that is, one value for vertically-polarized components and another value for horizontally-polarized components. A material with such properties is said to be birefringent.

Diffraction is the phenomenon whereby the direction of propagation of a wave is "bent" around an obstructing object. It is entirely unrelated to reflection or refraction and is best understood by applying Huygen's Principle. Huygen's Principle states that every point on the wave front of a propagating wave is, itself, a point source of spherically propagating waves. As illustrated in Figure 2.4.2.6, the contributions from the multitude of "point sources" on a broad wavefront add in such a way that all non-normal contributions to the projected wavefront cancel. The result is a projected wavefront moving in a direction perpendicular to the original wavefront, a result in agreement with reality. When the wavefront encounters a wall with a pinhole, however, the "point sources" to either side of the hole are obscured by the wall. Thus, their non-normal contributions to the projected wavefront do not cancel those from the sources just in front of the hole, thereby causing the wave to propagate "around the corner". As a result of diffraction, the "shadow" of an obstructing object is never entirely "sharp". The diffraction effects are, however, restricted to the portions of the wavefront near the edge of the obstruction (where there are no adjacent point sources to cancel the non-normal contributions). The term "near", in this case, means

(a) Broad Wavefront



(b) Wavefront at Wall with Pinhole

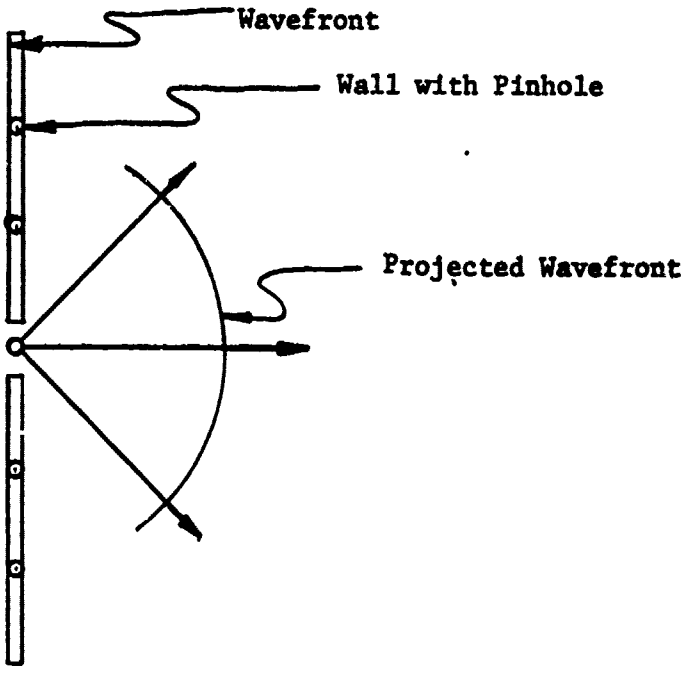


Figure 2.4.2.6 Huygen's Principle Applied to Wavefront Projection

"at a distance not large with respect to the wavelength of the wave." This restriction means that the effect of an obstructing object, (or aperture), on the propagation of a wave is a function of the relative size of the object and the wavelength of the wave. If the wavelength of the radiation falling on an object is large compared to the size of the object, the effect on the radiation field will be significant only in the immediate vicinity of the object. The reflected energy will be small and scattered, and the "shadow" behind the object will be "filled in" at points well beyond the object. On the other hand, if the wavelength is small in comparison with the size of the object, the reflection will be large and the "shadow" will be "sharp".

Interference is the most complex, and hence most difficult to predict, phenomenon associated with electromagnetic wave propagation. Electromagnetic waves are, in fact, oscillatory electromagnetic fields. It is to be expected that two such fields, if equal in amplitude and opposite in phase, will cancel. This phenomenon is known as destructive wave interference. Two such waves (fields), if in-phase, will, of course, add, creating constructive wave interference. Note that, in order to consistently add or subtract, sinusoidal waveforms must maintain their phase relationship. This means that they must be of exactly the same frequency. Thus, two electromagnetic waves can interfere, (constructively or destructively), only if they are coherent (have the same frequency).

A demonstration of wave interference is the arrangement shown in Figure 2.4.2.7. In this arrangement, a monochromatic point source is placed behind a screen with two slits as shown. In accordance with Huygen's Principle, the two slits, irradiated by the same source, and hence of exactly the same frequency, become coherent line sources of radiation. If another screen is placed in front of the first, the radiation from the two slits will illuminate the second screen as shown. At the center of the screen, the waves from the two slits will arrive in phase, having traveled exactly the same distance. The waves will, therefore, add producing a bright line down the center of the screen. At some distance off the centerline, however, the lengths of the paths traveled by the radiation from the two slits will be different by exactly one-half wavelength, producing waves exactly 180° out of phase at that point on the screen. There, the waves will cancel, producing a dark line to either side of the centerline as shown in the figure. Alternating bright and dark lines will be repeated at regular intervals from the center of the screen. This pattern, due to wave interference, is called a diffraction pattern because the second screen is illuminated by diffracted radiation.

2.4.3 Principal Factors Limiting Communication System Range -- The four major factors limiting the maximum useful range of a communication system are: free-space attenuation, land mass and atmospheric absorption, multipath interference, and background noise.

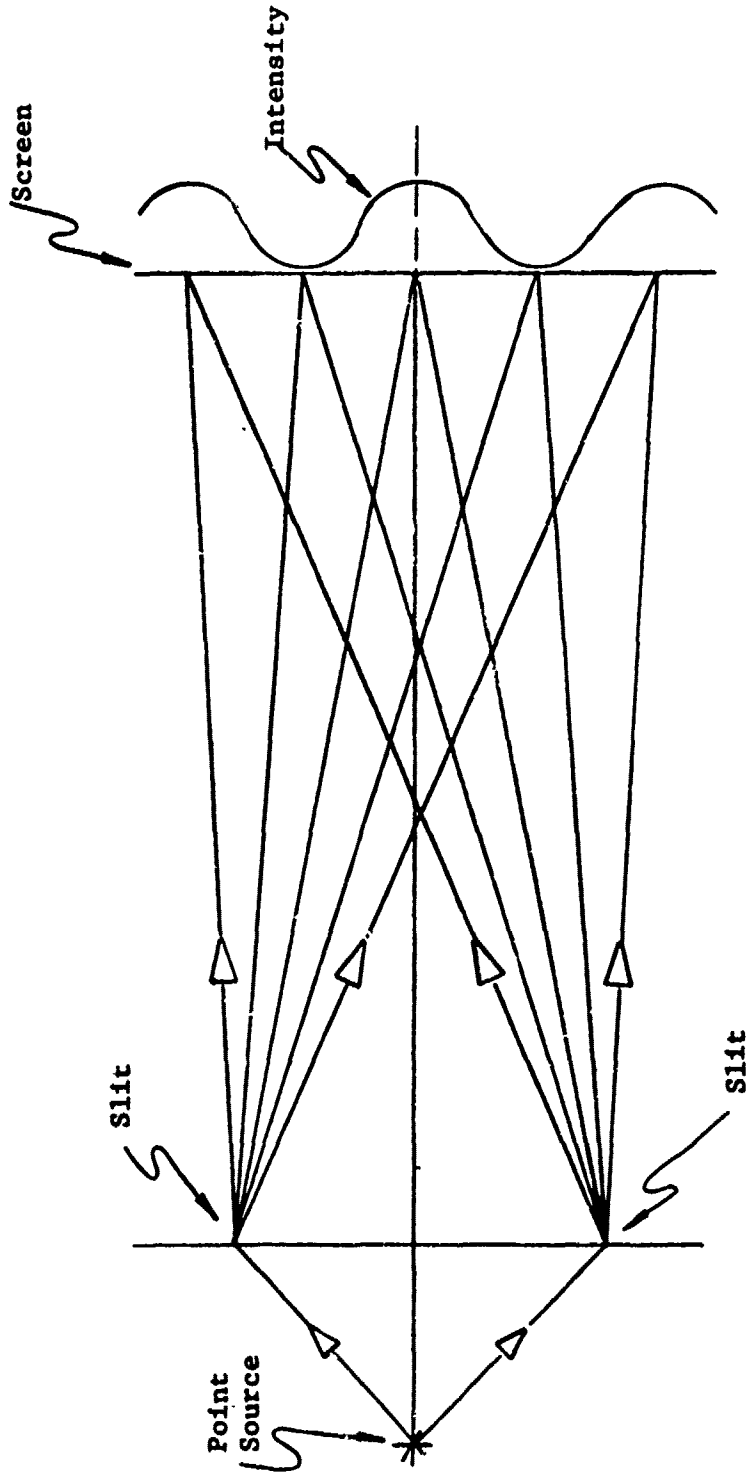


Figure 2.4.2.7 -- Interference Between Waves from Two Coherent Sources

Due to free-space attenuation (spherical spreading, discussed in Section 2.4.2), the signal power received at range R (Meters) from a transmitter emitting power P_T (watts) is given by:

$$P_R = \frac{P_T G_T A_R}{4\pi R^2} \qquad [2.4.3.1]$$

where G_T is the transmitting antenna directive gain (N.D.)[1.] and A_R is the capture area of the receiving antenna (Meters^2).

Equation [2.4.3.1], known as the communications Equation, considers only signal attenuation due to free-space attenuation. Absorption of the signal by the medium of propagation can be a more important cause of signal loss than free-space attenuation. This is especially true if the propagation path consists, in part, of land mass. (Land mass here includes earth, water, ice, and vegetation). Typical curves of signal strength versus distance, for a wave at the surface of the earth, are shown in Figure 2.4.3.1. The curves clearly indicate an increase in attenuation (decrease in signal strength) at higher frequencies. Atmospheric absorption becomes important at high signal frequencies. In Figure 2.4.3.2(a) are shown typical curves of signal attenuation, due to atmospheric gases, as a function of wavelength. The peaks are caused by resonances of tri-atomic molecules, as indicated. The curves shown in Figure 2.4.3.2(b) are typical for attenuation due to absorption by particulate moisture.

1. N.D. = Non-dimensional.

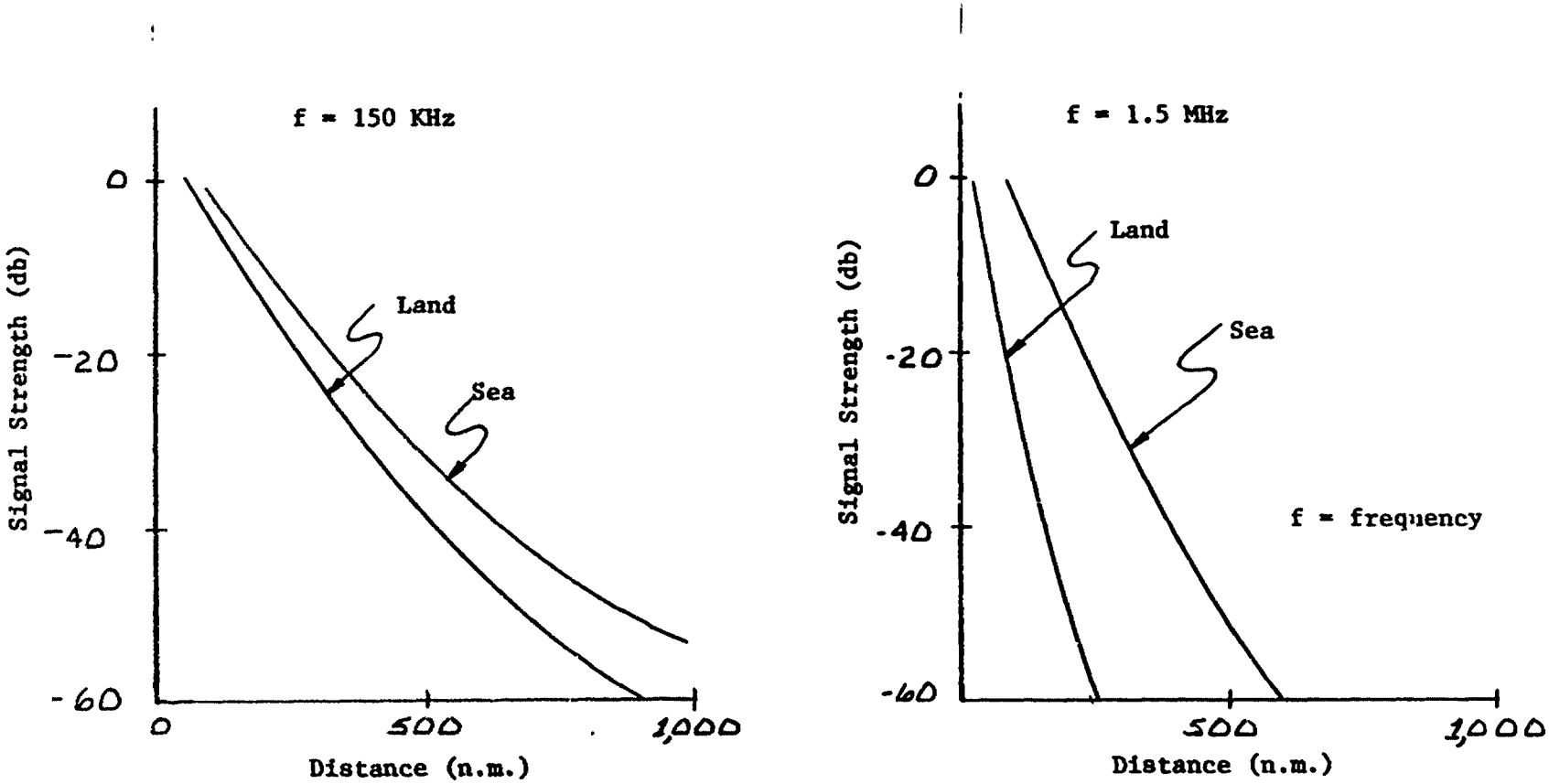


Figure 2.4.3.1 -- Attenuation of Signal Strength by Surface Waves Due to Absorption by Land and Sea Mass.

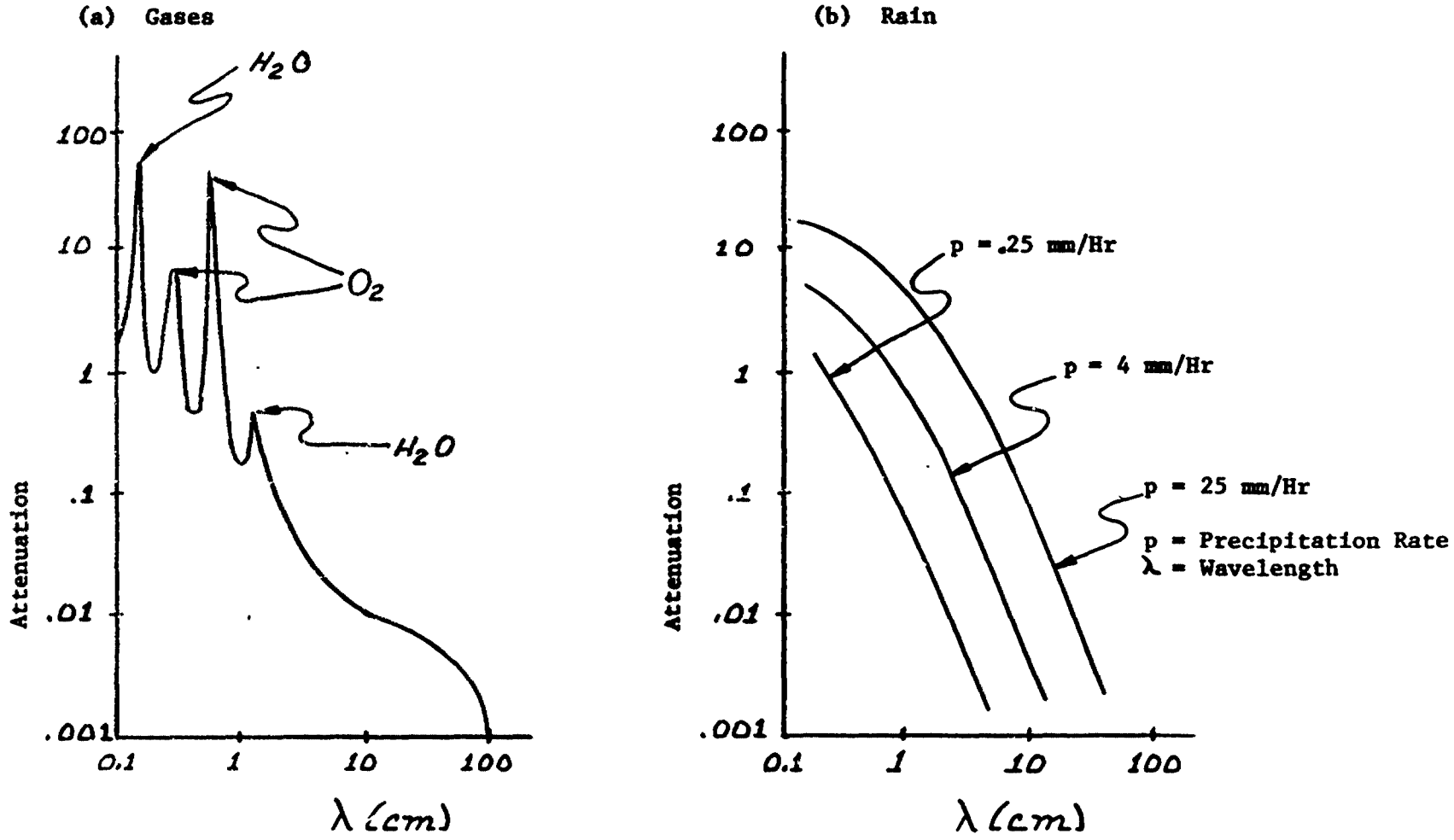


Figure 2.4.3.2 -- Attenuation of Signal Strength of Electromagnetic Waves Due to Absorption by the Atmosphere

Electromagnetic wave interference is discussed in Section 2.4.2 of this text. The phenomenon of interference plays a significant, often a dominant, role in determining the signal strength at the receiver in a communications system. The major sources of interference are reflections from the land mass, from man-made structures and vehicles, from the atmosphere, and from the ionosphere. The phenomenon of interference due to multiple signal paths is called multipath. Several such paths are illustrated in Figure 2.4.3.3. In that figure, the paths involved are a direct, unreflected path, one reflected from the ground, one reflected from the ionosphere, and one reflected from an aircraft. Due to the differences in path length, the signals may add or subtract. In the case of the aircraft reflection, the phase is continuously changing, resulting in alternately constructive and destructive interference, a phenomenon familiar to television viewers. Perhaps the most important instance of multipath interference is that between the direct path and the earth-reflected path for a transmitting antenna close to the ground. The result is a complex radiation (antenna) pattern which makes extremely difficult the prediction of communication system useful range. In Figure 2.4.3.4(a) is shown a typical radiation pattern for an antenna mounted just above the surface of the earth. The figure is a polar plot of signal strength, the distance from the origin to the curve representing signal strength in that direction. The size, shape, and number of "petals" in the pattern is a function of the distance of the antenna above the earth. In Figure 2.4.3.4(b) is a plot of signal strength as a function of distance from the antenna. Note the alternate constructive and destructive

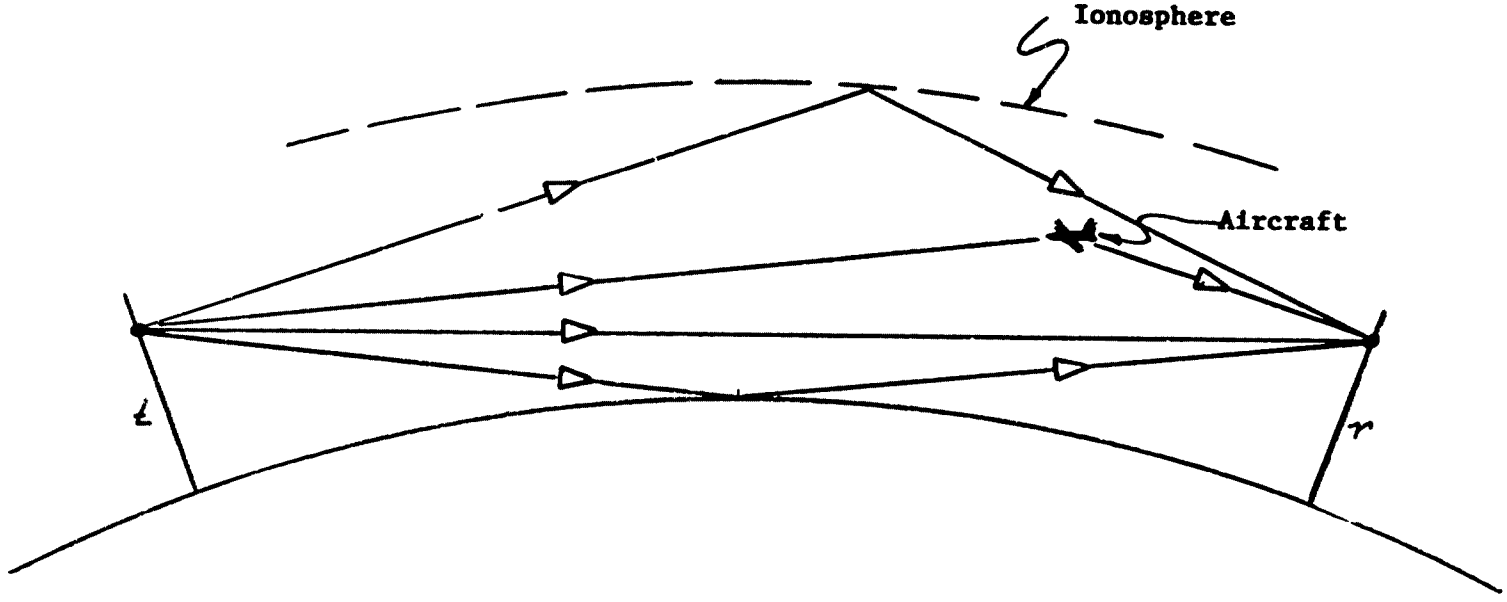


Figure 2.4.3.3 -- Multipath Interference

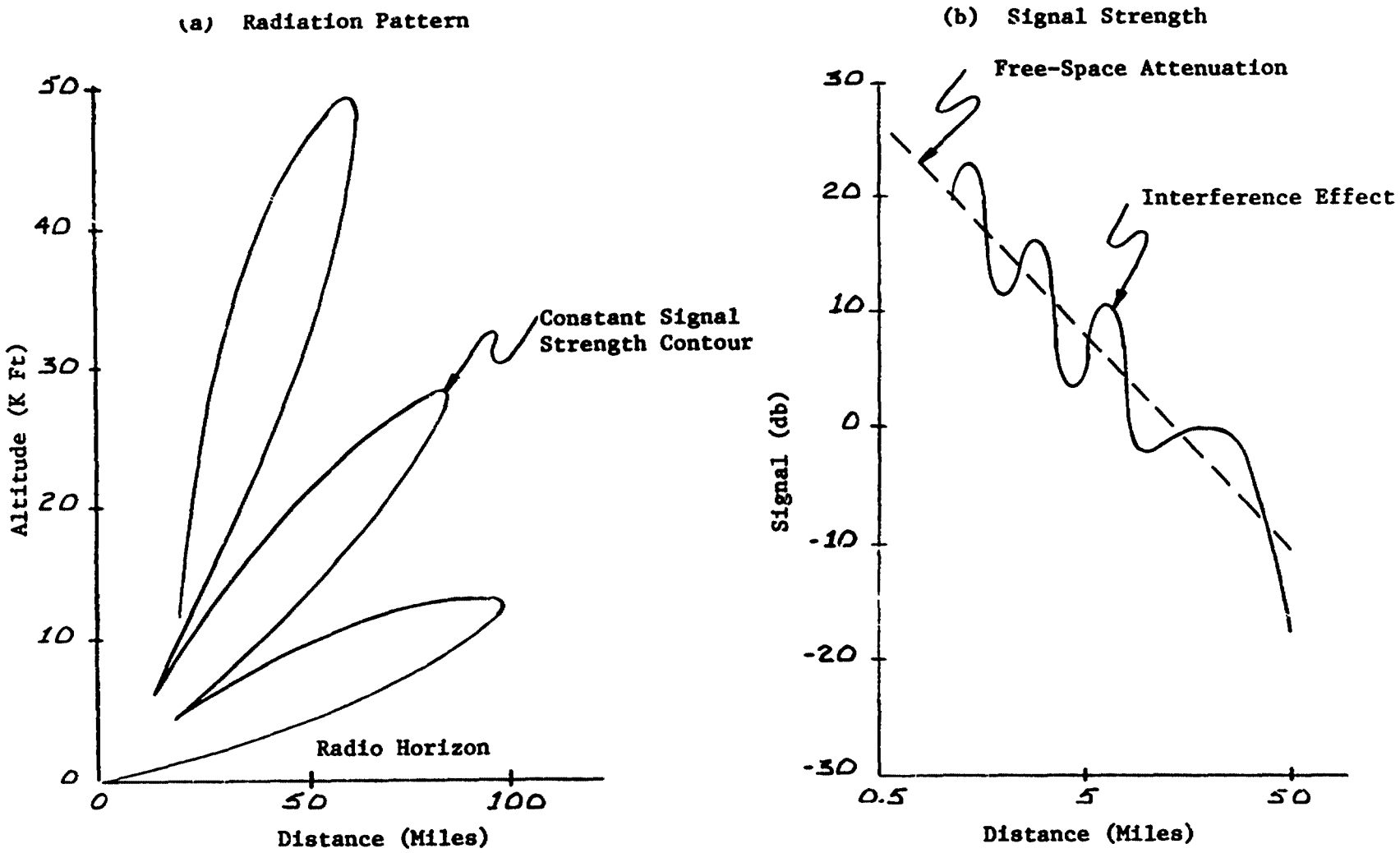


Figure 2.4.3.4 -- Multipath Interference due to Ground Reflection

interference superimposed upon the free-space attenuation.

In practice, it is not the received signal strength that determines the performance of a communications system; it is the signal-to-noise ratio.. Modern technology is such that it is generally possible to build an amplifier with more gain. The problem is that the noise is amplified along with the signal. Thus, it is, ultimately, the signal-to-noise ratio that is important. There are significant noise sources both internal to the system and external. It is the external sources with which we are here concerned since it is those sources that affect the choice of propagation modes and operating frequencies. The major external noise sources are atmospherics (principally lightning), man-made equipment, and galactic (principally solar) phenomena. Plots of typical noise signal strength versus frequency are shown in Figure 2.4.3.5. Since the magnitudes of these noise signals depend upon many factors, including location and time of day, the absolute signal strengths are not important. The two important features of the plots are: (1) the fact that they all decrease in amplitude with increasing frequency, and (2) the fact that atmospheric noise dominates at low frequencies and man-made noise dominates at high frequencies. There are, of course, special conditions under which other sources of noise are dominant. An important source of this nature is the extremely large-amplitude electromagnetic pulse due to a nuclear explosion. The main problem with respect to such a pulse is not preservation of signal-to-noise ratio but prevention of burn-out of the receiving apparatus.

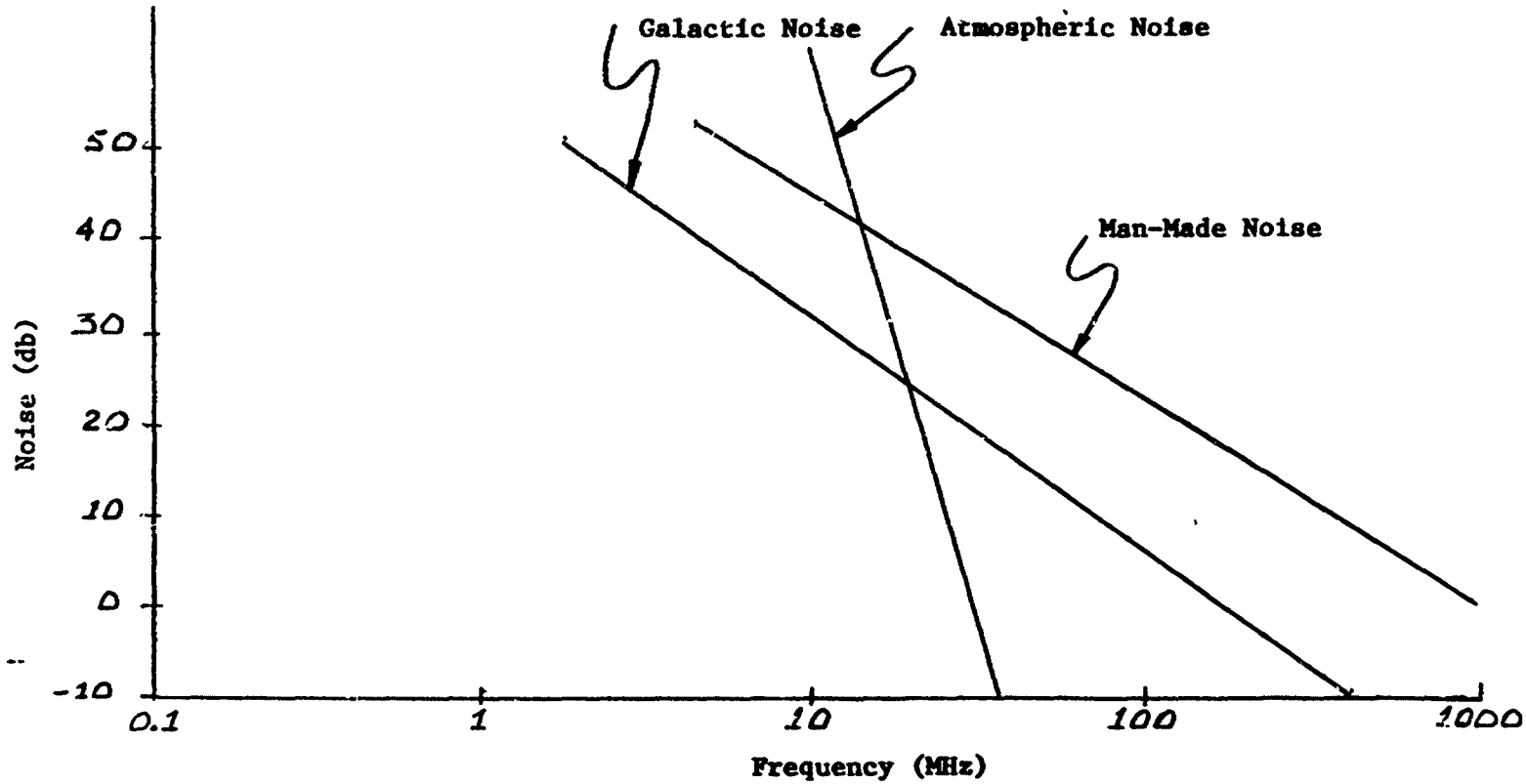
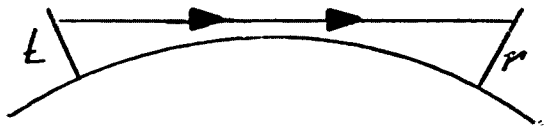


Figure 2.4.3.5 -- Electromagnetic Noise Source Spectra

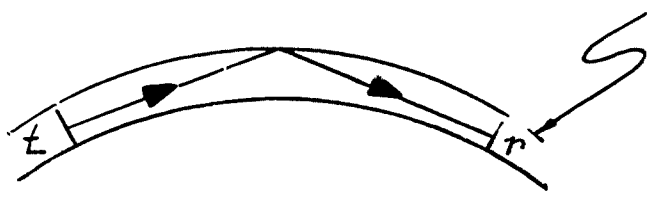
2.4.4 Electromagnetic Wave Propagation Modes -- There are four basic modes of terrestrial electromagnetic wave propagation: space wave, ground wave, sky wave, and ducted wave. For given conditions, all four modes will be involved, the amount of energy propagated by any one mode being determined by a number of factors. These factors include signal frequency, range, transmitter and receiver location, terrain, transmitting and receiving antennas, atmospheric conditions, and ionospheric conditions. The principal determinants of the dominant mode of propagation are the frequency and the range between transmitter and receiver.

At frequencies above 100 megahertz, the dominant mode of propagation is space wave (direct, line-of-sight propagation); as illustrated in Figure 2.4.4.1(a). No intentional reflection or refraction is involved, only free-space propagation. Space wave propagation offers reliable communication, over short distances, with relatively low power. The low power requirement results partly from the low noise prevalent at these frequencies and partly from the efficient antenna designs possible at these frequencies. The main disadvantage is the restriction to direct line-of-sight communication. Such a system is not only restricted by the radio horizon (farthest point visible from the transmitting antenna), but also by objects such as buildings and, at the higher frequencies, even vegetation. Atmospheric absorption is also a problem, at the higher frequencies, under conditions of high atmospheric moisture content.

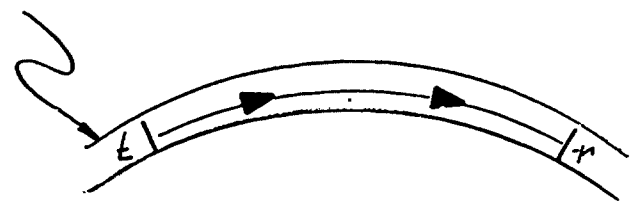


(a) Space Wave

(b) Ground Wave



(c) Sky Wave



(d) Duct Prop.

Figure 2.4.4.1 -- Principal Modes of Electromagnetic Wave Propagation

At frequencies below approximately 100 kilohertz, the dominant mode of propagation is ground or surface wave, as illustrated in Figure 2.4.4.1(b). At these low frequencies, an electromagnetic wave tends to propagate along the interface between two media such as the air and the earth. For that reason, the ground wave is able to travel beyond the radio horizon, thus providing long-range communication with a reliability much greater than that of the alternate methods of long-range communication, (excluding those methods, such as satellite communication, which employ repeater stations). The principal disadvantages of ground-wave communication are the high power required and, for airborne applications, the large antennas required. The large required power is a result of two factors: the high background noise at low frequencies and the large attenuation due to propagation partly in the land mass. The requirement for a large antenna is explained in Section 2.6 of this text. For some (high data rate) applications, the relatively small bandwidth of systems employing low-frequency carriers is another significant disadvantage. (Refer to Section 2.5 of this text for an explanation of the bandwidth limitation).

At intermediate frequencies (between approximately 1 megahertz and 100 megahertz), it is possible to "bounce" an electromagnetic wave off of the ionosphere or Heavyside Layer as shown in Figure 2.4.4.1(c) and 2.4.4.2. The term "bounce" is used because the phenomenon is the result of bending or refraction rather than reflection. The limitation of this mode of communication to a frequency band is a result of the fact that, at higher frequencies, the waves escape through the ionosphere into outer space; and, at lower frequencies, the waves are greatly attenuated by absorption

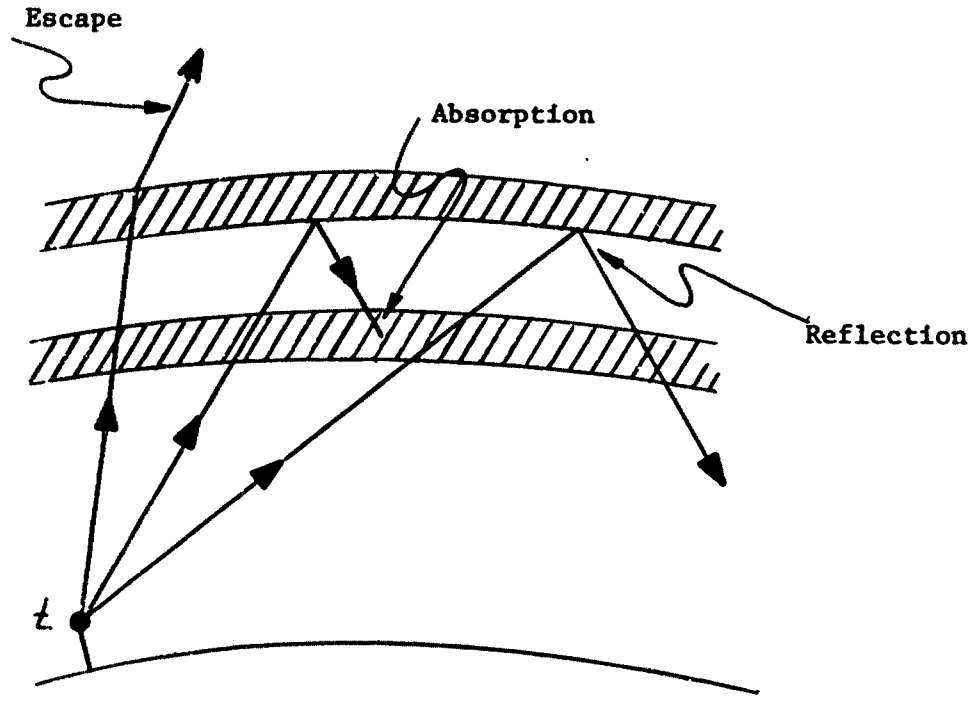


Figure 2.4.4.2 -- Ionospheric Effects on Electromagnetic Wave Propagation

in the ionosphere. The waves will also escape into space if directed at too steep an angle, thus limiting short-range communication. The principal advantage of sky-wave propagation is the ability to communicate beyond the radio horizon without the large expenditure of power involved in ground-wave propagation, (or the use of repeaters). The principal disadvantage is the difficulty of achieving reliable operation for any period of time, due to fluctuations in the ionosphere.

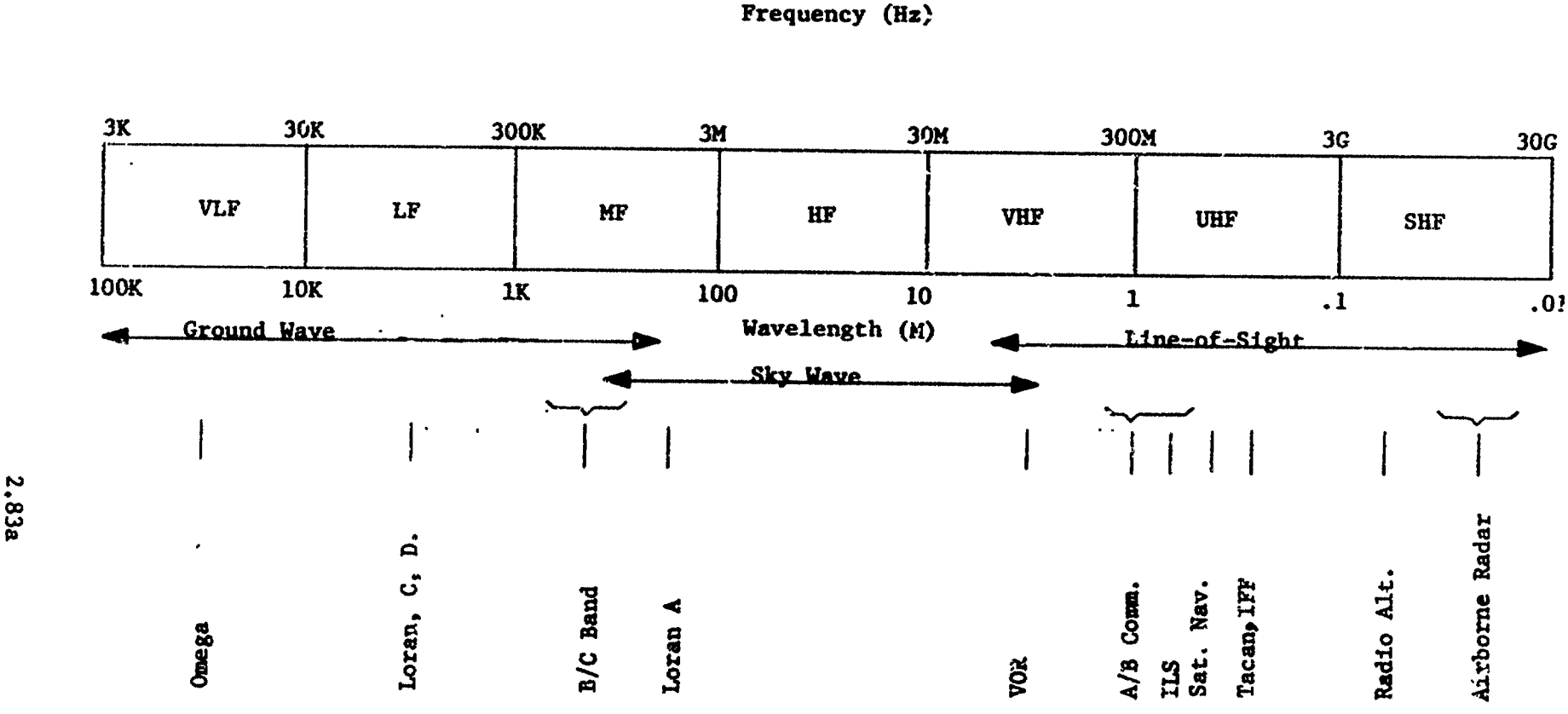
Under certain conditions the index of refraction of the atmosphere varys with altitude in such a way as to bend propagation electromagnetic waves downward, as shown in Figure 2.4.4.1(d). Under such conditions, the radio horizon can be greatly extended. This propagation mode is rarely utilized because of the unreliability of continuing communication. Search radars sometimes make use of this phenomenon to extend useful range.

The effect of various parameters on the propagation of electromagnetic waves is summarized in Figure 2.4.4.3. Presented in this chart is the information pertinent to the selection of propagation mode and frequency for a given communication requirement.

In Figure 2.4.4.4 is depicted the utilization of the radio frequency electromagnetic spectrum by some typical airborne systems. In general, ground waves at low frequencies are utilized when reliable coverage is required, sky waves at intermediate frequencies are utilized when long-range coverage at low power is required, and space waves at high frequencies are utilized when short range coverage at low power is required.

FREQ. (MHz)	BAND	PROPAGATION MODE	ADVANTAGES	DISADVANTAGES
0.015 to 0.030	VLF	Ground Wave	Reliability	High power required. Large antennas. Atmospheric noise. Small bandwidth.
0.030 to 3.0	LF & MF	Ground Wave Day; Sky Wave Night	Smaller antennas than at VLF. More reliability than at HF.	More absorption than at VLF. Unreliable at long range. Atmospheric noise.
3.0 to 30	HF	Sky Wave	Long Range. Low Power.	Ionospheric variations. More absorption than at LF. Man made noise.
30 to 300	VHF	Line-of-Sight with some sky wave	Reliability at short range with lower power. Less Atmospheric noise. Directive antennas.	Ground reflections. Man-made noise. Unreliable with sky wave.
300 to 3000	UHF	Line-of-Sight with some duct propagation	Low noise.	Multipath interference. Atmospheric absorption. Line-of-sight limit.
3000 to 30000	SHF	Line-of-Sight	Small, directive antennas. Low power required.	High absorption. Multipath interference. Receiver noise. Line-of-sight limit.

Figure 2.4.4.3 -- Electromagnetic Wave Propagation Characteristics



2.83a

Figure 2.4.4.4 — Utilization of Radio Frequency Spectrum

2.5 Signal Processing

2.5.1 Time Domain/Frequency Domain Duality -- Fundamental to an understanding of signal processing is the concept of the duality between the time domain and the frequency domain. The principle of duality can be stated as follows.

> Any time-varying quantity can be described in the frequency domain (as a function of frequency) as well as in the time domain (as a function of time).

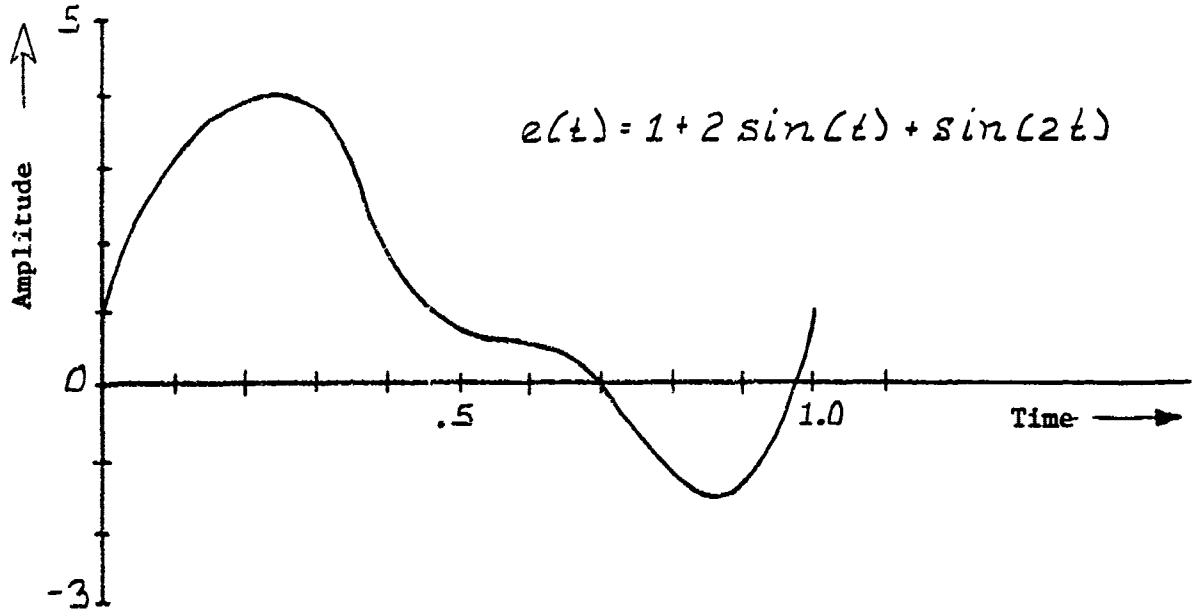
For example, the function

$$e(t) = 1 + 2\sin(t) + \sin(2t) \qquad [2.5.1.1]$$

can be described by the time plot shown in Figure 2.5.1.1(a). It can, equally well, be described by the frequency (spectral) plot of Figure 2.5.1.1(b). The time plot shows the magnitude of the function, e(t), as a function of time. The frequency plot (spectrum) shows the magnitudes of the frequency components as a function of frequency. (For a function containing only discrete frequencies, such as this one, the spectrum consists of impulse (or delta) functions, as shown here.)

The frequency-domain description illustrated in Figure 2.5.1.1 can be applied to general time functions by employing a Fourier series expansion of the time function. That is, an arbitrary time function can be represented, to any desired accuracy and over any desired time interval, by a series of purely sinusoidal terms.

(a) Time Domain Representation of e(t)



(b) Frequency Domain Representation of e(t)

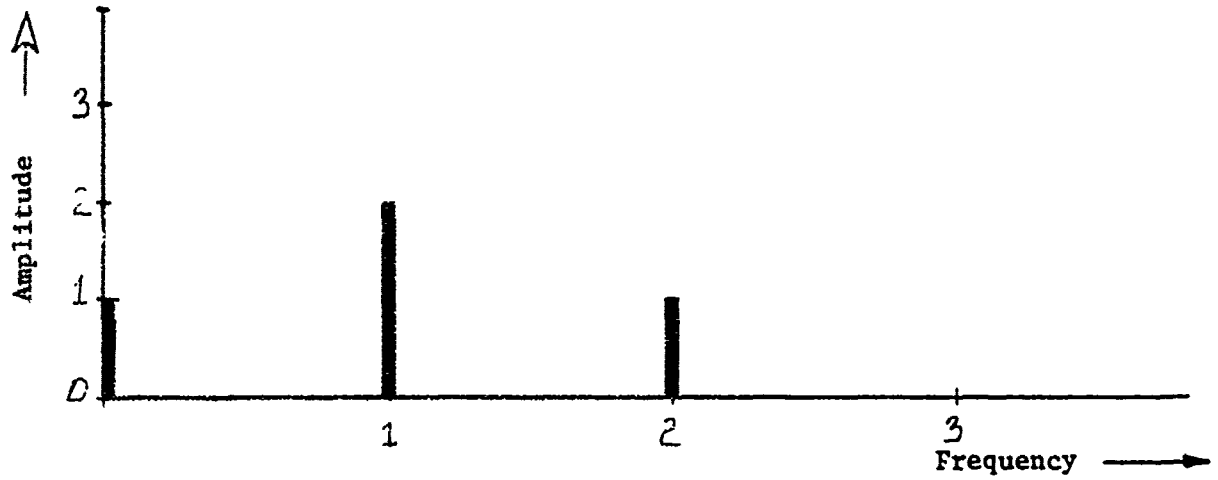


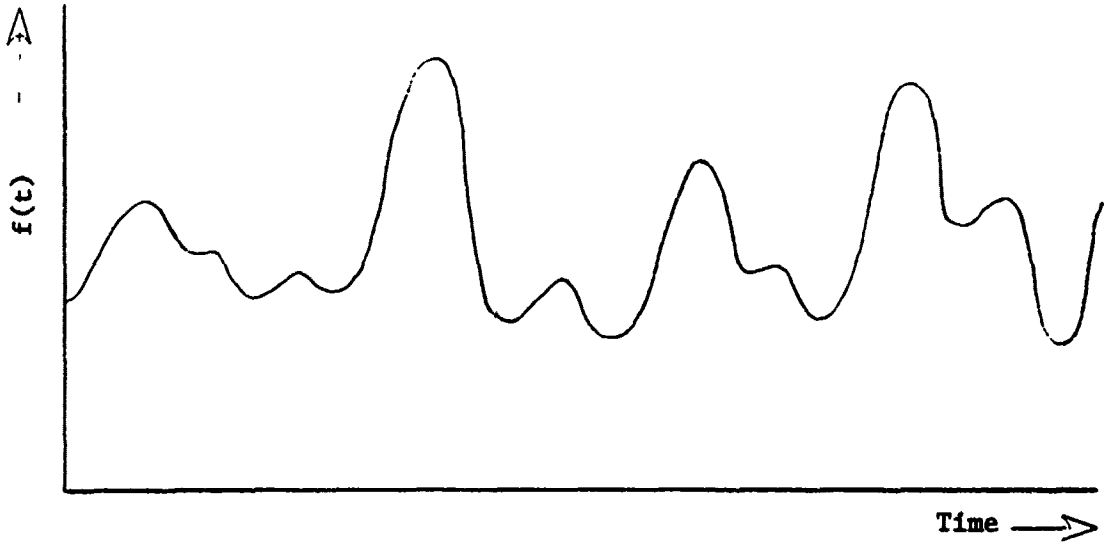
Figure 2.5.1.1 — Time and Frequency Domain Representations of a Time-Varying Quantity

In order to represent an arbitrary time function perfectly, or over all time, would require, in general, an infinite number of terms in the series. Fortunately, few practical applications require an impractical number of terms. When an impractical number of discrete-frequency terms would be required, the technique can be extended by assuming an infinite number of terms, replacing the Fourier series sum by an integral, and representing the frequency spectrum by a continuous function (continuum) rather than a series of impulses. Thus, the spectrum of the time function shown in Figure 2.5.1.2(a) might be that shown in Figure 2.5.1.2(b).

As previously indicated, the design and/or analysis of a system often requires the application of the principle of time/frequency duality. The spectral distribution of the energy in a signal, (spectrum), is often more significant than are the temporal characteristics (waveform), of that signal. Furthermore, it is often more convenient to process the signal in the frequency domain, (e.g. by means of a band-pass frequency filter) than in the time domain (e.g. by means of a time-varying gain).

When the system is linear, the response of the system to a complex waveform can be determined by: (1) breaking the input waveform down into its Fourier components, (2) altering the amplitude and phase of each component according to the frequency response characteristics of the system, and (3) re-combining the altered components to obtain the overall response. The frequency response characteristics of a system describe the system in the frequency domain, specifying the amplitude attenuation and phase shift,

(a) Arbitrary Time Function f(t)



(b) Spectrum of f(t)

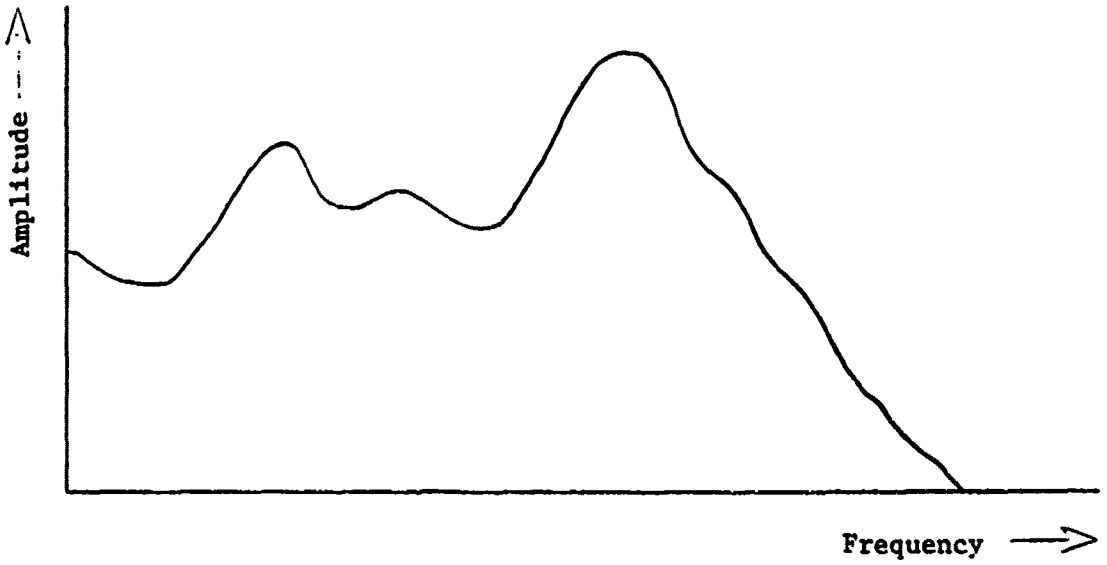


Figure 2.5.1.2 -- Time and Frequency Domain Representations of a Non-Discrete Frequency Time-Varying Quantity

imposed by the system, as functions of frequency. An important corollary of the principle of time/frequency duality is:

> The amplitude and phase characteristics of a system are not independent but are uniquely related. Consequently, it is impossible to alter the amplitude of a signal as a function of frequency without altering its phase.

The analysis of a time-varying system is often vastly simplified by the use of Fourier and Laplace transforms. These transforms, based upon time/frequency duality, allow the analyst to employ algebra rather than calculus to determine the response of a system to a given input. In general, the output of a linear, time-varying system is determined, in the time domain, by a convolution integral. That is, the output, $y(t)$, is related to the input, $x(t)$, by the integral:

$$y(t) = \int_0^t x(\tau)\, g(t-\tau)\, d\tau \qquad [2.5.1.2]$$

where the Green's Function, $g(t - \tau)$, represents the characteristics of the system. The convolution integral defined in Equation [2.5.1.2] is integrable only for a limited number of known functions. It is impossible to integrate in general. If, however, the Fourier Transform is employed, the output (Fourier Transform), $Y(w)$, is related to the input (Fourier Transform), $X(w)$, by the equation:

$$Y(\omega) = G(\omega)\, X(\omega) \qquad [2.5.1.3]$$

That is, $Y(w)$, is obtained simply by multiplying $X(w)$ by the Transfer Function of the system, $G(w)$. In system analysis, block diagrams are

often employed, based upon the Fourier or Laplace Transform, as shown in Figure 2.5.1.3. In such a block diagram, the output of a block is obtained from the input by multiplying the input by the transfer function of the block.

2.5.2 Linear and Nonlinear Systems -- As previously indicated, the use of operational transforms and block diagrams is valid only for linear systems; that is, systems for which the Principle of Linear Superposition applies. This principle can be stated as follows.

> For linear systems, the response of the system to the sum of two or more simultaneous inputs is equal to the sum of the individual responses to the inputs applied one at a time.

Linear operations of interest are addition, subtraction, time integration, time differentiation, amplification, and frequency filtering. Nonlinear operations of interest are multiplication, division, raising to a power, exponentiation, limiting, rectification, modulation, and demodulation.

In the time domain, system linearity decouples the responses to simultaneous inputs. An important consequence of system linearity, in the frequency domain, is:

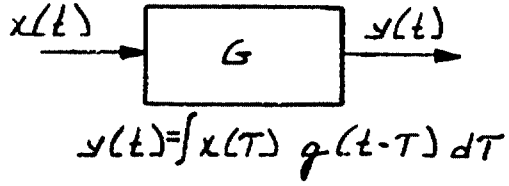
> The steady-state output of a linear system will exhibit no frequency components not present in the inputs.

conversely:

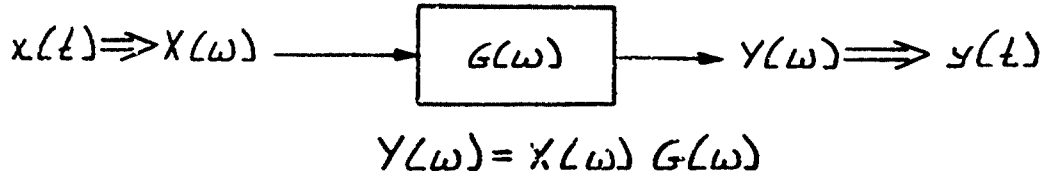
> A non-linear system (or operation), in general, creates frequencies in the output not present in the input.

2.5.3 Spectral Filtering -- Spectral (frequency) filtering is the selective alteration of the amplitude and phase characteristics of a signal as a

(a) Convolution Internal (Time Domain)



(b) Fourier Transform (Frequency Domain)



(c) Laplace Transform (Frequency Domain)

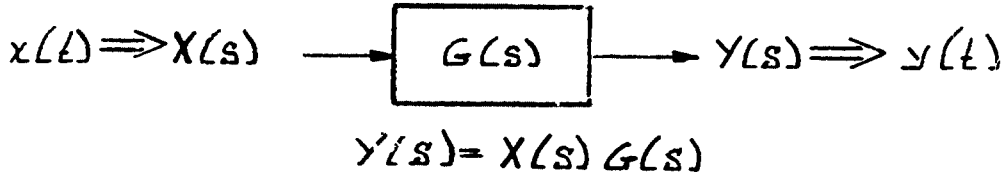


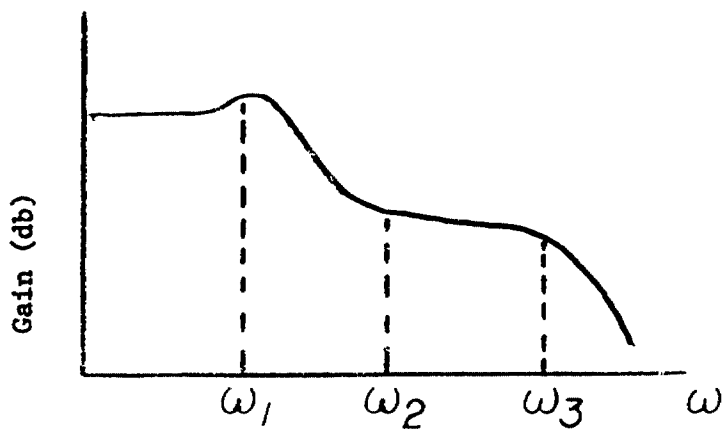
Figure 2.5.1.3 -- Operational Transforms and Block Diagrams

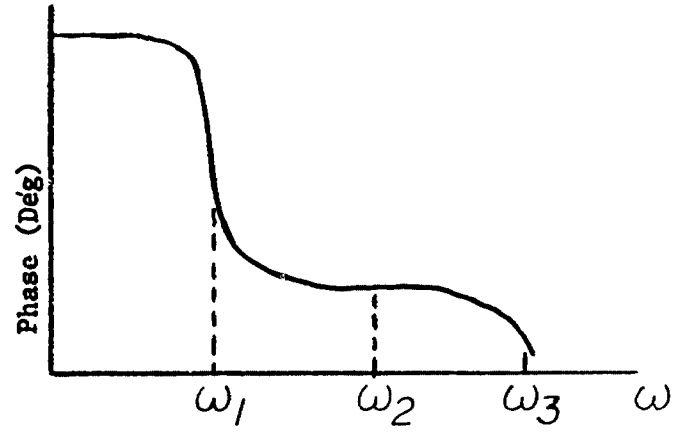
function of frequency. Frequency filtering is a linear operation and, therefore, introduces no new frequency components in the signal. There are numerous applications (and designs) for frequency filters, some intended primarily to alter signal amplitude and some to alter phase. Any frequency-selective alteration of one will, of course, affect the other.

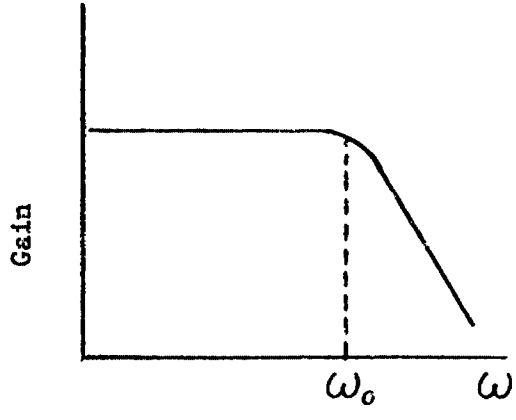
The filters of principal interest here are the low-pass, high-pass, and band-pass filters. In Figure 2.5.3.1(a) are shown the amplitude and phase-shift characteristics of a generalized filter with"break" frequencies w_1, w_2, amd w_3. Note that a change in amplitude characteristics is accompanied by a change in phase characteristics. The amplitude characteristics of low-pass, high-pass, and band-pass filters are shown in Figures 2.5.3.1(b), (c), and (d), respectively. The low-pass filter attenuates signal components with frequencies above the cut-off frequency, w_o. The high-pass filter attenuates signal components with frequencies below the cut-off frequency, w_o. The band-pass filter attenuates signal components with frequencies below the low cut-off frequency, w_L, and above the high cut-off frequency, w_H.

In an analog system, spectral filters can be constructed utilizing circuit components with frequency-sensitive impedances, such as capacitors and inductors. Simple, passive low-pass, high-pass, and band-pass filter circuits are shown in Figures 2.5.3.2 (a), (b), and (c), respectively. In a digital system, (which processes discrete data), spectral filters can be constructed using samplers and delay lines. The delay-canceller used in a digital moving-target indicator radar is an example of a digital high-pass filter.

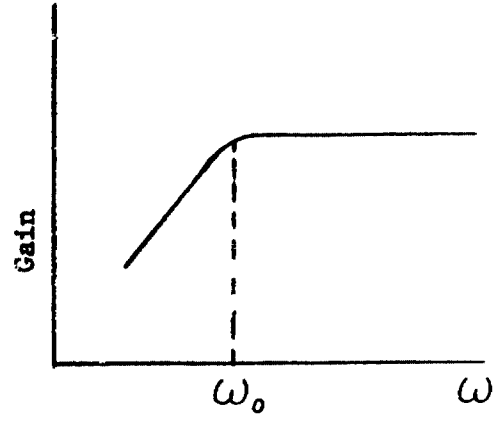
(a) General Filter



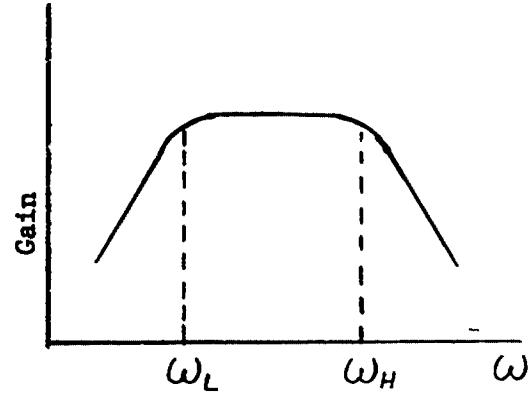




(b) Low-Pass Filter



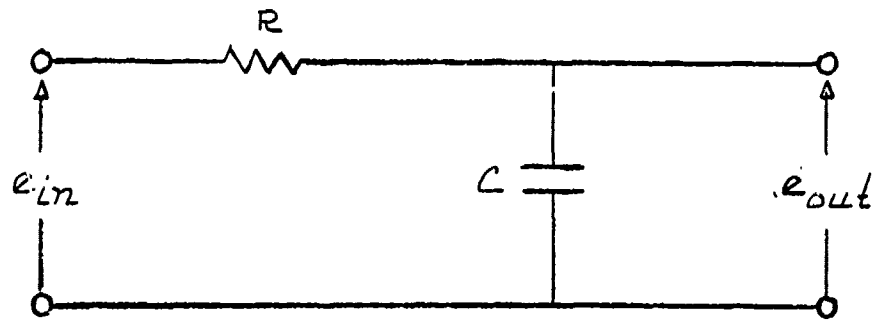
(c) High-Pass Filter



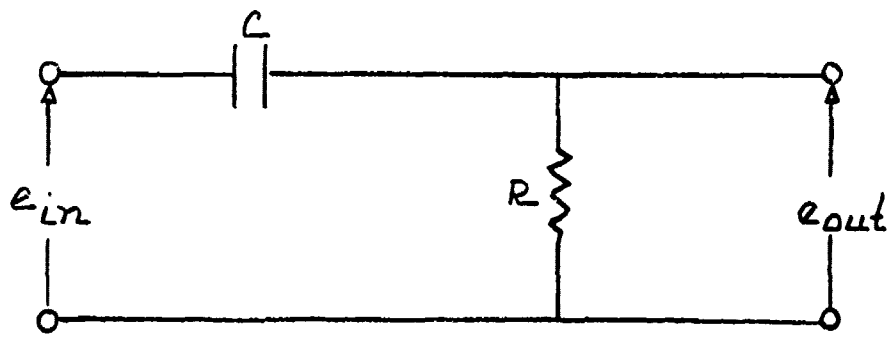
(d) Band-Pass Filter

Figure 2.5.3.1 -- Spectral Filter Responses

(a) Low-Pass Filter



(b) High-Pass Filter



(c) Band-Pass Filter

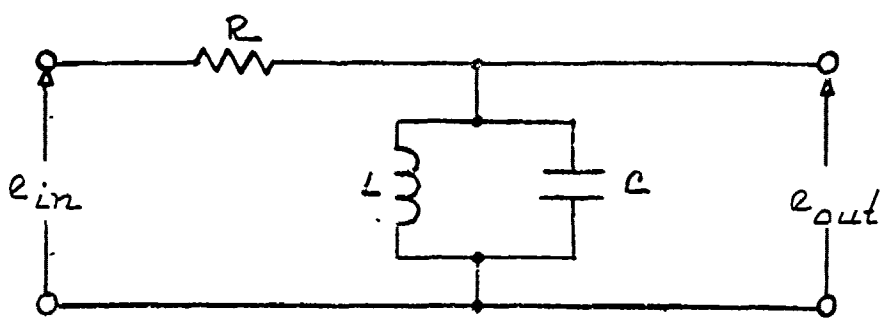


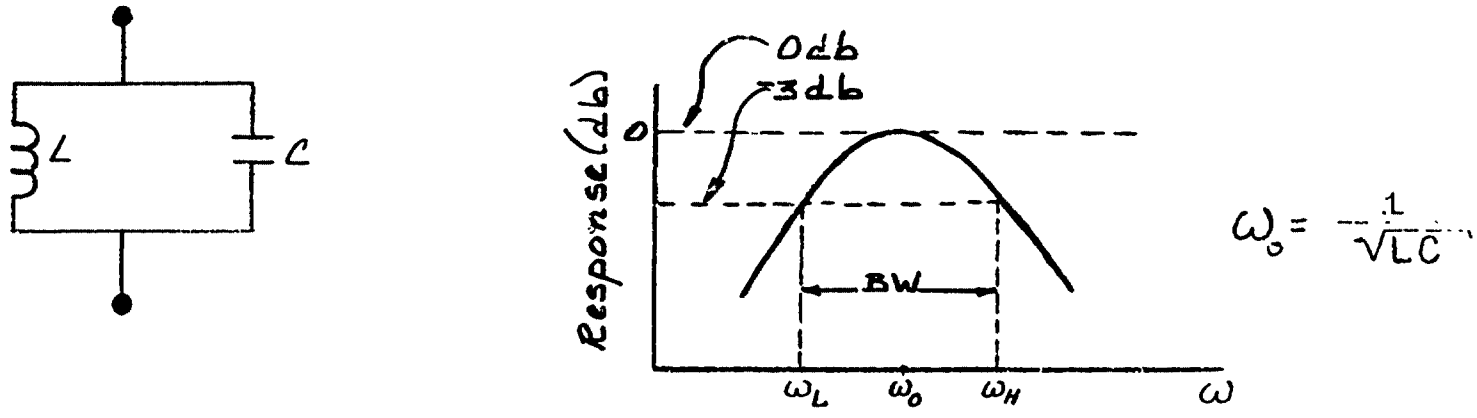
Figure 2.5.3.2 -- Spectral Filter Circuits

The frequency range between the low and high cut-off frequencies (that is, the pass-band) of a band-pass filter is called the bandwidth of the filter. The cut-off frequencies are defined as indicated in Figure 2.5.3.3(a). The cut-off frequencies are those frequencies for which the (power) response of the filter is one-half of the value at the peak (3 db down from the peak). This definition of bandwidth is applied to systems in general, as indicated in Figure 2.5.3.3 (b).

2.5.4 Signal Amplification and Attenuation -- Perhaps the most common signal processing operation is that of amplification or attenuation; that is, multiplying the signal amplitude by a non-frequency-dependent constant. Multiplication by a constant is, of course, a linear operation. Thus, no change is effected in the spectral content of the signal. If the gain (multiplying factor) of the operation varies as a function of signal magnitude, however, the process is nonlinear, and spectral changes occur. Compression, clipping, limiting, and rectifying involve such nonlinear gains.

2.5.5 Modulation and Demodulation -- Modulation is the process of altering some characteristic of one signal in accordance with another signal. The intent usually is to impress upon the first signal (the carrier) the information content of the other (the modulating signal). The result is a modulated carrier incorporating the desired properties of both signals. Demodulation is the process of extracting the modulating signal, and hence the information, from the modulated carrier. The carrier can be of many forms; however, we will, in this text, be almost always concerned with a pure sinusoidal carrier or a pulsed-sinusoidal carrier.

(a) Tuned Circuit Bandwidth



(b) System Bandwidth

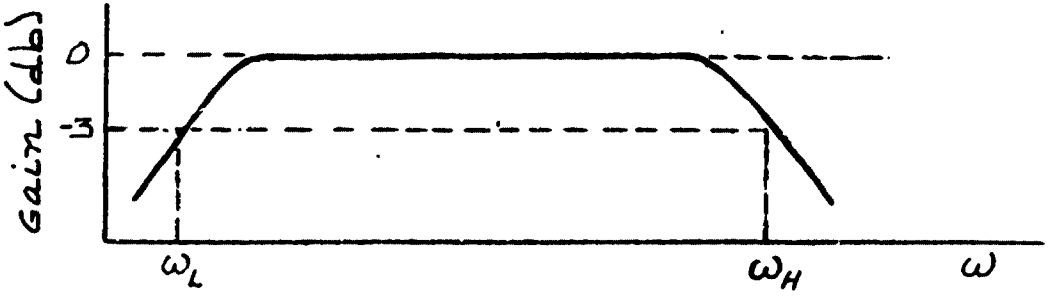


Figure 2.5.3.3 — Bandwidth Definitions

The characteristics of the carrier that can be altered by modulation are the amplitude and the frequency or phase. Note that frequency and phase angle are intimately related. That is:

$$\theta = \int \omega(t)\, dt$$

$$\omega = \frac{d\theta}{dt} \qquad [2.5.5.1]$$

where θ is phase angle and w is frequency.

Modulation is a nonlinear operation and, therefore, generates frequencies in the modulated signal not present in the carrier or the modulating signal. Thus, modulation always alters the bandwidth of the signal. As with any process, modulation affects both the time domain and the frequency domain characteristics of the signal.

As previously indicated, the basic forms of modulation are amplitude modulation and frequency modulation. Amplitude modulation varies the amplitude of the carrier in proportion to the magnitude of the modulating signal, as shown in Figure 2.5.5.1. The actual modulation process is fundamentally the same for an analog and a digital signal. For the binary digital case shown in Figure 2.5.5.1 (b), the magnitude of the modulated carrier has only two possible values.

As an example of amplitude modulation, consider the modulation of a pure sinusoidal carrier with a pure sinusoidal modulating signal, as shown in Figure 2.5.5.2. Let the carrier be given by:

$$e_c(t) = E_c \sin(\omega_c t) \qquad [2.5.5.2]$$

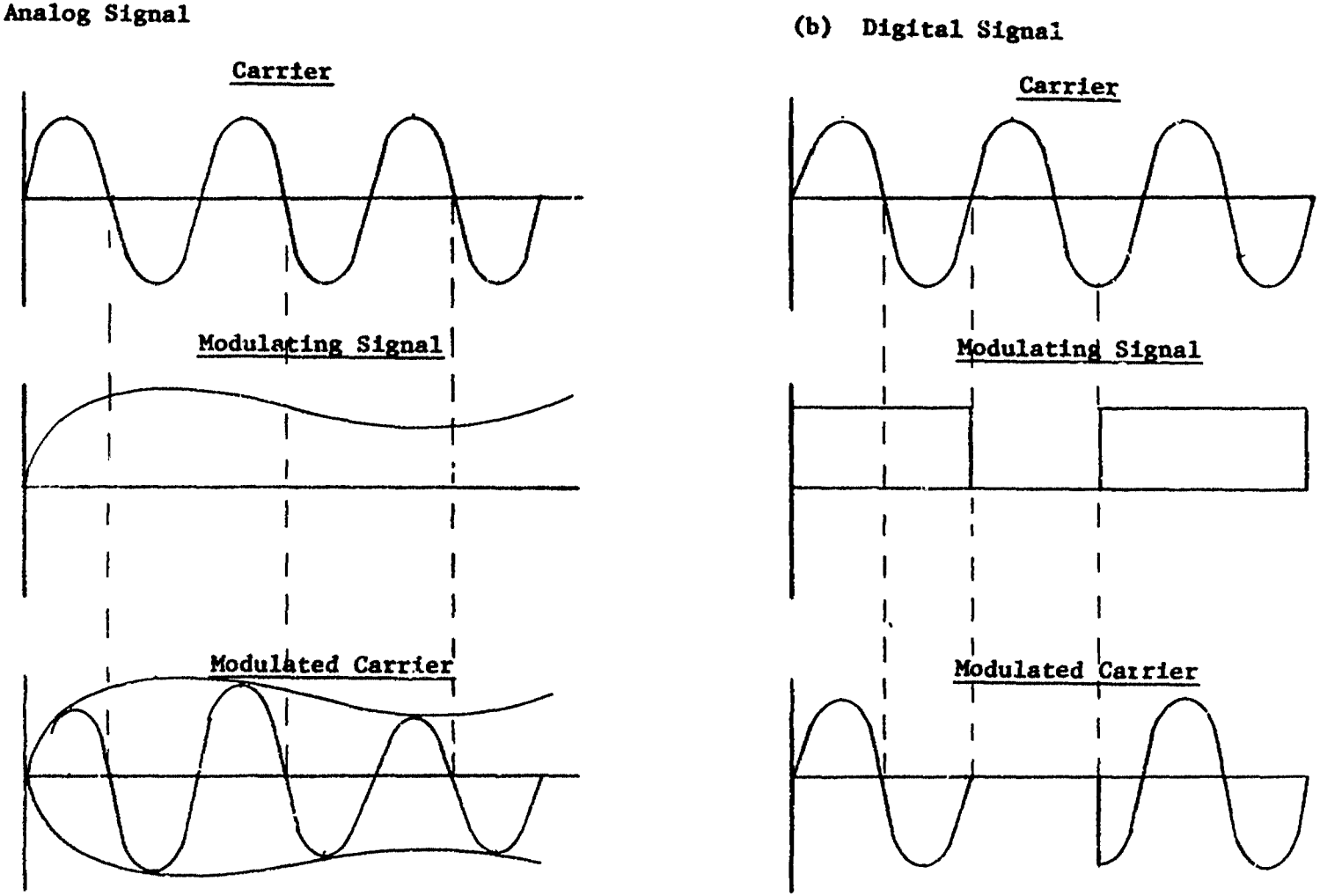
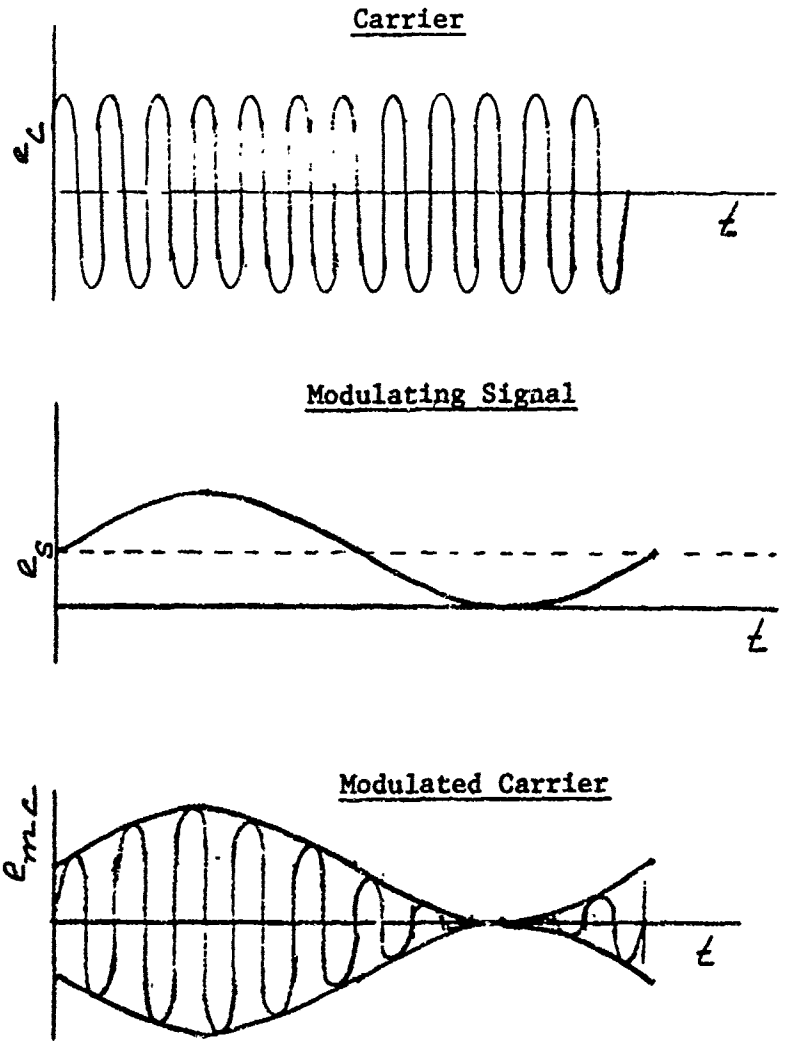


Figure 2.5.5.1 -- Amplitude Modulation

(a) Time Domain



(b) Frequency Domain (Spectrum)

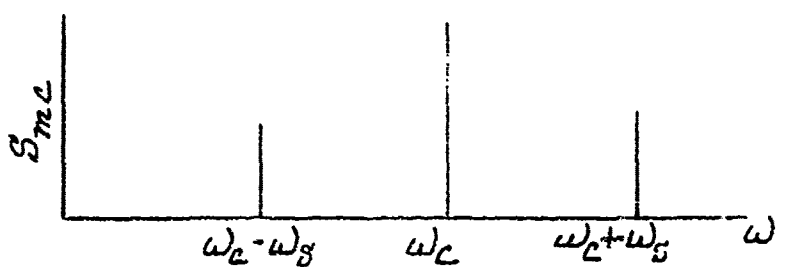


Figure 2.5.5.2 — Amplitude Modulation of a Pure Sinusoid with a Pure Sinusoid

and the modulating signal be given by:

$$e_s(t) = E_s \sin(\omega_s t) \qquad [2.5.5.3]$$

Define the modulation process by the equation:

$$e_{mc}(t) = [1 + m\, e_s(t)]\, e_c(t) \qquad [2.5.5.4]$$

where m is the modulation index.[1] The modulated carrier is, then, given by:

$$e_{mc}(t) = E_c \sin(\omega_c t) + m E_s E_c \sin(\omega_c t) \sin(\omega_s t) \qquad [2.5.5.6]$$

or, employing the trigonometric identity:

$$\sin\alpha \sin\beta = \tfrac{1}{2}\cos(\alpha-\beta) - \tfrac{1}{2}\cos(\alpha+\beta),$$

the modulated carrier is given by:

$$e_{mc}(t) = E_c \sin[\omega_c t] + \left(\frac{m E_s E_c}{2}\right)\cos[(\omega_c-\omega_s)t] - \left(\frac{m E_s E_c}{2}\right)\cos[(\omega_c+\omega_s)t] \qquad [2.5.5.7]$$

The time-domain representation of the modulated carrier is shown in Figure 2.5.5.2 (a). From Equation [2.5.5.7], it can be seen that there are three frequency components present in the modulated carrier: the carrier frequency, w_c, the upper sideband frequency, $(w_c + w_s)$, and the lower sideband

1. Amplitude modulation defined in this manner avoids the possibility of overmodulation, which results in a reversal of carrier phase. In suppressed-carrier modulation, the carrier is multiplied by the modulating signal alone. That is:

$$e_{mc}(t) = e_s(t)\, e_c(t) \qquad [2.5.5.5]$$

frequency, $(w_c - w_s)$, as shown in Figure 2.5.5.2 (b). The generation of sum and difference frequencies (sidebands) in this manner is typical of the amplitude modulation process. For a complex modulating signal, the spectrum of the modulating signal is replicated above and below the carrier frequency as shown in Figure 2.5.5.3. As a result of this frequency domain replication, the bandwidth of an amplitude modulated signal is given by:

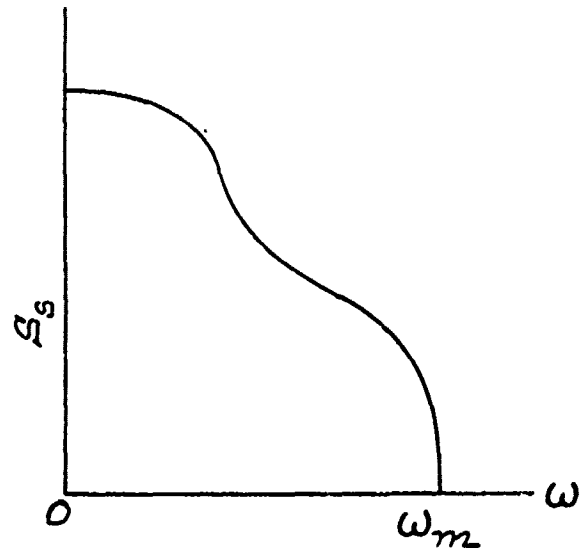
$$\text{(Bandwidth)} = 2\ \text{[Highest Modulating Frequency]} \qquad [2.5.5.8]$$

Note that all the necessary information is present in a single sideband to reconstruct the entire modulated carrier and, hence, the modulating signal. Single sideband systems make use of this fact to conserve system bandwidth and transmitting power.

Amplitude modulation and demodulation are identical processes. That is, both involve multiplying an information-carrying signal by a purely sinusoidal carrier signal. In Equations [2.5.5.2] through [2.5.5.7], the multiplicative modulation process was demonstrated by deriving the expression for a sinusoidal carrier amplitude modulated by a sinusoidal information-carrying signal. In the following development, demodulation (synchronous detection) will be demonstrated utilizing the same multiplicative process. Starting with the modulated carrier from Equation [2.5.5.7], we have:

$$e_{mc}(t) = E_c \sin[\omega_c t] + \left(\frac{m E_s E_c}{2}\right)\cos[(\omega_c - \omega_s)t] - \left(\frac{m E_s E_c}{2}\right)\cos[(\omega_c + \omega_s)t] \qquad [2.5.5.9]$$

(a) Spectrum of Modulating Signal



(b) Spectrum of Modulated Carrier

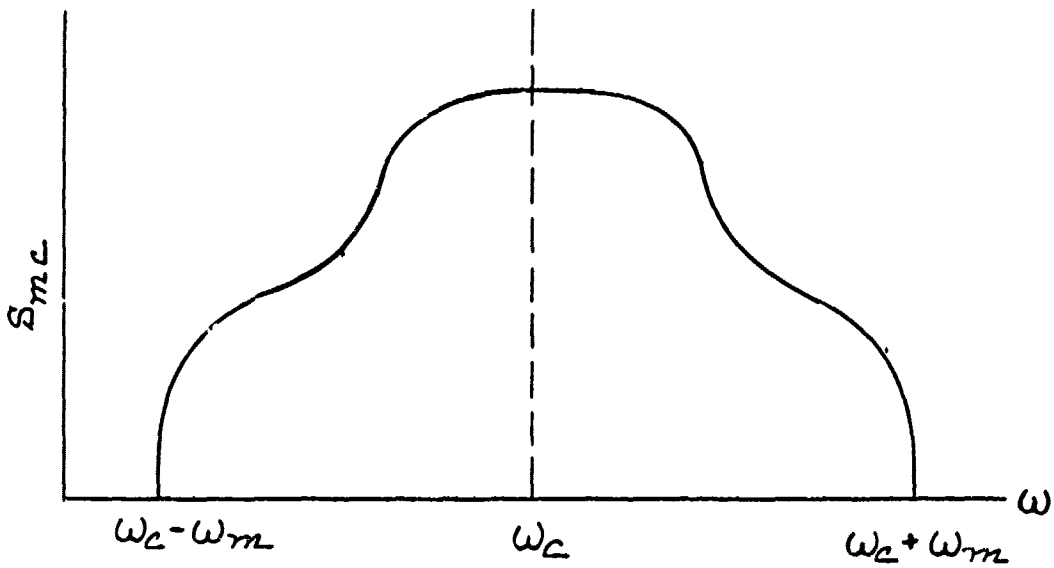


Figure 2.5.5.3 -- Spectrum of Amplitude Modulated Carrier

Multiplying by a reference (carrier) signal, $e_r(t)$, given by:

$$e_r(t) = E_r \sin[\omega_c t]$$

we have, for the demodulated signal, $e_d(t)$:

$$e_d(t) = E_c E_r \sin^2[\omega_c t] + \left(\frac{m E_c E_r E_s}{2}\right) \sin[\omega_c t] \cos[(\omega_c - \omega_s)t] - \left(\frac{m E_c E_r E_s}{2}\right) \sin[\omega_c t] \cos[(\omega_c + \omega_s)t] \quad [2.5.5.10]$$

or, employing the trigonometric identities:

$$\sin\alpha \cos\beta = \tfrac{1}{2}\sin(\alpha+\beta) + \tfrac{1}{2}\sin(\alpha-\beta)$$

$$\sin^2\alpha = \tfrac{1}{2} - \tfrac{1}{2}\cos(2\alpha)$$

we have:

$$e_d(t) = \left(\frac{E_c E_r}{2}\right)\{1 - \cos[2\omega_c t]\} + \left(\frac{m E_c E_s E_r}{4}\right)\{\sin[(2\omega_c - \omega_s)t] + \sin[\omega_s t]\} - \left(\frac{m E_c E_s E_r}{4}\right)\{\sin[(2\omega_c + \omega_s)t] - \sin[\omega_s t]\} \quad [2.5.5.11]$$

where:

$$\omega_s << (2\omega_c - \omega_s) < 2\omega_c < (2\omega_c + \omega_s) \quad [2.5.5.12]$$

Employing a low-pass filter to remove all components with frequencies much greater than ω_s, we have, for the demodulated signal:

$$e_d(t) = \left(\frac{E_c E_r}{2}\right)[1 + m E_s \sin(\omega_s t)] \quad [2.5.5.13]$$

which, except for an arbitrary gain and a D.C. offset, is the desired modulating signal. Similar multiplicative processes are used to substitute

one carrier frequency for another by offsetting the frequency of the reference signal. This process is called carrier frequency conversion and is used in the tuner and intermediate-frequency sections of communications equipment as well as in the detectors of Doppler radar equipment.

A simple, and frequently employed, type of amplitude demodulation is simple rectification of the modulated signal. Rectification is the deletion of all portions of a signal that reverse sign. In order to demonstrate demodulation by means of rectification, we will demodulate the sinusoidally modulated sinusoid previously demodulated by synchronous detection. Starting with the form of the modulated carrier given in Equation [2.5.5.6], we have, for the modulated carrier:

$$e_{mc}(t) = [1 + m \sin(\omega_s t)] E_c \sin(\omega_c t) \qquad [2.5.5.14]$$

Since the term $(1 + m \sin(\omega_s t)) E_c$ never goes negative, the rectified form of e_{mc} (t), denoted by Rect $\{e_{mc}$ (t)$\}$, is given by:

$$\text{Rect}\{e_{mc}(t)\} = [1 + m \sin(\omega_s t)] E_c \, \text{Rect}\{\sin(\omega_c t)\} \qquad [2.5.5.15]$$

The waveform of Rect $\{\sin(\omega_c t)\}$ is shown in Figure 2.5.5.4. Rect$\{\sin(\omega_c t)\}$ can be expanded in a Fourier series to give:

$$\text{Rect}\{\sin(\omega_c t)\} = \frac{1}{\pi} + \frac{1}{2}\sin(\omega_c t) - \frac{2}{3\pi}\cos(2\omega_c t) + \cdots \qquad [2.5.5.16]$$

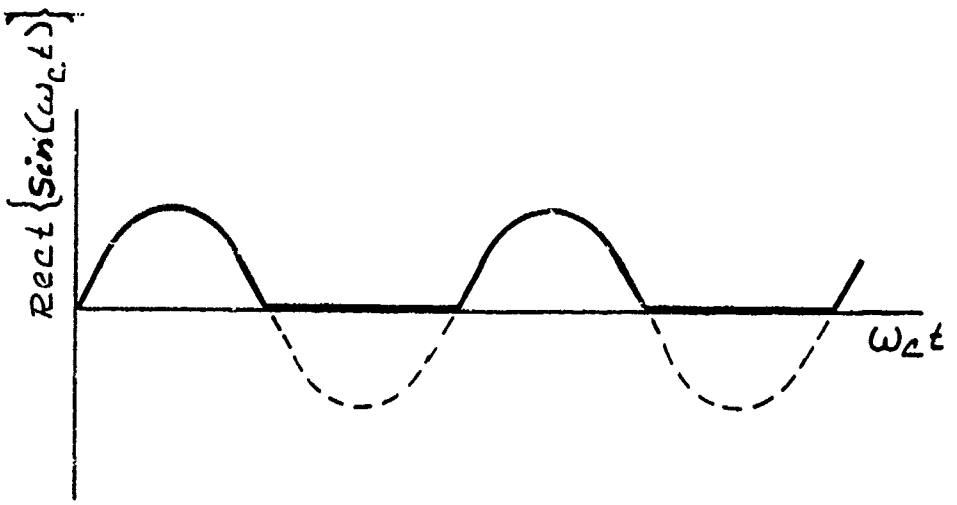


Figure 2.5.5.4 Waveform of $Rect\{sin(\omega_c t)\}$

Thus:

$$Rect\{e_{mc}(t)\} = E_c[1 + m\,Sin(\omega_s t)]\left[\frac{1}{\pi} + \frac{1}{2}Sin(\omega_c t) + \cdots\right] \quad [2.5.5.17]$$

where $\omega_s << \omega_c$. Employing a low-pass filter to remove all components with frequency much greater than ω_s, we have, for the demodulated signal, $e_d(t)$:

$$e_d(t) = \frac{E_c}{\pi}\left[1 + m\,Sin(\omega_s t)\right] \quad [2.5.5.18]$$

which, except for an arbitrary gain and a D.C. offset, is the desired modulating signal.

Frequency modulation varies the instantaneous frequency of the carrier in proportion to the magnitude of the modulating signal, as shown in Figure 2.5.5.5. As in the case of amplitude modulation, the modulation process is fundamentally the same for analog and digital signals.

As an example of frequency modulation, consider, again, the modulation of a pure sinusoidal carrier with a pure sinusoidal modulating signal, as shown in Figure 2.5.5.6. Let the carrier be given by:

$$e_c(t) = E_c\,Sin(\omega_c t) \quad [2.5.5.19]$$

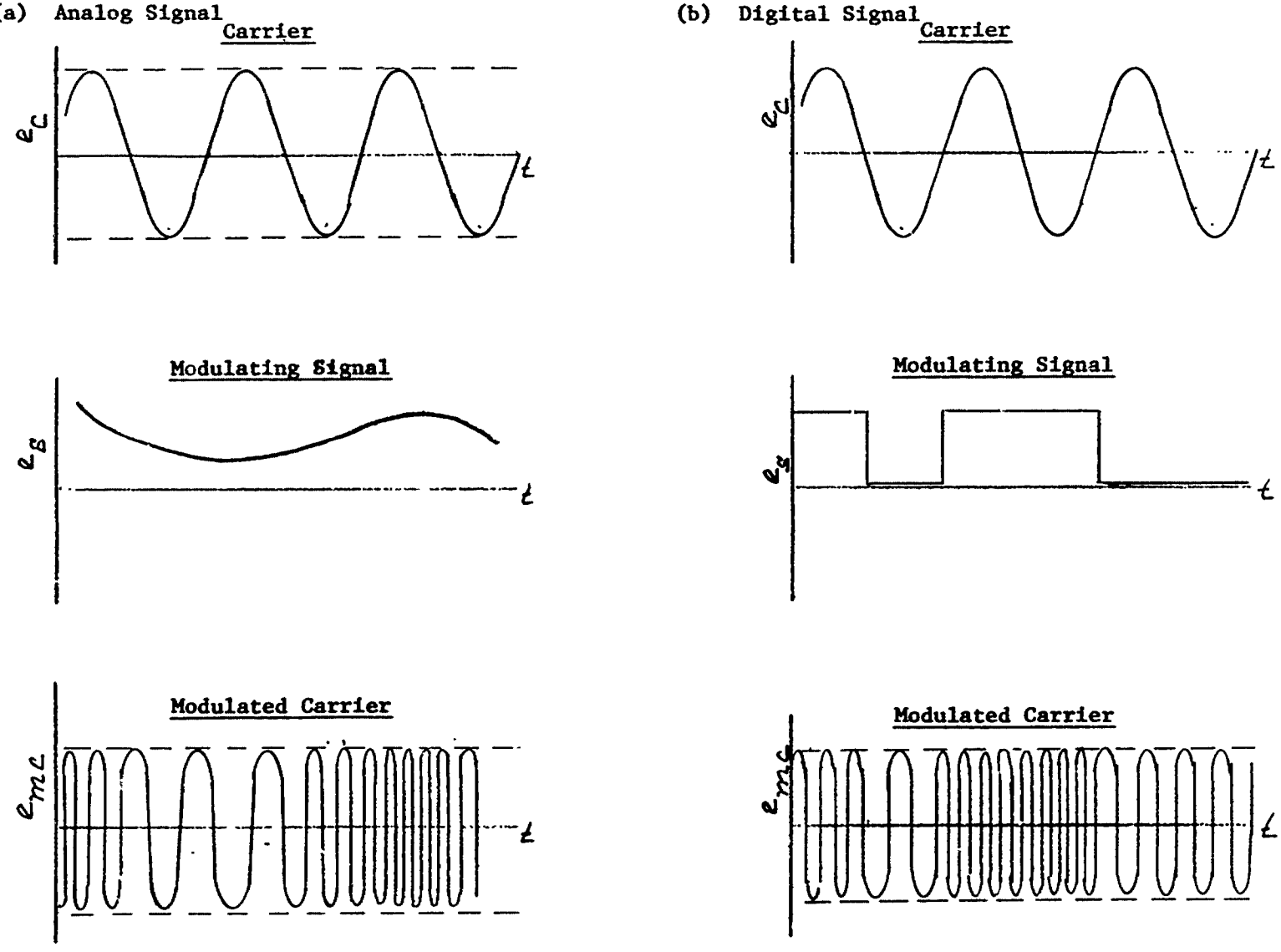
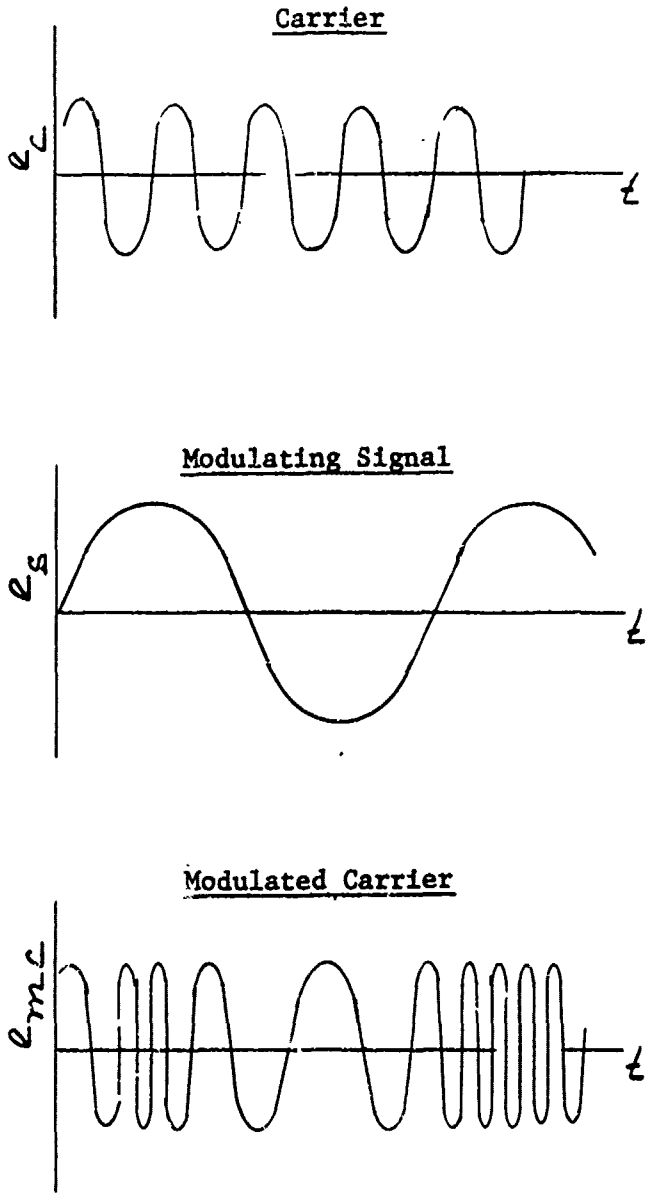


Figure 2.5.5.5 -- Frequency Modulation

(a) Time Domain



(b) Frequency Domain (Spectrum)

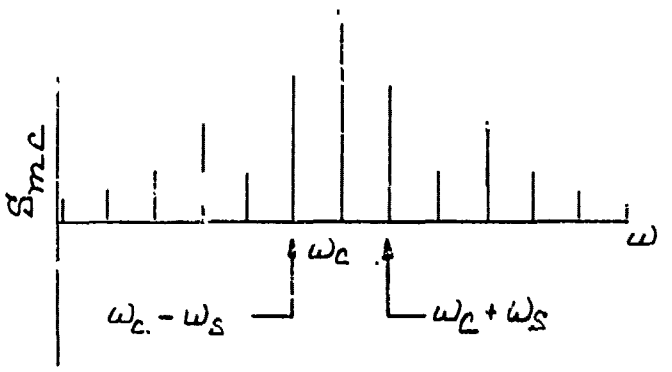


Figure 2.5.5.6 -- Frequency Modulation of a Pure Sinusoid with a Pure Sinusoid

and the modulating signal be given by:

$$e_s(t) = E_s \cos(\omega_s t) \qquad [2.5.5.20]$$

Define the modulation process by the equation:

$$e_{mc}(t) = E_{mc} \sin\left[\int (\omega_c + k\, e_s)\, dt\right] \qquad [2.5.5.21]$$

where k is a modulation coefficient. The modulated carrier is then given by:

$$e_{mc}(t) = E_{mc} \sin\left[\omega_c t + \left(\frac{\Delta\omega}{\omega_s}\right) \sin(\omega_s t)\right] \qquad [2.5.5.22]$$

where the deviation ratio, $(\Delta w/w_s)$, is given by:

$$\frac{\Delta\omega}{\omega_s} = \frac{k E_s}{\omega_s} \qquad [2.5.5.23]$$

Equation [2.5.5.22], which involves a trigonometric function of a trigonometric function, can be expanded, in terms of Bessel functions, to the form:

$$\begin{aligned} e_{mc}(t) = E_{mc} \{ & J_0 \cos[\omega_c t] + J_1 \cos[(\omega_c+\omega_s)t] \\ & - J_1 \cos[(\omega_c-\omega_s)t] + J_2 \cos[(\omega_c+2\omega_s)t] \\ & - J_2 \cos[(\omega_c-2\omega_s)t] + J_3 \cos[(\omega_c+3\omega_s)t] \\ & + \cdots \} \end{aligned} \qquad [2.5.5.24]$$

where the J's are Bessel functions of the first kind and are functions only of the deviation ratio. Thus, for the simple case of a sinusoidally frequency modulated sinusoid, the spectrum of the modulated carrier consists of numerous sidebands at frequencies separated by an interval equal to w_s, above and below w_c, as shown in Figure 2.5.5.6 (b). This result demonstrates the fact that, for the same information content, a frequency modulated carrier requires, in general, a much greater system bandwidth than does an amplitude modulated carrier. A rule-of-thumb formula for the required system bandwidth is:

$$\text{Bandwidth} = 2\ [\text{Highest Modulating Frequency} + \text{Greatest Frequency Deviation}] \qquad [2.5.5.25]$$

Obviously, this formula does not include all sidebands. (There are an infinite number with ever-decreasing magnitudes). It does, however, include enough sidebands that the remaining ones contain only a small fraction of the total power. The spectral complexity of a frequency modulated carrier is responsible for its greatest advantage over an amplitude modulated carrier, for radio communication. Due to its complexity, an FM signal is difficult for natural phenomena (e.g. atmospherics) to mimic. For that reason, FM communication is relatively free of interfering noise.

Phase modulation varies the instantaneous phase angle of the carrier in proportion to the magnitude of the modulating signal, as shown in Figure 2.5.5.7. For the digital (discrete) case, shown in Figure 2.5.5.7 (b), the instantaneous phase reversals are evident.

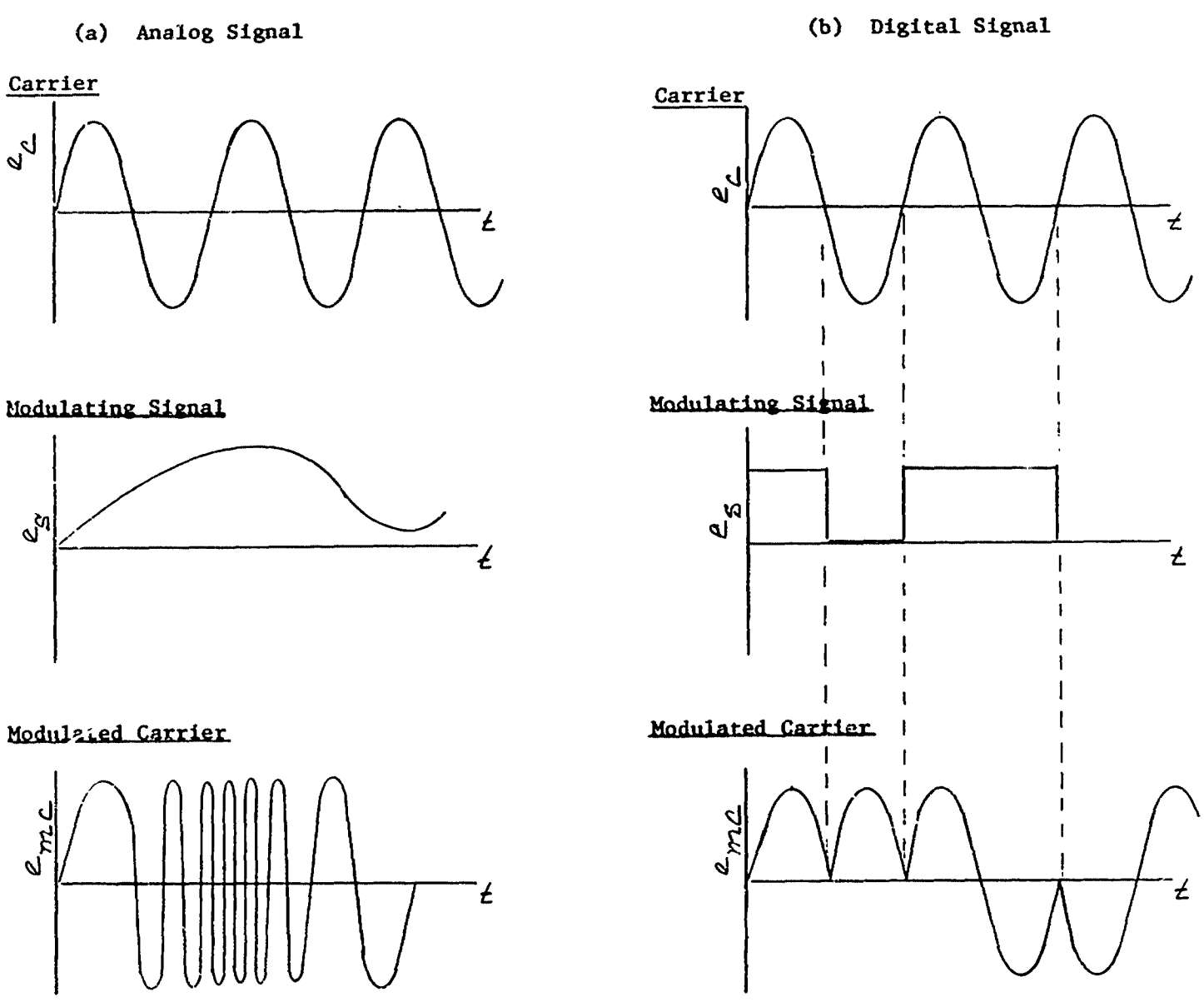


Figure 2.5.5.7 -- Phase Modulation

As an example of phase modulation, consider, again, the case where both the carrier and the modulating signal are pure sinusoids. Thus, let:

$$e_c(t) = E_c \sin(\omega_c t) \qquad [2.5.5.26]$$

and:

$$e_s(t) = E_s \sin(\omega_s t) \qquad [2.5.5.27]$$

Define the modulation process by the equations:

$$e_{mc}(t) = E_{mc} \sin[\omega_c t + k e_s(t)] \qquad [2.5.5.28]$$

where k is a modulation coefficient. The modulated carrier is then given by:

$$e_{mc}(t) = E_{mc} \sin[\omega_c t + m \sin(\omega_s t)] \qquad [2.5.5.29]$$

where the modulation index m is given by:

$$m = k E_s \qquad [2.5.5.30]$$

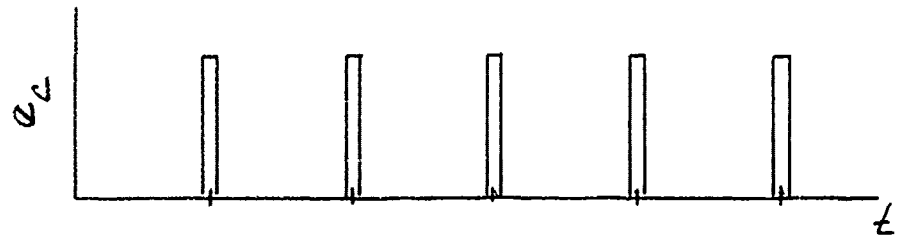
A comparison of Equations [2.5.5.22] and [2.5.5.29] indicates that, for the case of sinusoidal modulation of a sinusoid, the only difference between the result for phase modulation and that for frequency modulation is in the definitions of the modulation coefficients. In frequency modulation, the deviation ratio is inversely proportional to signal frequency. In phase modulation, the modulation index is not. This result demonstrates

the great similarity between frequency and phase modulation. The sidebands are the same for the two cases, for the same modulation coefficient. Identical results from frequency and phase modulation can be obtained by phase modulating the carrier with the time derivative of the signal used to frequency modulate the carrier. The most significant differences between frequency and phase modulation are those concerned with hardware implementation.

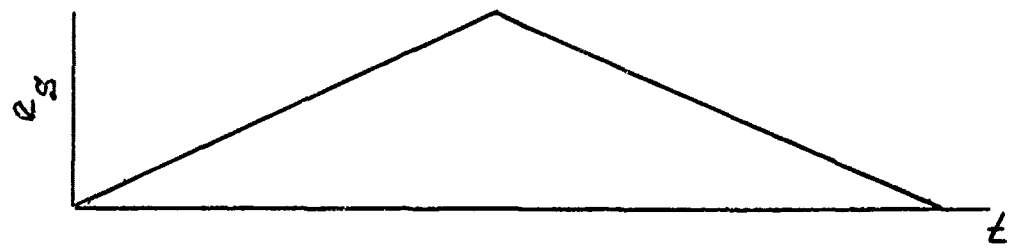
Pulse modulation differs from those types of modulation discussed previously in that the carrier is a pulse train, (periodic bursts of sinusoid), rather than a continuous sinusoid. The parameters of the pulse train that can be modulated (varied) are the amplitude (PAM), width (PWM), and position (PPM) or frequency (PFM) of the pulses. The three basic types of pulse modulation are illustrated in Figure 2.5.5.8. It should be noted that, in pulse modulation, the pulse amplitude, width, and position are the analog quantities for the modulating signal magnitude. Thus, pulse modulation is a discrete or sampled-data process but it is not a digital process.

The pulsing of the carrier in pulse modulation greatly complicates the spectrum of the modulated carrier. The spectrum of the pulse train itself, with no modulation, is a uniformly spaced array of frequency components at intervals equal to the pulse repetition rate, f_r, to either side of the sinusoidal carrier frequency, f_o, as shown in Figure 2.5.5.9 (a). When a pure sinusoidal amplitude modulation of frequency f_s is applied to the pulse train, sidebands are generated on both sides of each of the frequency components contained in the pulse train spectrum, as shown in Figure 2.5.5.9 (b).

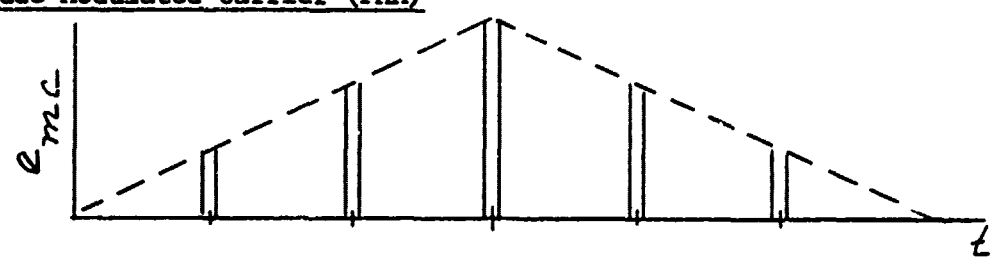
Carrier (Pulse Envelope only)



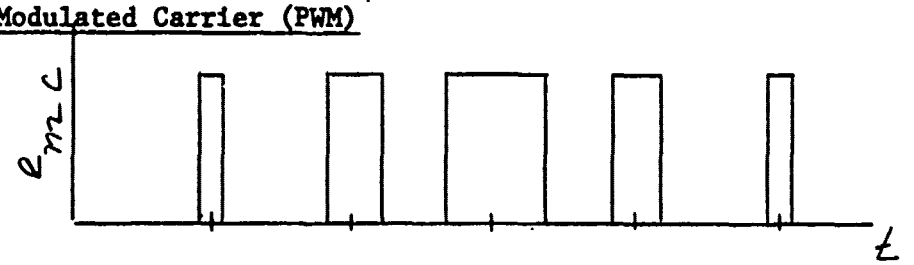
Modulating Signal



Amplitude Modulated Carrier (PAM)



Width Modulated Carrier (PWM)



Position Modulated Carrier (PPM)

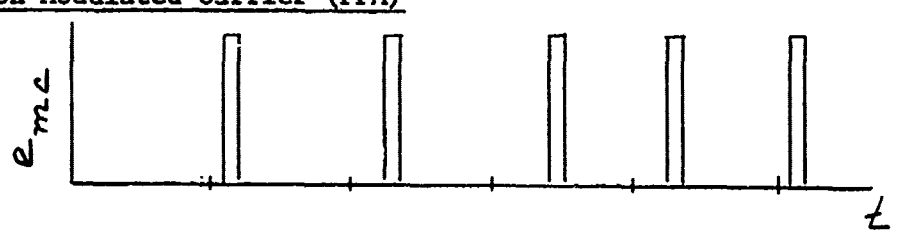
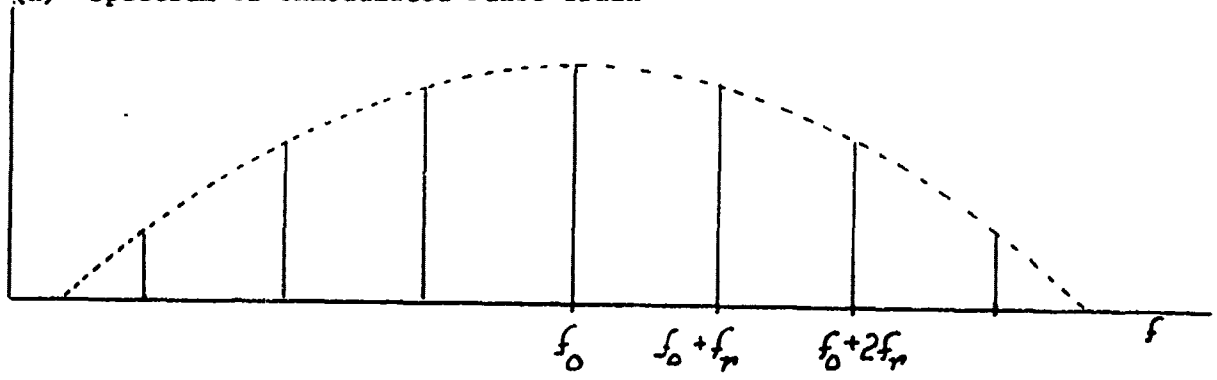


Figure 2.5.5.8 -- Pulse Modulation

(a) Spectrum of Unmodulated Pulse Train



(b) Spectrum of Sinusoidally Modulated Pulse Train

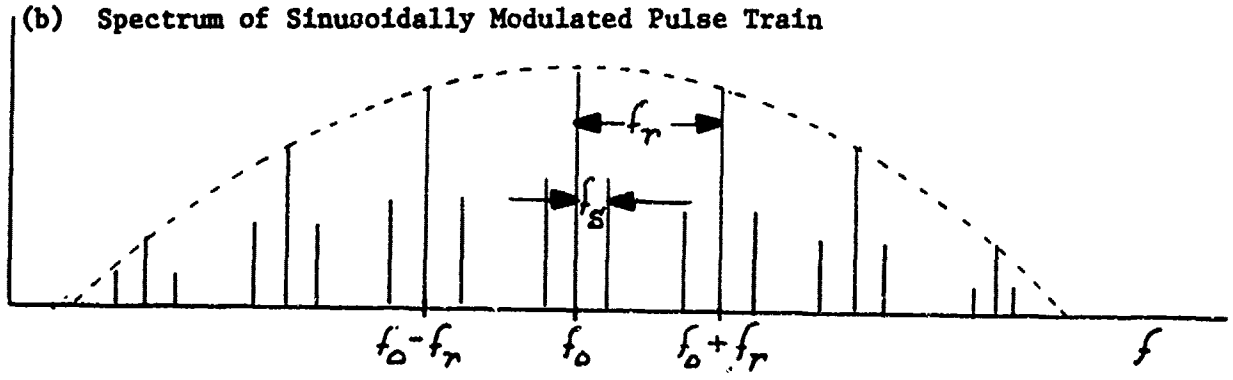


Figure 2.5.5.9 -- Spectrum of Pulse Train with Pulse Amplitude Modulation

Pulse code modulation (PCM) employs a symbolic code. It is, therefore, a digital signal processing technique. Like all digital methods, it deals with discrete (sampled) quantities. These "samples" may be pulses in amplitude (Amplitude Shift Keying or ASK), such as those illustrated in Figures [2.5.5.8], or they may be "pulses" in the frequency (Frequency Shift Keying or FSK) or phase (Phase Shift Keying or PSK) of an otherwise continuous carrier. In the illustrations, we will assume pulses in amplitude. That is, we will assume that the carrier consists of bursts of sinusoidal carrier.

In pulse code modulation, the magnitude of the modulating signal is quantized as shown in Figure 2.5.5.10. Each quantum of magnitude is then assigned a symbol consisting of a string of pulses which, if the code is binary, are either present or absent, in their assigned position in the symbol. In the example of Figure 2.5.5.10, a three-bit straight binary code is assigned to the quantum values zero through seven. The modulated carrier, shown in the figure is a train of pulses, in groups of three, in which the absence of a pulse denotes a zero and the presence of a pulse denotes a one. The pulse train is seen to include the sequence 000,001,100, and 110, signifying values of the modulating signal, for successive sample periods, of 0,1,4, and 6.

The principal advantages of pulse code modulation derive from the digital nature of the process. They are, as stated in Section 2.2 of this text, accuracy and excellent noise rejection capability. The most important

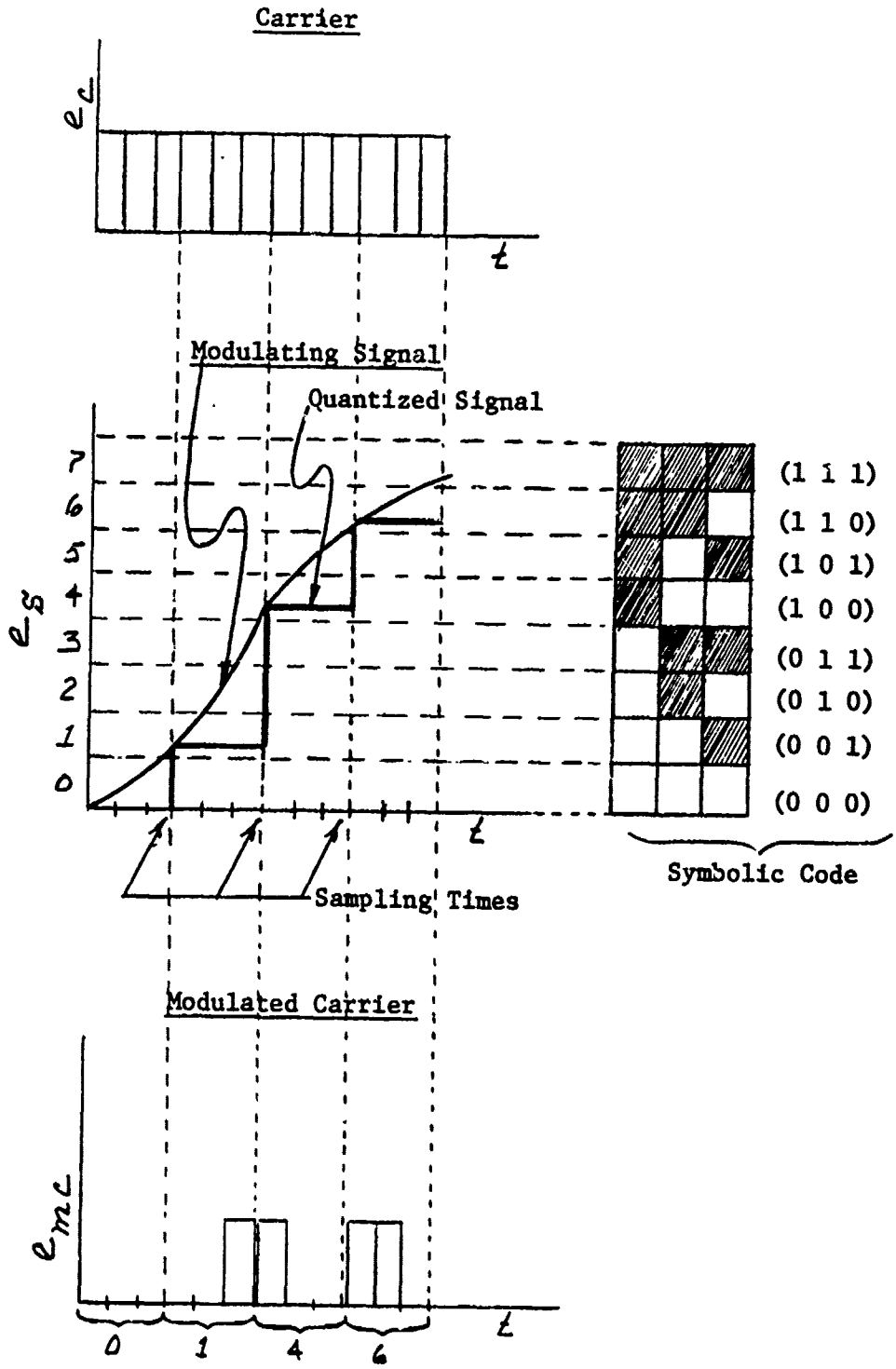


Figure 2.5.5.10 -- Pulse Code Modulation

disadvantage of PCM is the large required system bandwidth. An approximate formula for the bandwidth of a PCM signal is given by:

$$\text{Bandwidth} = 2 \left[\begin{array}{l} (\text{Highest Modulating Frequency}) \\ \times\ (\text{Number of Quantum Levels}) \end{array} \right] \qquad [2.5.5.31]$$

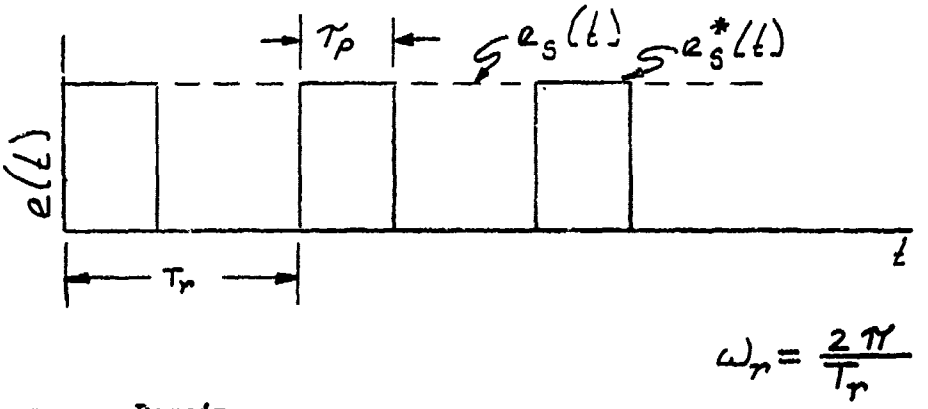
Other disadvantages of PCM are the sampled-data effects, also discussed in Section 2.2 of this text.

2.5.6 Pulsing or Time Sampling -- In many airborne systems of interest, the signals are not continuously available. Digital systems are, by nature, sampled-data systems. Scanning sensors are also time-sampling devices. Time-division multiplex renders output signals which are sampled or discrete-time quantities. In the time domain, the effects of pulsing or sampling are reasonably obvious. The sampled signal is quantized and/or discontinuous. In Figure 2.5.6.1 (a), a non-time-varying signal, $e_s(t)$, is shown sampled, with sampling interval (period) T_r and sampling duration (pulse width) τ_p. Note that sampling is not necessarily synchronous (periodic). We will not, however, consider non-periodic (asynchronous) sampling here.

The frequency domain representation (spectrum) of the unsampled (D.C.) signal, $e_s(t)$, is simply a single line at zero frequency. The spectrum of the sample signal, $e_s^*(t)$, however, is quite complex. As shown in Figure 2.5.6.1 (b), it consists of an infinite number of lines, beginning at zero frequency and, in general, occurring at intervals of the sampling frequency, w_r, for all frequencies. The amplitudes of the spectral lines vary with frequency as shown by the dotted line in the figure. The amplitudes decrease with increasing frequency at a rate sufficient to yield a finite total power in the signal spectrum.

When the unsampled signal, $e_s(t)$, is time varying as shown in Figure 2.5.6.2 (a), the corresponding spectrum is as shown in Figure 2.5.6.2 (b). The spectrum for a time-varying $e_s(t)$, like that for a constant (unmodulated)

(a) Time Domain



(b) Frequency Domain

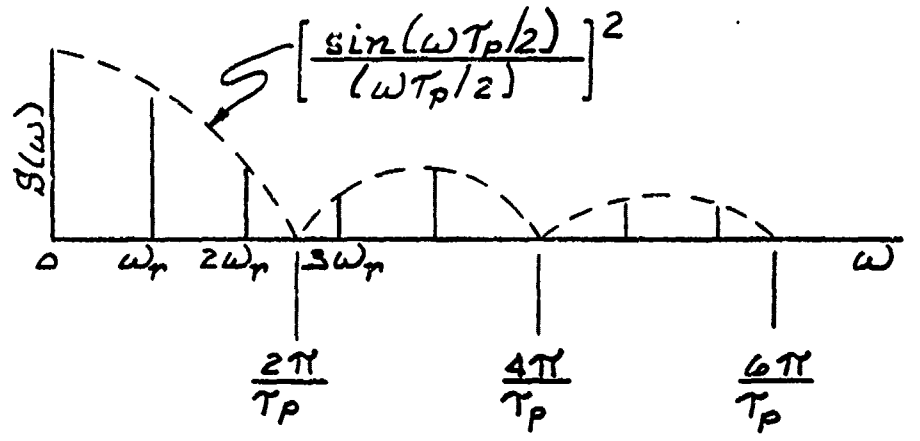
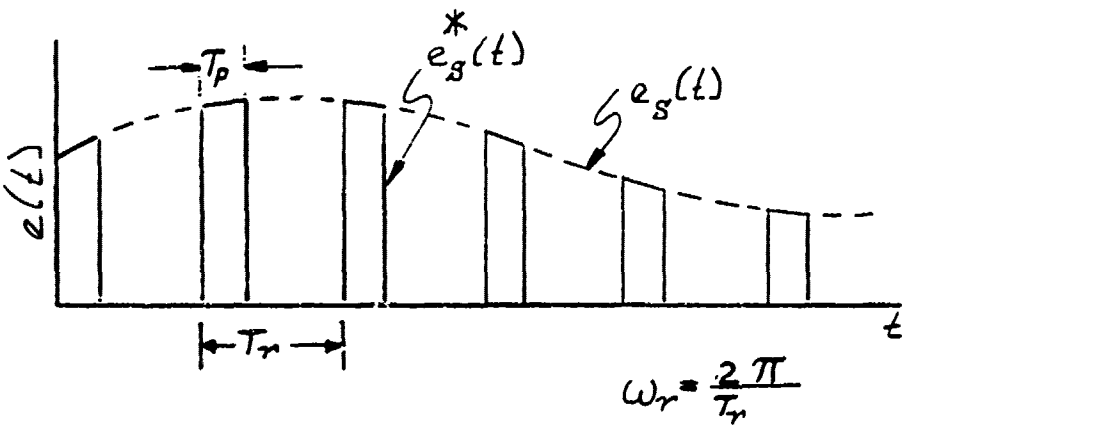


Figure 2.5.6.1 -- Pulsed Non-Time-Varying Waveform With No Carrier

(a) Time Domain



ω_{sm} = Max. Signal Frequency

(b) Frequency Domain

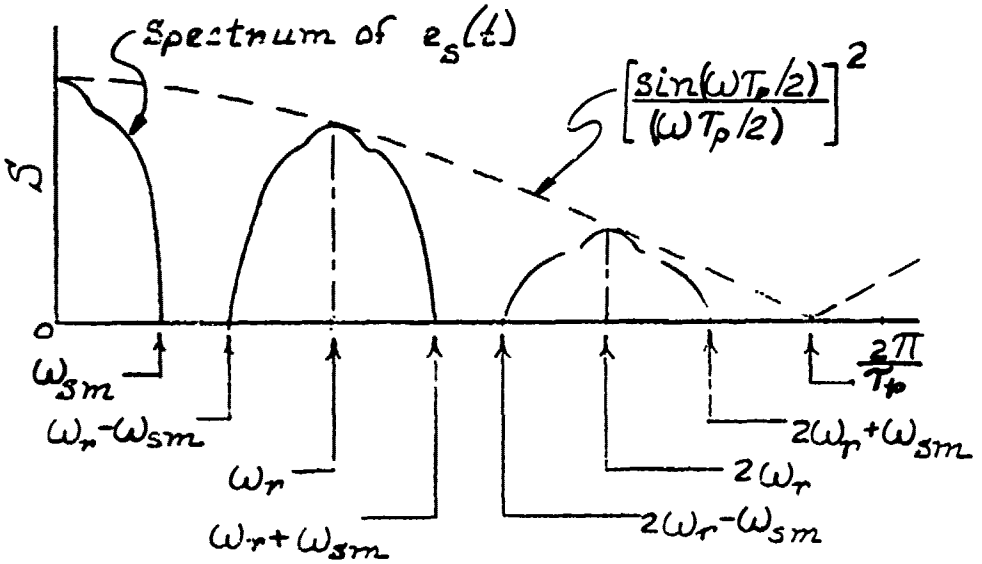


Figure 2.5.6.2 -- Pulsed Time - Varying Waveform with No Carrier

$e_s(t)$, has components at intervals of w_r for all frequencies; but with the entire spectrum of $e_s(t)$ reflected above and below each line at $w=n\ w_r$. Thus the spectrum of the modulating signal, $e_s(t)$, is replicated, and reflected, at intervals of w_r throughout the spectrum. The amplitudes of these replicated spectra vary with frequency, as did the single lines for the unmodulated case, decreasing rapidly with increasing frequency. The replication of the spectrum of the modulating signal, due to time-sampling or pulsing, is called frequency aliasing or folding because the signal spectrum appears in the same "shape" but at new positions on the frequency scale.

An important criterion for time-sampling a signal can be deduced from Figure 2.5.6.2 (b). Note that if the maximum frequency component in the modulating signal, w_{sm}, is large enough, the lower portion of the first folded spectrum,(at $w = w_r$), will overlap the upper portion of the primary (unfolded) spectrum. When the folded spectra overlap, the overlapping signal components cannot be separated by any means and modulating signal information is irretrievably lost. In order to prevent such loss of information, the frequency gap between the two spectra must be greater than zero. That is:

$$(w_r - w_{sm}) - w_{sm} > 0 \qquad [2.5.6.1]$$

or:

$$w_r \geq 2\ w_{sm}$$

This inequality, known as the sampling theorem, states that: As long as the sampling rate, w_r, is at least twice the highest frequency component in the signal to be sampled, the sampled signal contains all of the information present in the unsampled signal. Thus, in theory, if $w_r \geq 2\ w_{sm}$, the unsampled signal, $e_s(t)$, can be reconstructed from the sampled signal, $e_s^*(t)$. An illustration of the manner in which an insufficiently high sampling rate (insufficiently small T_r) can alias or fold frequency components of the sampled spectrum into a different portion of the frequency scale is shown in Figure 2.5.6.3. In that figure, the original signal, $e_s(t)$, is given by:

$$e_s(t) = E \sin(\omega_s t) \qquad [2.5.6.2]$$

That signal is sampled, as shown, with period T_r. The sampled wave form, $e_s^*(t)$, consists, in this case, of instantaneous sampled values at $t = n\ T_r$. The aliased waveform, $e_A(t)$, is shown faired through the sampled values and is given by:

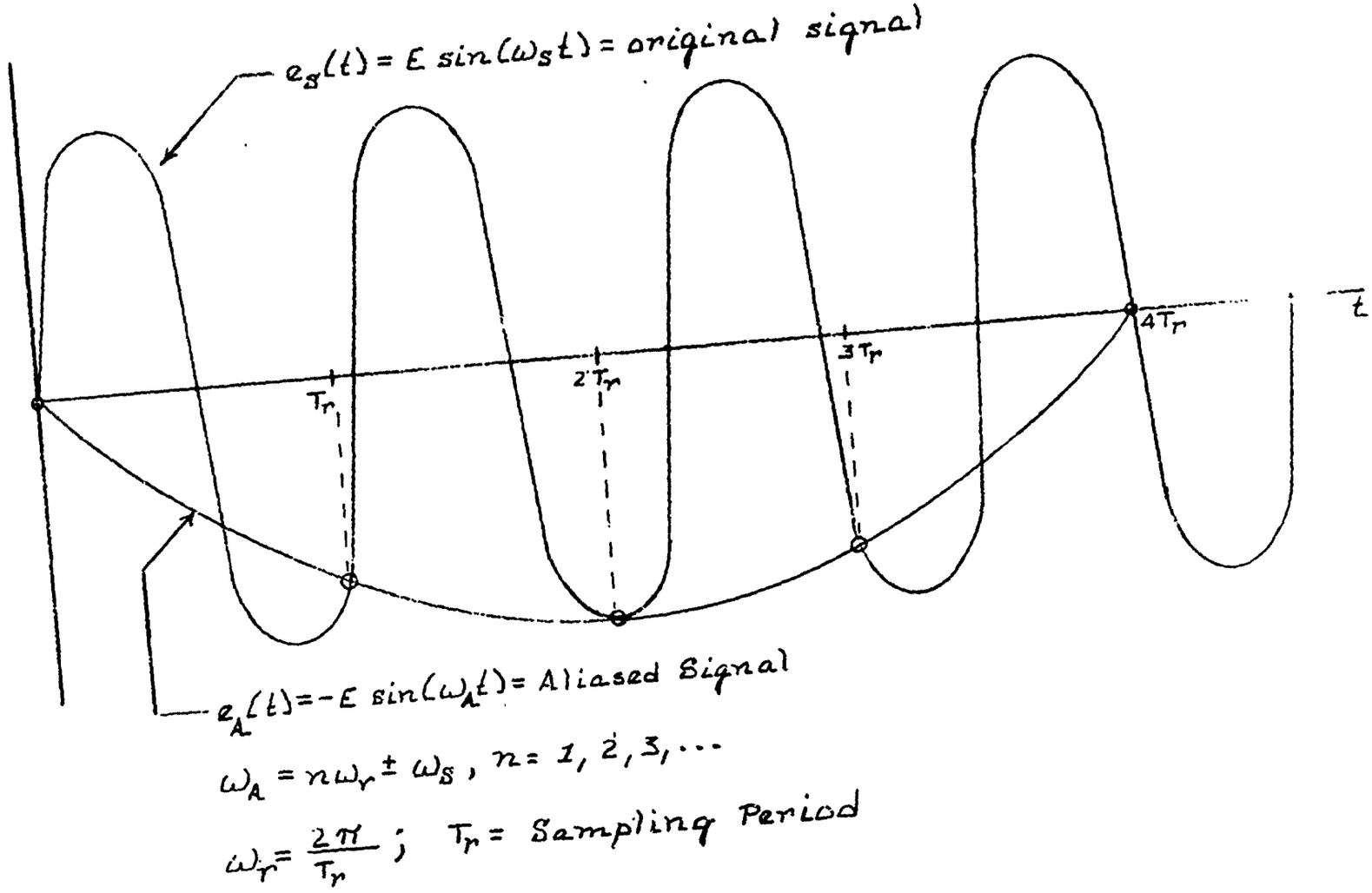
$$e_A(t) = -E \sin(\omega_A t) \qquad [2.5.6.3]$$

where the folded frequencies are given by:

$$\omega_A = n\omega_r \pm \omega_s \quad ; \quad n = 0, 1, 2, \cdots \qquad [2.5.6.4]$$

The particular aliased frequency illustrated in the figure is the lower sideband frequency for n = 1. Thus:

$$\omega_A = \omega_r - \omega_s \qquad [2.5.6.5]$$



$$\omega_A = n\omega_r \pm \omega_s, \quad n = 1, 2, 3, \ldots$$

$$\omega_r = \frac{2\pi}{T_r}; \quad T_r = \text{Sampling Period}$$

Figure 2.5.6.3 -- Frequency Aliasing Due to Time Sampling

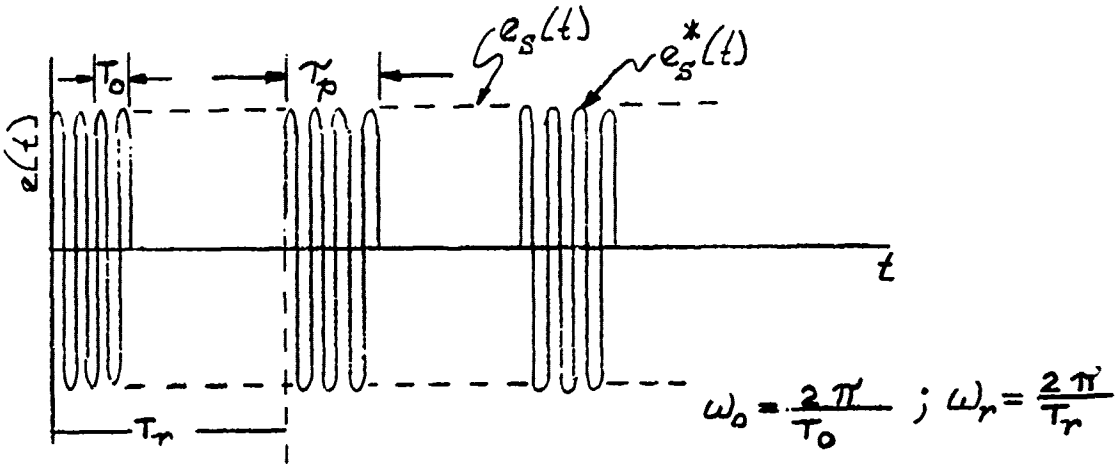
Frequency folding due to time sampling is especially troublesome in a pulse Doppler radar system where it can cause a fast-closing target to appear (alias) as an opening target or vice versa.

When the signal to be sampled consists of a carrier with constant (unmodulated) amplitude as shown in Figure 2.5.6.4 (a), the corresponding spectrum of the sampled signal is that shown in Figure 2.5.6.4 (b). Note that this spectrum is identical to that shown in Figure 2.5.6.1 (b) for a constant amplitude signal with no carrier, except that the entire spectrum, with carrier, is shifted upward in frequency by w_o, the carrier frequency. There are now folded frequency components both above and below w_o at intervals of w_r, the sampling or pulsing frequency.

When the signal to be sampled consists of a modulated carrier as shown in Figure 2.5.6.5 (a), the spectrum of the sampled signal is that shown in Figure 2.5.6.5 (b). Here, the spectrum is identical to that shown in Figure 2.5.6.2 (b) for a time-varying signal with no carrier, except that the entire spectrum is shifted upward in frequency by w_o, the carrier frequency. Now the spectrum of the modulating signal appears aliased and folded about frequencies above and below w_o, the carrier frequency, at intervals of w_r, the sampling frequency. This waveform and its corresponding spectrum are especially important because they represent exactly the situation encountered in a pulse Doppler radar.

2.5.7 Signal Correlation -- The time cross-correlation function, $R_{e1,e2}(\tau)$, for two signals, e_1 (t) and e_2(t) is defined by the equation:

(a) Time Domain



(b) Frequency Domain

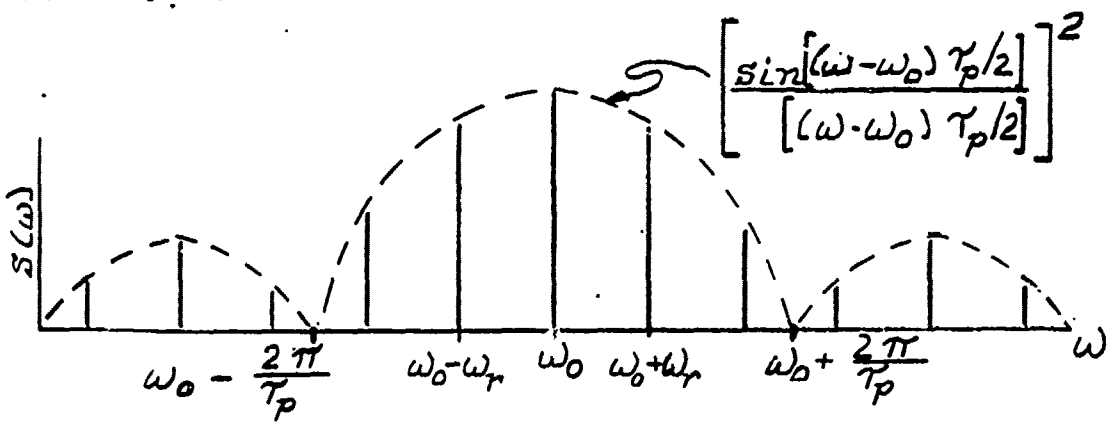
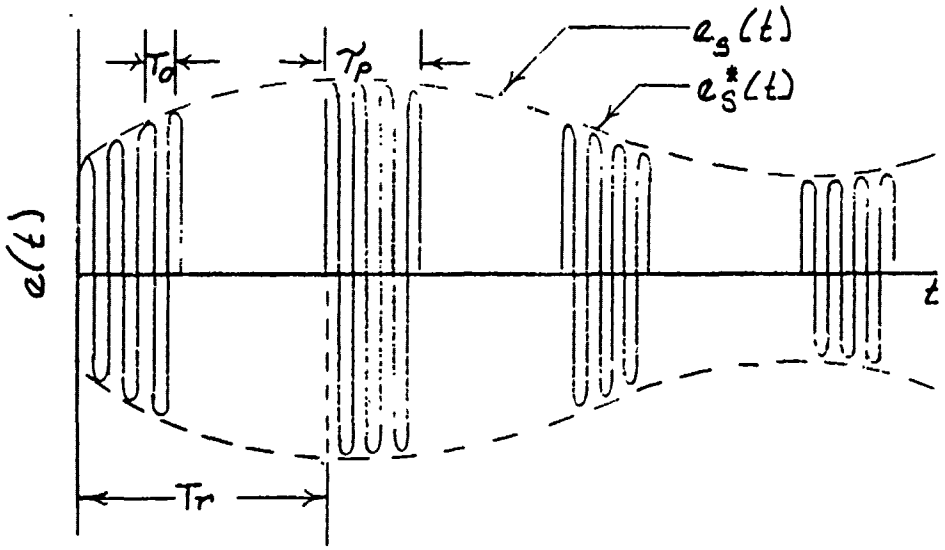


Figure 2.5.6.4 -- Pulsed Non-Time-Varying Waveform with Carrier

(a) Time Domain



(b) Frequency Domain

ω_{sm}=Max. Signal Frequency

$\omega_o = \frac{2\pi}{T_o}$, $\omega_r = \frac{2\pi}{T_r}$

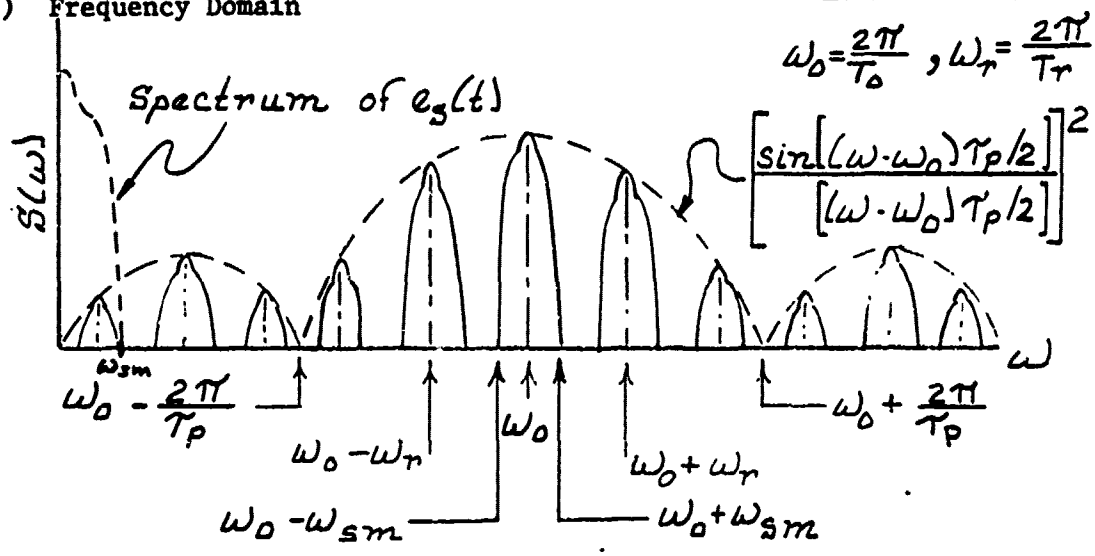


Figure 2.5.6.5 -- Pulsed Time-Varying Waveform with Carrier

$$R_{e1,e2}(\tau) = \lim_{T\to\infty}\left\{\frac{1}{2T}\int_{-T}^{+T} e_1(t)\, e_2(t+\tau)\, dt\right\} \qquad [2.5.7.1]$$

That is, the correlation function is the long-term time average of the product of the two signals. For practical applications, of course, the integration limit, T, is finite. The correlation between two signals is a measure of the coherence or functional dependence between those signals. The time-delay parameter, τ, allows a comparison of the two signals as a function of a variable difference in time of observation. A zero correlation coefficient for a particular value of τ indicates no functional dependence for the two signals, for that particular difference in time-of-observation. A non-zero value for $R(\tau)$, either positive or negative, indicates functional dependence. Thus, for example, if $e_1(t)$ and $e_2(t)$ represented transmitted and received radar signals, respectively, one would expect non-zero correlation for a value of τ equal to the two-way time of flight for a pulse to the target and back.

If $e_1(t)$ and $e_2(t)$ are the same signal, $R(\tau)$ is the autocorrelation function. That is:

$$R_{e1}(\tau) = \frac{1}{2T}\int_{-T}^{+T} e_1(t)\, e_1(t-\tau)\, dt \qquad [2.5.7.2]$$

It should be noted that the correlation between two sinusoids of different frequencies is zero. It should also be noted that the correlation between Sin [wt] and Cos [w (t-τ)] varies as Sin (τ), and is therefore zero for $\tau = 0$. Further, it should be noted that the correlation between a random (non-deterministic) signal and any other signal is zero. The block diagram

for a practical signal correlation detector is shown in Figure 2.5.7.1. The low-pass filter approximates the function of an integrator. Its time constant determines the integration period, T, and must be large in comparison with the maximum periods associated with the signal.

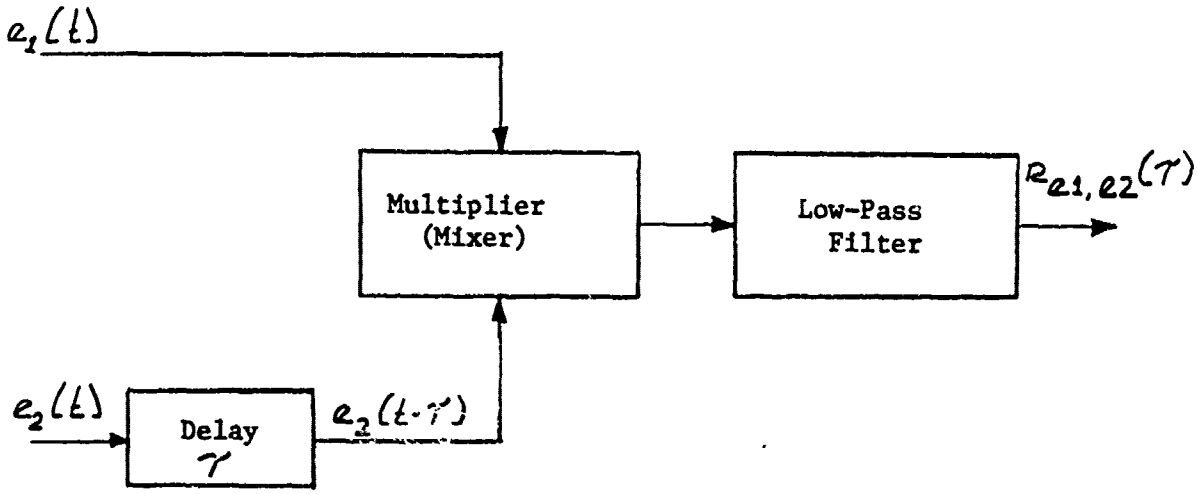


Figure 2.5.7.1 — Signal Correlator

2.5.8 Spread Spectrum Signals -- In Section 2.5.5 of this text, it is demonstrated that an information-carrying signal with a given bandwidth can be conveyed by a communications system with a bandwidth capability approximately equal to the information signal bandwidth, (e.g., by means of single-sideband amplitude modulated transmission). Under most circumstances, conservation of system and signal bandwidth is desirable, in order to conserve power and make efficient use of the available RF spectrum. There are situations, however, where it is desirable to expand the bandwidth of a signal artifically; that is, for reasons unrelated to the information bandwidth. Such methods are called spread-spectrum techniques.

There are a number of possible reasons for the use of spread spectrum signals. There are also a number of methods of generating them. Some of these reasons and methods are best considered in the time domain. Others are best considered in the frequency domain. Always, however, it should be noted than any signal processing in the time domain is accompanied by effects in the frequency domain, and vice versa. In particular, all of the processes classified as spread spectrum techniques, whether time oriented or frequency oriented, result in an extensive, artificial broadening of the signal bandwidth.

The principal purposes of spread spectrum processing are those listed below.

Encryption of Information
Covert (Undetected) Communication
Selective Addressing
Multiplexing of Information
Rejection of Background Noise
Rejection of Jamming
High Resolution Radar Ranging

All methods of spectrum spreading apply some sort of more-or-less complex modulation to the transmitted signal. The result is a time and frequency "scrambling" of that signal. The original signal can be recovered only by reversing the process, which requires a knowledge of the original modulation applied. Thus, spread spectrum processing can be employed to encrypt information.

Covert communication refers to undetected communication rather than encrypted information. By spreading the signal power over a broad frequency band, spread spectrum processing reduces the power density in any given region. If spread enough, the signal spectrum can be reduced below the level of the background noise. When the spreading process is reversed, the signal spectrum is de-spread and the signal-to-noise ratio is restored. Other spread spectrum techniques continually change the carrier frequency or time-of-transmission according to a complex schedule, thus making it difficult for an unprogrammed receiver to detect the transmission.

By encoding (modulating) each message with a unique code, multiple receivers can simultaneously be selectively addressed using a single communication channel. Each addressee decodes the transmission using his own code, thereby recovering only his intended message. The result of the "addressing" operation, in the frequency domain, is to spread the spectrum of the signal.

Multiple information signals can simultaneously be transmitted to a single receiver utilizing the same techniques employed for selective addressing. The various information signals are separated at the receiver by demodulating

the transmitted signal using multiple demodulators, each with its own code.

An important result of spread spectrum processing is the great improvement in the signal-to-noise ratio that can be achieved by the spectrum "de-spreading" process. When a spread spectrum signal is de-spread (demodulated) utilizing the unique code (signal) previously used to spread (modulate) it, the result is to re-concentrate the signal into a relatively narrow frequency band. Any external interference accompanying the spread spectrum signal, however, has not been modulated by the encoding signal. The effect of the de-spreading process on the (unspread) interference is, then, to spread it. When the de-spread signal is passed through a narrow-band-pass filter, only a small part of the (spread) interference power is passed. The signal-to-noise ratio after spread spectrum processing to that before processing is called the process gain. That is, the process gain, G_p, is defined as:

$$G_p = \frac{(S/N)_{out}}{(S/N)_{in}} \qquad [2.5.8.1]$$

In terms of the signal bandwidths involved, the process gain is given by:

$$G_p = \frac{[BW]_{ss}}{R_s} \qquad [2.5.8.2]$$

where [BW]ss is the bandwidth of the spread spectrum signal in Hertz and R_I is the information signal information rate in bits per second. Spread

spectrum techniques are commonly used in the processing of the signals from communications satellites because of the small signal-to-noise ratios involved.

The interference referred to above can be natural background noise or it can be intentional jamming. For the case of a jamming signal, the process gain is defined as:

$$G_p = \frac{(S/J)_{out}}{(S/J)_{in}} \qquad [2.5.8.3]$$

where (S/J) is the signal-to-jamming power ratio. Spread spectrum processing is a powerful ECCM (Electronic Counter-Counter Measures) technique because it does not require detailed information concerning the jamming signal. It is most effective against narrow band noise jamming since wide band noise is already "spread" to some extent. (Spread spectrum processing would be ineffective against true white noise.)

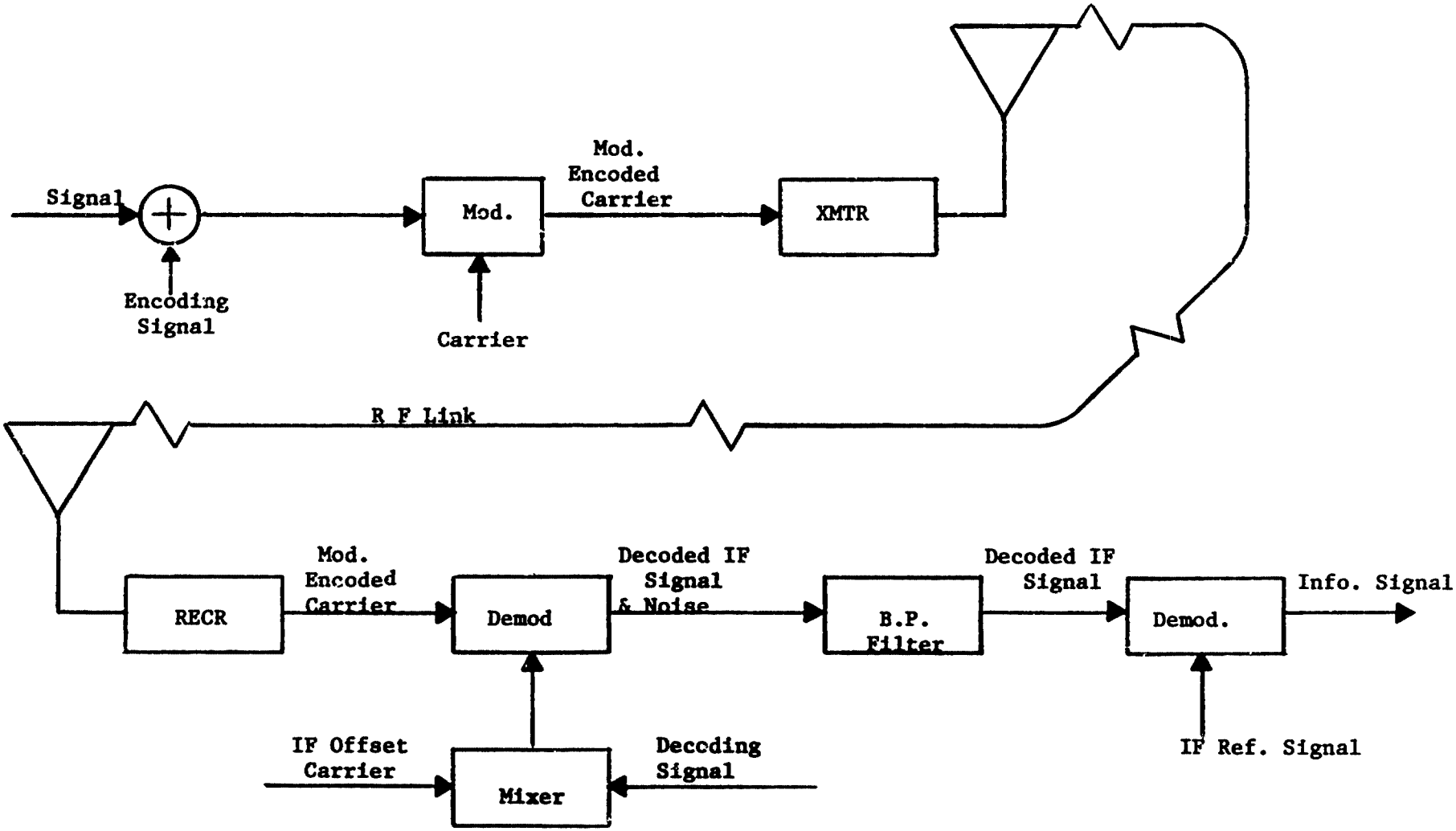
One of the most important applications of spread spectrum signal processing is in radar ranging. The range resolution of a pulsed radar, (without spread spectrum processing), is determined by the pulse width. The pulse width determines how closely spaced in range two targets can be and still produce returns that can be resolved. (See Section 2.3.3 of the text on radar systems for a discussion of radar range resolution.) By providing "time coloring" within the pulse, spread spectrum processing allows target echo time-of-return (and hence target range) to be determined to a finer

resolution than possible with an uncolored pulse. The "time coloring" of the pulse results in spreading the radar signal spectrum. For a discussion of pulse time-coloring techniques, see Section 2.10 of the text on radar systems.

The principal spread spectrum techniques are the following.

Direct Sequence Modulation
Frequency Hopping
Swept Frequency Modulation
Time Hopping

Direct sequence modulation entails modulation of the carrier by a combination of the information signal and a pseudo-random encoding (spectrum spreading) signal, as shown in Figure 2.5.8.1. The encoded (spread spectrum) carrier is then transmitted to the receiver. In the decoder, a signal, at the carrier frequency offset by the intermediate frequency, is combined with the decoding signal (identical to the encoding signal) and the combination is used to de-spread the received signal and reduce it to intermediate frequency. The decoded signal plus noise is then passed through an IF band-pass filter to remove all signal components outside the bandwidth of the original information signal. Finally, the decoded signal, (still at intermediate frequency), is demodulated to remove the IF carrier and recover the original information signal. The name "direct sequence" modulation refers to the fact that, in most modern systems, the encoding/decoding signal is a sequence of binary digits. An analog encoding signal can also be used. The most difficult aspect of direct sequence spread spectrum communications is the synchronization of the decoding signal



2.112a

Figure 2.5.8.1 -- Direct Sequence Spread Spectrum Communication System

with the encoding signal modulating the spread spectrum carrier. Various schemes employing pattern-recognition devices are used for this purpose.

The diagram in Figure 2.5.8.1 shows the information signal and encoding signal combined by binary (modulo-2) addition. This process, of course, assumes that the encoding signal is a binary sequence and that the information signal has been binary encoded. In modulo-2 addition, 0 + 1 = 1 and 1 + 1 = 0.

To demonstrate that modulo-2 addition can be used for both encoding and decoding, consider the case of a binary information signal, 101010, being encoded (and decoded) by a binary sequence, 100110. Thus, 101010 is modulo-2 added to 100110 to yield the encoded signal as follows.

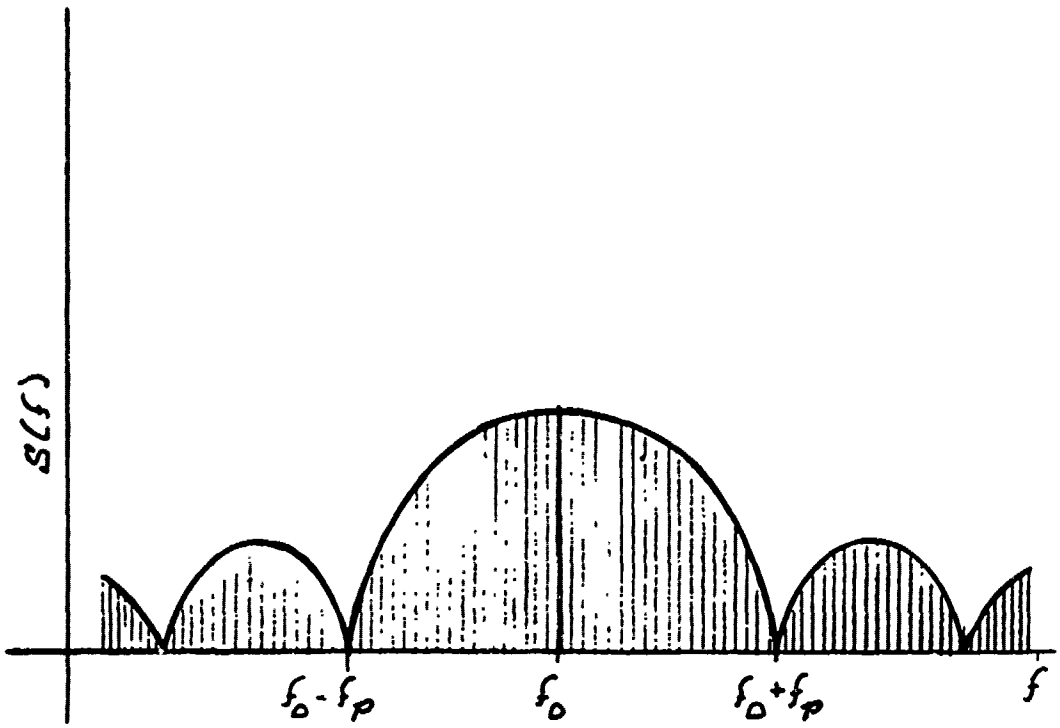
```
  101010    (Information Signal)
+ 100110    (Encoding Signal)
  001100    (Encoded Signal)
```

The encoded signal is thus the binary word (sequence) 001100. The encoded signal, 001100, is decoded by modulo-2 adding it to the decoding (and encoding) signal, 100110, thus:

```
  001100    (Encoded Signal)
+ 100110    (Decoding Signal)
  101010    (Decoded Signal)
```

The result of the second modulo-2 addition is seen to be the original information signal, as required.

The spectrum of a direct sequence spread spectrum signal is shown in Figure 2.5.8.2 and is seen to resemble that shown in Figure 2.5.6.5 for a pulse modulated carrier.

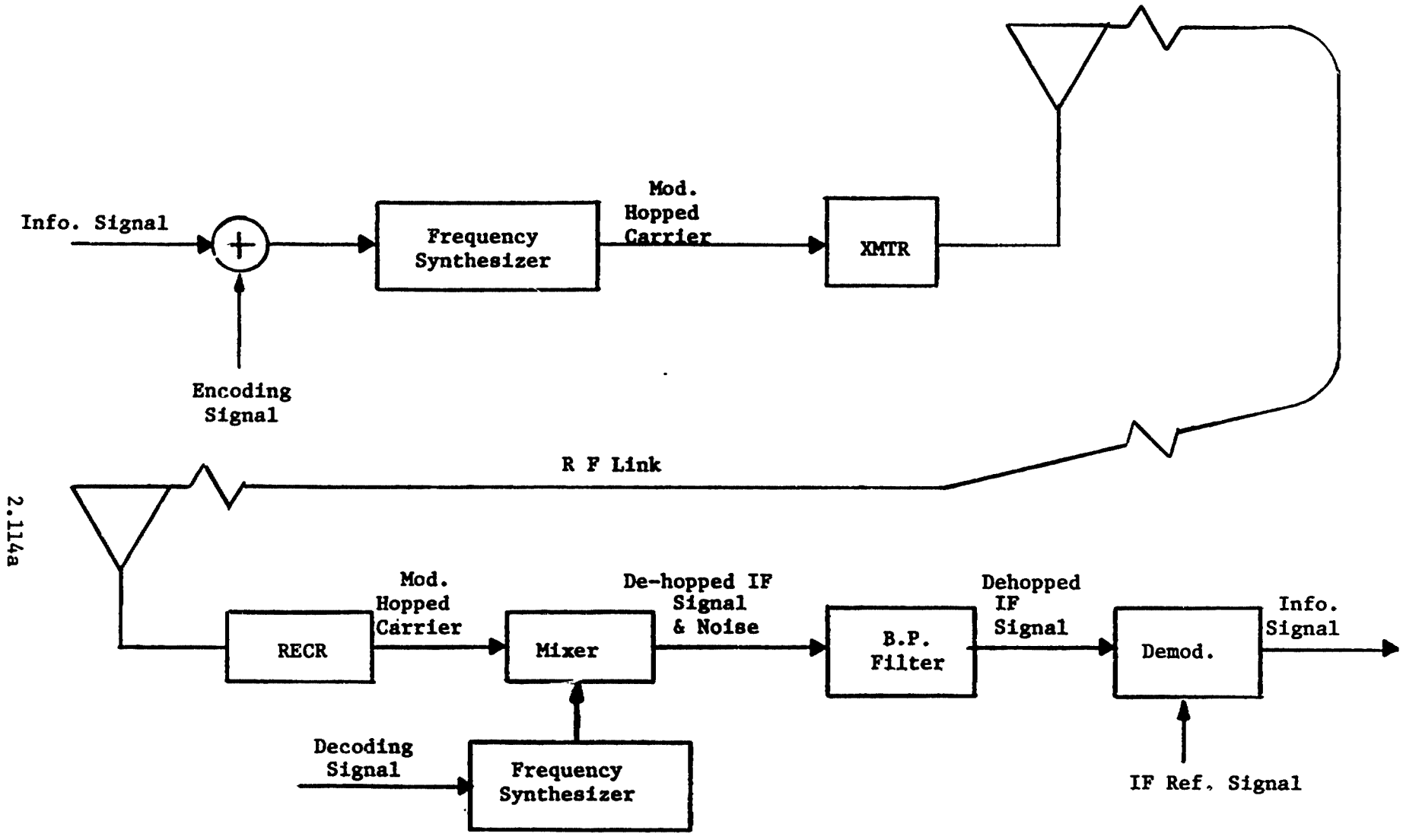


f_p = Pulse Sequence Rate

Figure 2.5.8.2 -- Spectrum of Direct Sequence Spread Spectrum Signal

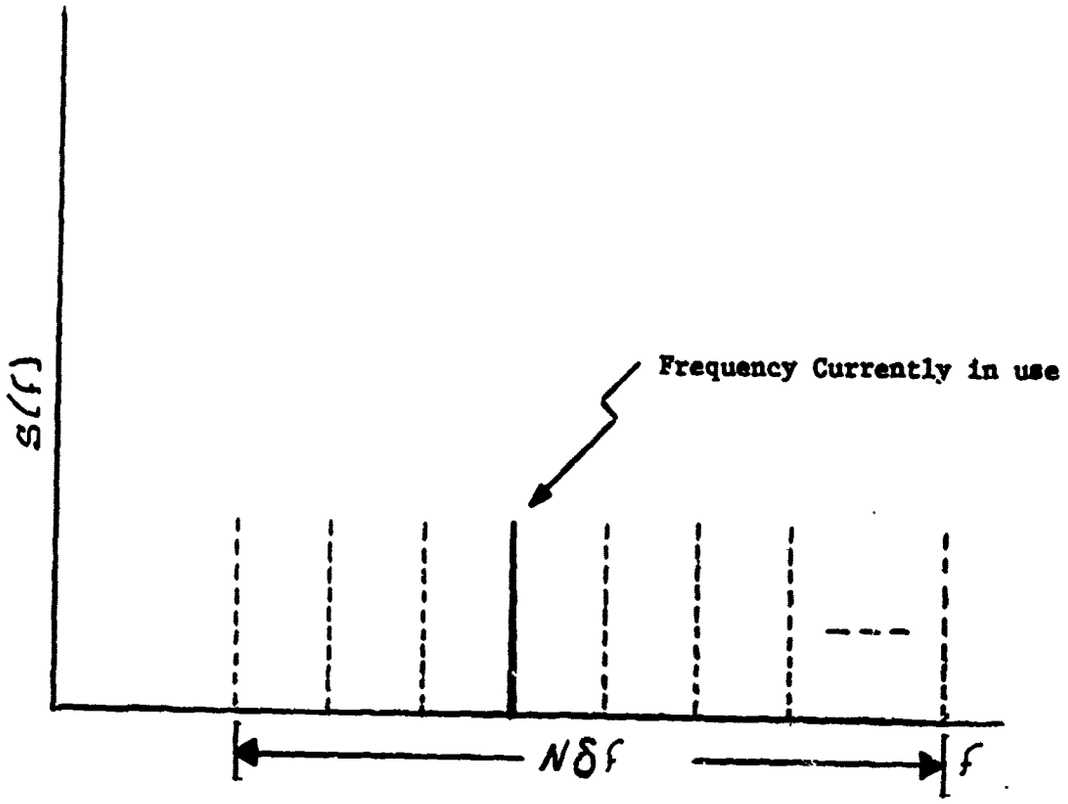
Frequency hopping is, as the name implies, a system by means of which a spread spectrum is attained by abruptly shifting the carrier frequency of the signal to one of numerous discrete values within a prescribed band, in accordance with a pseudo-random sequence similar to the encoding signal used in direct sequence modulation. This process is shown in the block diagram of Figure 2.5.8.3. The frequency hopping sequence and the information signal are combined and the resulting digital signal is used to drive a digitally controlled frequency synthesizer. The result is an information-carrying frequency-hopping carrier. The spread spectrum carrier is transmitted to the receiver. In the decoder, the decoding signal, identical to the previously employed encoding signal, is used to control a frequency synthesizer, the output of which is mixed with the incoming signal to de-hop the carrier and reduce it to intermediate frequency. The IF signal is then passed through a band-pass filter and demodulated to recover the information signal. Ths spectrum of a frequency-hopping spread spectrum signal is shown in Figure 2.5.8.4. The number of discrete carrier frequencies for an actual system would be many times the number depicted and the frequency aliasing effects of pulse modulation would create additional frequencies.

The swept frequency spread spectrum system attains a broad spectrum by sweeping the frequency of the carrier during the pulse interval. As shown in Figure 2.5.8.5, the chirp program signal controls the frequency of a voltage controlled oscillator, the output of which is the chirped carrier. The chirped carrier is then modulated by the information signal. The modulated,



2.114a

Figure 2.5.8.3 -- Frequency Hopping Spread Spectrum Communication System



N = Number of Frequencies Available

δf = Frequency Spacing

Figure 2.5.8.4 — Spectrum of Frequency Hopping Spread Spectrum Signal

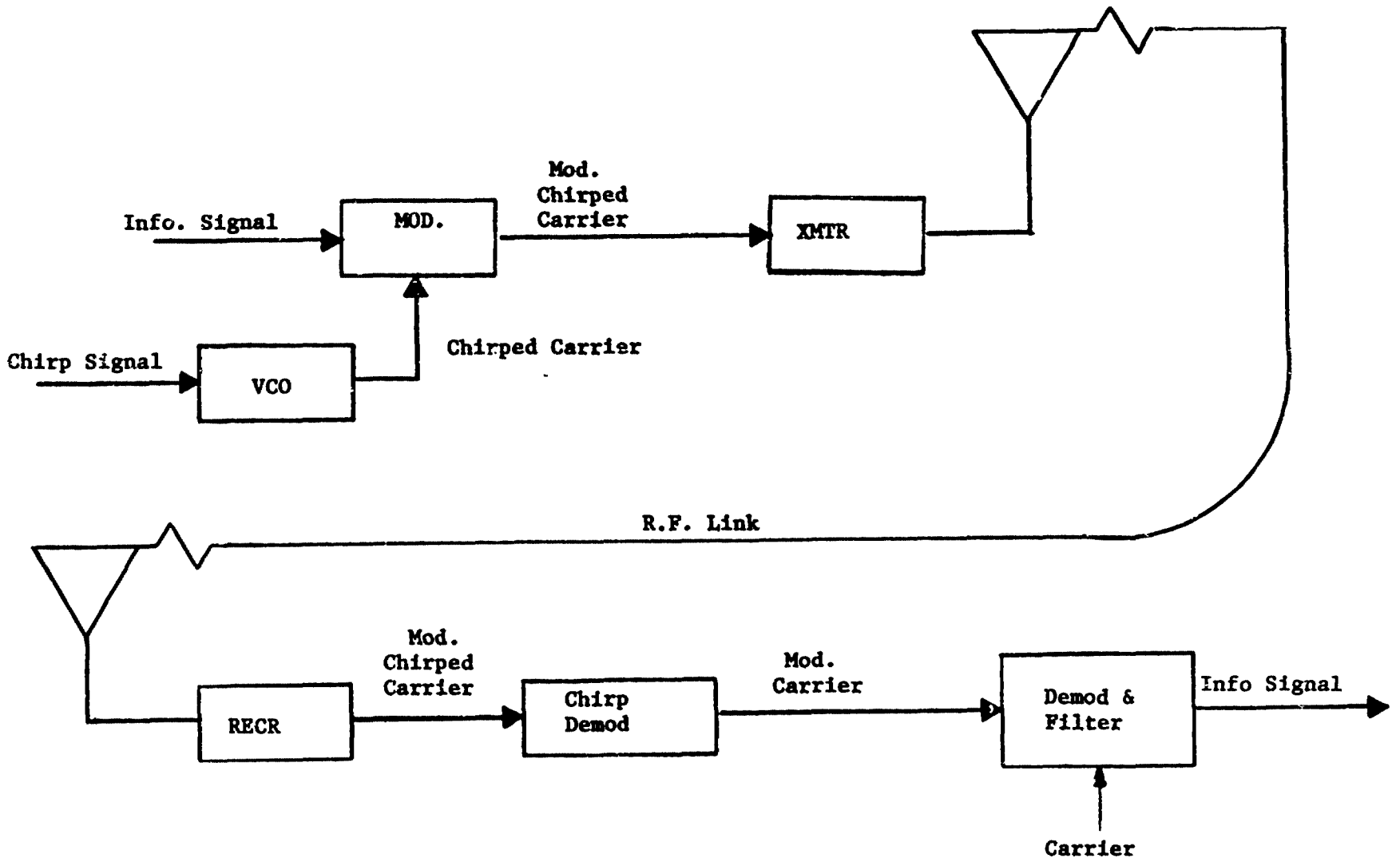


Figure 2.5.8.5 -- Swept Frequency Spread Spectrum Communication System

chirped carrier is transmitted to the receiver. At the receiver, the transmitted signal is de-chirped, (usually by passing it through a dispersive delay line). The de-chirped carrier is then demodulated and filtered to yield the information signal.

Considered in the frequency domain, the swept frequency yields the same advantages as those offered by frequency hopping. In the time domain, the swept frequency time colors the transmitted signal, thus allowing high-resolution time-correlation measurements. Swept frequency modulation is the technique employed for pulse compression radar ranging as discussed in Section 2.10 of the text on radar systems. The same technique is also used in other communications applications.

A time hopping spread spectrum system abruptly changes pulse transmission time and period in a manner similar to that in which a frequency hopping system changes carrier frequency. The block diagram for a time hopping system is shown in Figure 2.5.8.6. The information signal is combined with a pseudo-random encoding signal and used to modulate a carrier. The carrier is then switched on and off in a pseudo-random manner controlled by the encoding signal. (The system as shown is both direct sequence modulated and time hopped). The modulated, time hopped carrier is transmitted to the receiver. After reception, the transmitted signal is dehopped by an RF switch driven by the same pseudo-random sequence used to encode the signal. It is then reduced to IF frequency, filtered, and demodulated to yield the information signal.

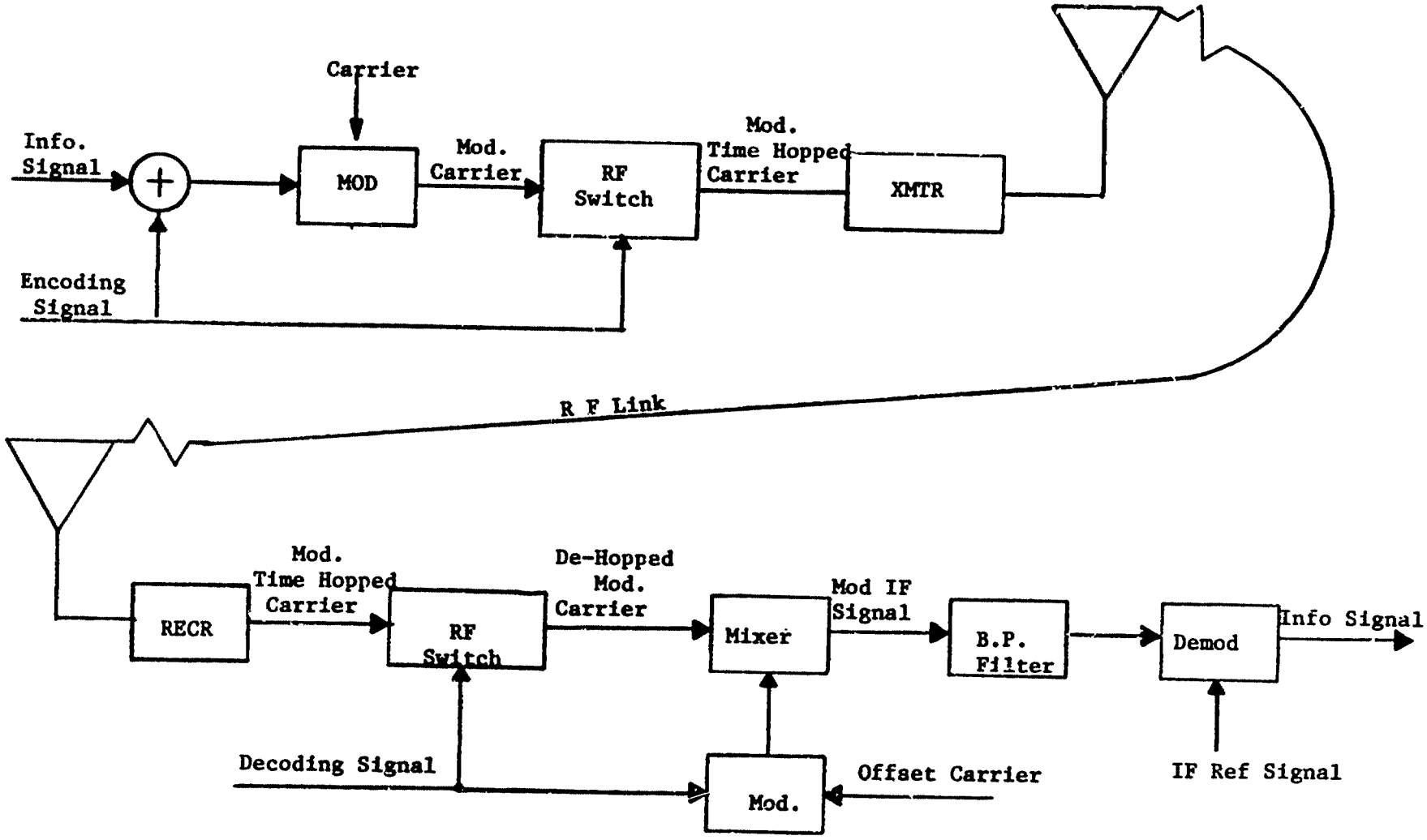


Figure 2.5.8.6 -- Time Hopping Spread Spectrum Communication System

2.5.9 Communication System Noise -- In all systems involved with the transmission and processing of information, it is ultimately the signal-to-noise ratio, rather than absolute signal magnitude, that limits system performance. Noise sources are of two types: those internal to the system and those external. Internal noise sources include contact noise, shot noise, generation/recombination noise, induction noise, and thermal noise.

Contact noise is caused by fluctuations in the contact resistance of internal current-carrying interfaces. The RMS value of the contact noise voltage is given by:

$$V_{cn} = K\ I\ R\ [\Delta f / f]^{1/2} \qquad [2.5.9.1]$$

where K is a constant depending upon the material, I is the average current flowing through the contact, R is the nominal contact resistance, f is the frequency of the noise, and Δf is the bandwidth over which the noise is measured. Due to the fact that contact noise power is inversely proportional to the frequency, it is sometimes called "1/f" noise. It is also called current noise, excess noise, and modulation noise. Contact noise is generally the dominant noise in semiconductor circuits at very low frequencies and negligible above about 100 hertz.

Shot noise is produced by fluctuations in the rate of arrival of electrons at internal junctions and is given by:

$$V_{sn} = [2\ I\ R^2\ g_e\ \Delta f]^{1/2} \qquad [2.5.9.2]$$

where I is the average current flowing through the junction, R is the resistance of the junction, q_e is the charge on an electron, and Δf is the bandwidth of noise measurement. Shot noise is generally dominated by other noise sources.

Generation/recombination noise is a result of the random production and destruction of charge carriers (free electrons and holes) in semiconductor materials. The RMS value of the noise voltage due to generation/recombination is given by:

$$V_{grn} = 2 I R \left[\frac{\tau \Delta f}{N(1+4\pi^2\tau^2 f^2)} \right]^{1/2} \qquad [2.5.9.3]$$

where I is the average current, R is the resistance, τ is the carrier lifetime, N is the total number of carriers, f is the frequency of the noise voltage, and Δf is the bandwidth of the noise measurement. Generation/recombination noise is generally dominated by other noise sources, especially at high frequencies.

Induction noise is generated by internal inductive and capacitive induction (pickup). The induced noise voltage depends upon the degree of unwanted coupling within the circuit and is generally reduced by electric and magnetic shielding to a value less than that due to other noise sources.

Thermal noise is produced by the random thermal motion of charged particles in electronic devices. The RMS value of the thermal noise voltage is given by:

$$V_{tn} = \left[4 k T_o R \Delta f \right]^{1/2} \qquad [2.5.9.3]$$

where k is Boltzmann's constant; T_o is the reference absolute temperature (usually taken as 290°, Kelvin); R is the noise source resistance; and Δf is the bandwidth of noise measurement. Thermal noise, sometimes called Johnson noise or resistance noise, is always present at temperatures above absolute zero and, for that reason, is dominant when all other noise sources have been eliminated. Thus, thermal noise is the ultimate limiting factor in achieving a large signal-to-noise ratio. Because of its prevalence, thermal noise is used as a standard against which to measure the noise levels in systems. For example, the noise present at the output of an amplifier is often stated in terms of the "noise figure", F_n, for the amplifier, defined by the equation:

$$F_n = \frac{V_n^2}{V_{tn}^2} \qquad [2.5.9.4]$$

where V_n is the actual noise voltage and V_{tn} is the thermal noise, defined by Equation [2.5.9.3], for a given reference temperature and for a bandwidth equal to the effective noise bandwidth of the amplifier. Both noise voltages are equivalent values, referred to the input. Thus, a device in which all noise sources except thermal noise at the input have been eliminated would have a noise figure of unity. The noise figure can be expressed in terms of the output rms noise voltage, $V_{n(out)}$, by the equation:

$$F_n = \frac{V_{n(out)}^2}{V_{tn}^2 G_p} \qquad [2.5.9.5]$$

where G_p is the power gain of the device and V_{tn} is given by Equation [2.5.9.3].

It is also possible to express the noise figure for a device in terms of the input and output signal-to-noise ratios. Thus:

$$F_n = \frac{(S/N)_{in}}{(S/N)_{out}} \qquad [2.5.9.6]$$

where (S/N) is the signal power-to-noise power ratio. It also is common practice to specify noise sources in terms of equivalent "noise temperatures". That is, a noise generating device with noise voltage V_n can be specified as having an "effective noise temperature", T_e, given by:

$$T_e = \frac{V_n^2 - V_{tn}^2}{4kR\Delta f} \qquad [2.5.9.7]$$

where all quantities are as defined for Equations [2.5.9.3] and [2.5.9.4]. Again, all quantities are referred to the input. Effective noise temperature is a measure of the noise generated internally by the device; that is, a measure of the noise at the output in excess of that due to thermal noise at the input. In terms of the noise figure, F, the effective noise temperature is given by:

$$T_e = T_o (F_n - 1) \qquad [2.5.9.8]$$

The bandwidth of a device was defined in Section 2.5.3 of this text as the frequency span between the one-half power points on the response (gain) curve. (See Figure 2.5.3.3). For many purposes, however, another definition of bandwidth is more meaningful. That definition is based upon the total power contained in the response of the device to white noise and yields the "equivalent noise bandwidth". The equivalent noise bandwidth, B_n, is defined by the equation:

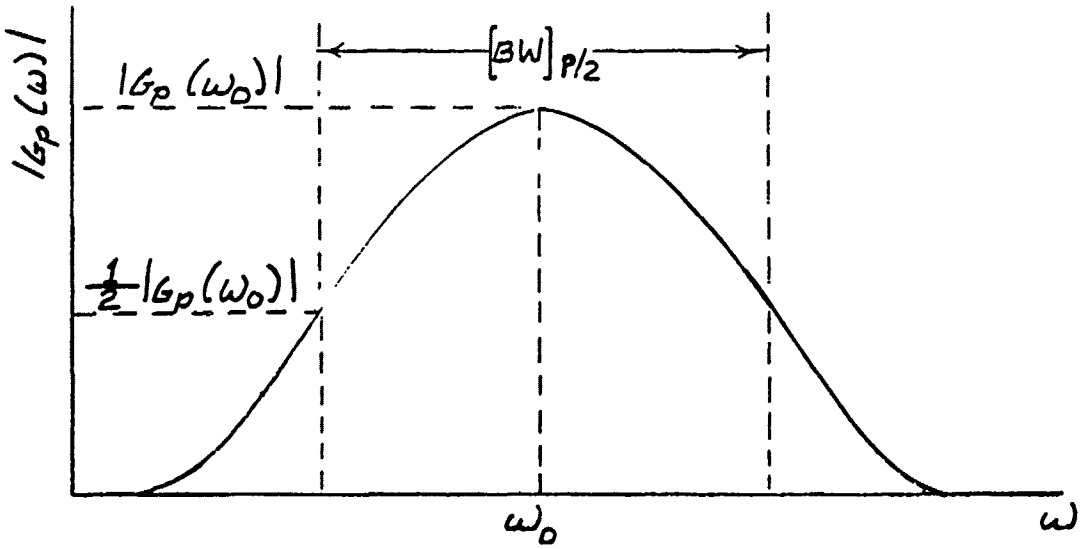
$$B_n = \frac{\int_0^\infty |G_p(\omega)|\, d\omega}{|G_p(\omega_o)|} \qquad [2.5.9.9]$$

where $/G_p(w)/$ is the magnitude of the power gain of the device as a function of frequency, w, and $/G_p(w_o)/$ is the value of $/G_p(w)/$ at the response peak. Both the one-half power and the equivalent noise bandwidths are illustrated in Figure 2.5.9.1.

The term "white noise", as used in the definition of equivalent noise bandwidth, refers to noise containing a uniform distribution of power at all frequencies. That is, the spectrum (power spectral density), S(w) of the noise is that shown in Figure 2.5.9.2(a). Because white noise would require an infinite total power, it is physically unrealizable. Often, however, band-limited white noise, as shown in Figure 2.5.9.2(b), is employed in the communications field.

External noise sources include atmospheric noise, galactic noise, and man-made noise. Atmospheric noise is due to lightning, corona discharges, and precipitation static. As shown in Figure 2.5.9.3, it is generally dominant

(a) One-Half Power Bandwidth



(b) Equivalent Noise Bandwidth

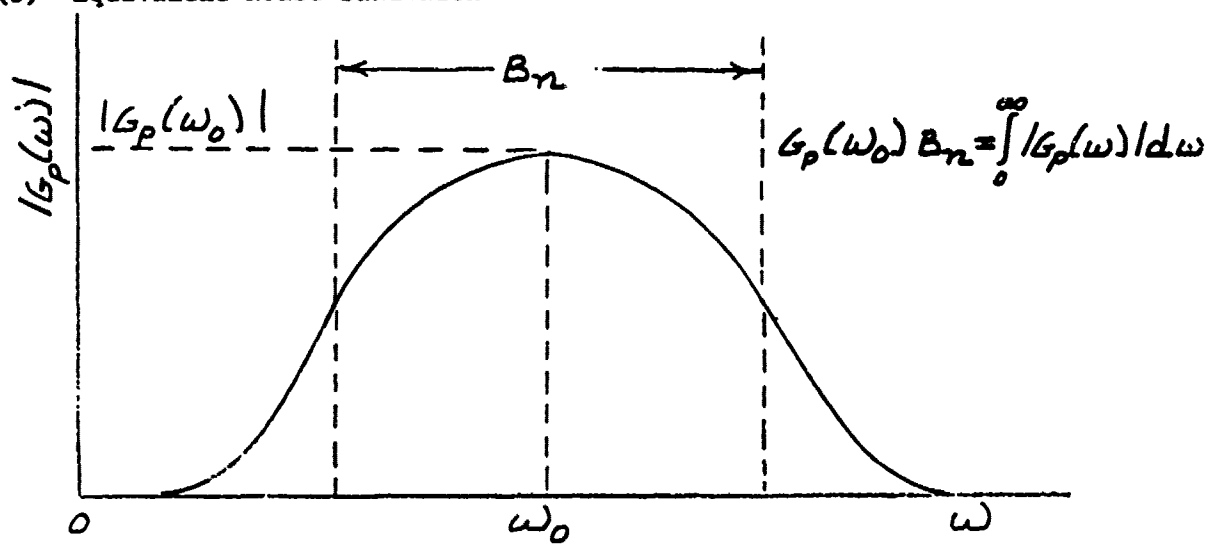
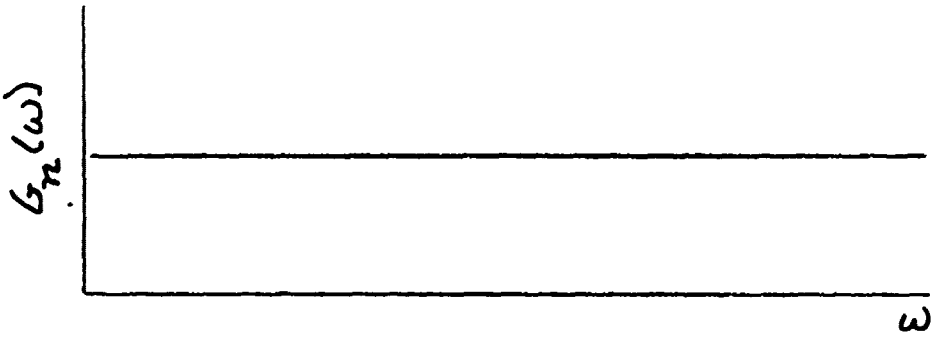


Figure 2.5.9.1 -- Bandwidth Definitions

(a) White Noise



(b) Band-Limited White Noise

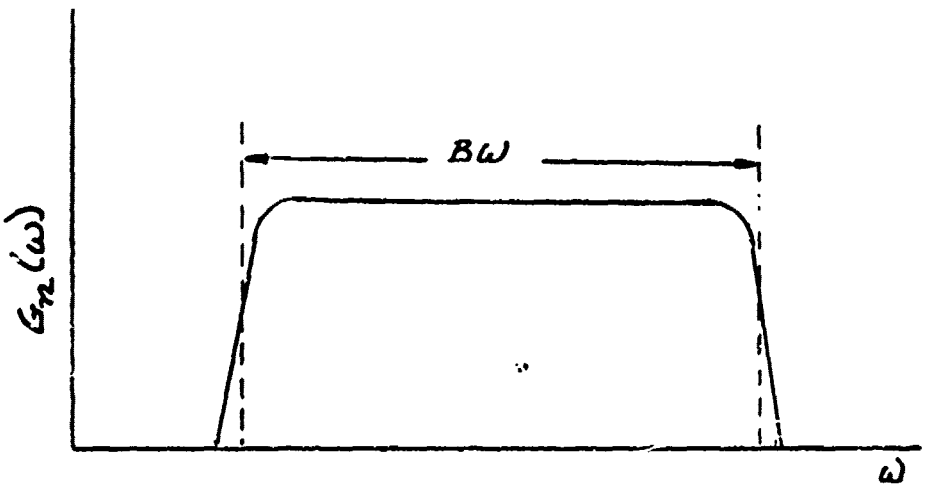


Figure 2.5.9.2 — White Noise Spectra

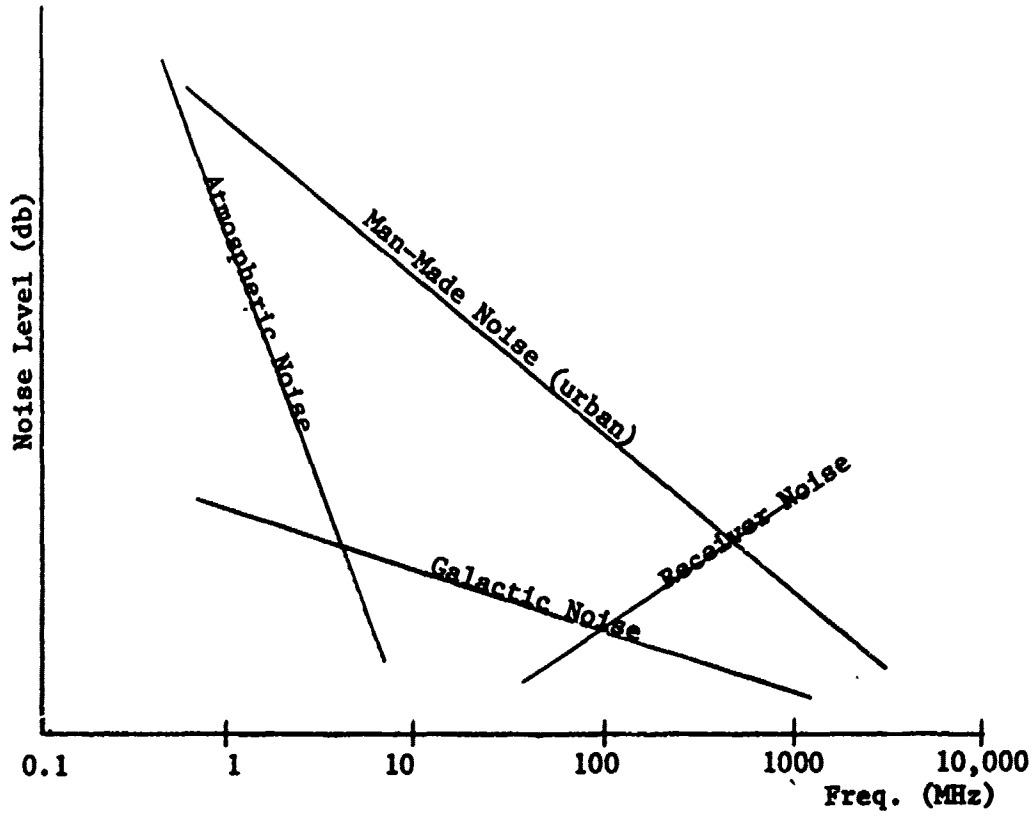


Figure 2.5.9.3 -- External Noise Source Spectra

at low frequencies. Galactic noise is produced by extra-terrestrial bodies, primarily the sun. Galactic noise is not usually a problem for terrestrial applications due to its relatively small magnitude. In the vicinity of man-made electrical and electronic equipment, man-mode noise can be dominant. The curve in Figure 2.5.9.3 shows the man-made noise level in a typical urban location. Even removed from populated areas, however, interfering signals from other transmitting equipment can be a problem. For purposes of comparison, the internal noise of a typical receiver is shown in the figure. As indicated, it becomes dominant at very high frequencies. Receiver noise is thermal noise and increases with frequency because the internal losses in a receiver increase with frequency, thereby raising the effective noise temperature.

As previously stated, the important measure of signal strength in a communications system is the signal-to-noise ratio. For that reason, the elimination of noise is a major objective of communication system design. Several general methods of noise reduction are available. One is to avoid the low frequency end of the spectrum. Except for internal receiver noise, all internal and external noise levels decrease with increasing frequency. Another method of noise reduction is to employ frequency modulation or spread spectrum techniques. As previously mentioned, the broad, complex spectra associated with those techniques make it relatively easy to distinguish between signal and noise. Another possibility is the use of digital signals. The natural noise resistance of digital systems, and the use of error correcting codes affords digital systems a high noise-rejection capability. Signal correlation techniques, as discussed in Section 2.5.7 of this text, were specifically devised to provide signal recognition in the presence of noise. Frequency filtering of the signal and

electric and magnetic shielding of hardware are, of course, standard noise reduction techniques. Ultimately, noise reduction depends upon some difference, in either the frequency or the time domain, between the signal and the noise. Any technique is useful that exploits that difference.

2.6 Antennas

2.6.1 Antenna Performance - Antenna characteristics often limit the performance of a communication system. This is especially true for airborne systems because of the limited space available for mounting an efficient antenna. Antenna efficiency and directivity require that the physical size of the elements be, (at least), an appreciable portion of a wave length. At the low end of the VHF band, (30 MHz), the wavelength is ten meters. In addition, the proximity of conducting surfaces, (creating reflected-wave interference), makes extremely difficult the design of an airborne antenna with an acceptable radiation pattern.

The principal function of an antenna is that of providing a suitable coupling between the transmission line (feeder) and the medium of propagation. That function involves not only impedance matching, but also energy transduction from electrical power to electromagnetic radiation. (An appropriate resistor at the end of the line would match the impedance of the line but the energy would be dissipated as heat rather than being radiated as electr magnetic waves.) Thus, the antenna should be considered an energy transducer as well as a device for directing electromagnetic waves.

2.6.2 Origin of Antenna Fields -- As stated in Section 2.3 of this text, an electrical charge creates an electric field in the surrounding space. If the charge is in motion, (an electrical current exists), a magnetic field is also created. When the current is time-varying, the electrical and magnetic fields interact, in accordance with Maxwell's equations, (as discussed in Section 2.3 of this text), to produce the propagating electric and magnetic fields described by Schrodinger's wave equations, (presented in Section 2.4 of this text). A

sinusoidally time-varying current, i(t), flowing in a conductor of incremental length ∂L, as shown in Figure 2.6.2.1, produces sinusoidal fields given by the equations:

$$i(t) = I_o \sin[\omega t] \quad \text{(Amperes)} \qquad [2.6.2.1]$$

$$\mathcal{E}_t(t) = \left(\frac{30\,\partial L\, I_o\, \lambda}{r^3}\right)\cos(\theta)\left\{\left(\frac{1}{2\pi}\right)\cos[\omega(t-r/c)]\right.$$
$$\left. -\left(\frac{r}{\lambda}\right)\sin[\omega(t-r/c)] - 2\pi\left(\frac{r}{\lambda}\right)^2\cos[\omega(t-r/c)]\right\} \quad \text{(Volts/Meter)} \qquad [2.6.2.2]$$

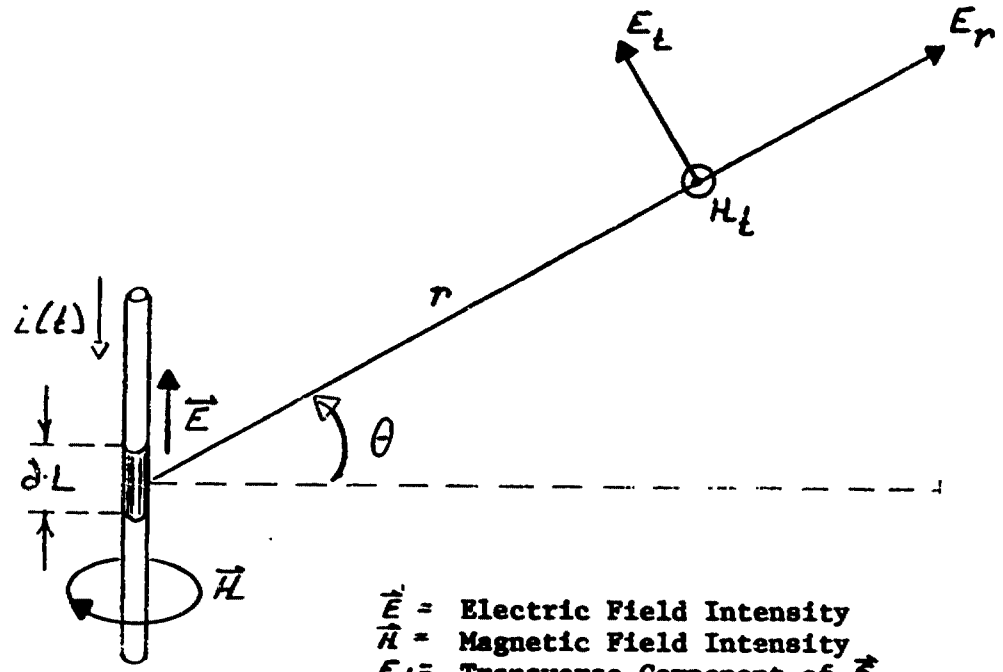
$$\mathcal{E}_r(t) = \left(\frac{60\,\partial L\, I_o\, \lambda}{r^3}\right)\sin(\theta)\left\{\left(\frac{1}{2\pi}\right)\cos[\omega(t-r/c)]\right.$$
$$\left. -\left(\frac{r}{\lambda}\right)\sin[\omega(t-r/c)]\right\} \quad \text{(Volts/Meter)} \qquad [2.6.2.3]$$

$$H_t(t) = \left(\frac{30\,\partial L\, I_o\, \lambda}{120\,\pi\, r^3}\right)\cos(\theta)\left\{\left(\frac{r}{\lambda}\right)\sin[\omega(t-r/c)]\right.$$
$$\left. -2\pi\left(\frac{r}{\lambda}\right)^2\cos[\omega(t-r/c)]\right\} \quad \text{(Volts/Meter)} \qquad [2.6.2.4]$$

As can be seen from Equations [2.6.2.2] through [2.6.2.4], the various terms of the field equations, (in braces), contain factors of $(1/2\pi)$, (r/λ), and $2\pi\,(r/\lambda)^2$. The field components due to terms containing $(1/2\pi)$ factors are called the static field; those containing (r/λ) factors the induction field; and those containing $2\pi\,(r/\lambda)^2$ the radiative field. At distances small in comparison with a wavelength, the static and induction fields are predominant. At distances large in comparison with a wavelength, the radiative fields are predominant. Two important conclusions can be drawn from Equations[2.6.2.2] through [2.6.2.4]:

(1) The higher the frequency (the shorter the wavelength), the more energy is radiated.

(2) At distances large in comparison with a wavelength, the electric and magnetic fields are transverse and vary as (1/r).



$\vec{E}$ = Electric Field Intensity
$\vec{H}$ = Magnetic Field Intensity
E_t = Transverse Component of $\vec{E}$
E_r = Radial Component of $\vec{E}$
H_t = Transverse Components of $\vec{H}$
$i(t)$ = Current in Element
∂l = Incremental Length of Element
r = Distance from Elemental Antenna
θ = Angle from Perpendicular to Element

Figure 2.6.2.1 — Electric and Magnetic Fields Produced by an Elemental Antenna

The region dominated by the static and induction fields is called the "near field"; that dominated by the radiative field is called the "far field". The boundary between the near and far fields is generally defined as:

$$R = \frac{2L^2}{\lambda} \text{ (Meters)} \qquad [2.6.2.5]$$

where L is the length (extent) of the conductor (antenna) and λ is the wavelength, as shown in Figure 2.6.2.2.

In the radiative (far field) region, the fields constitute spherically-spreading, transverse waves represented by the field equations shown in Figure 2.6.2.3. The electric field is in the plane containing the antenna element and the point of measurement. The magnetic field is perpendicular to that plane. The direction of propagation (energy flow) is indicated by the Poynting vector, $\overline{P}$, as shown. The field intensity pattern is the cos(θ) pattern indicated by the equations and shown, in polar coordinates, in the figure. The cos(θ) pattern actually represents a cross-section, in the plane of the antenna element, of a doughnut-shaped, three-dimensional radiation pattern.

2.6.3 Fields of Finite Antennas -- The fields produced by even the most complex antenna (or antenna array) can be determined by integrating (summing) the fields produced by incremental lengths of the antenna conducting elements. Expressions for the far-field electric and magnetic fields produced by an antenna can be derived by integrating the incremental field equations shown in Figure 2.6.2.3. By integrating these expressions along the paths of the antenna elements, the total fields can be obtained in accordance with the principle of linear superposition. In general, the amplitudes and phases of the currents in the incremental

Near & Far Fields

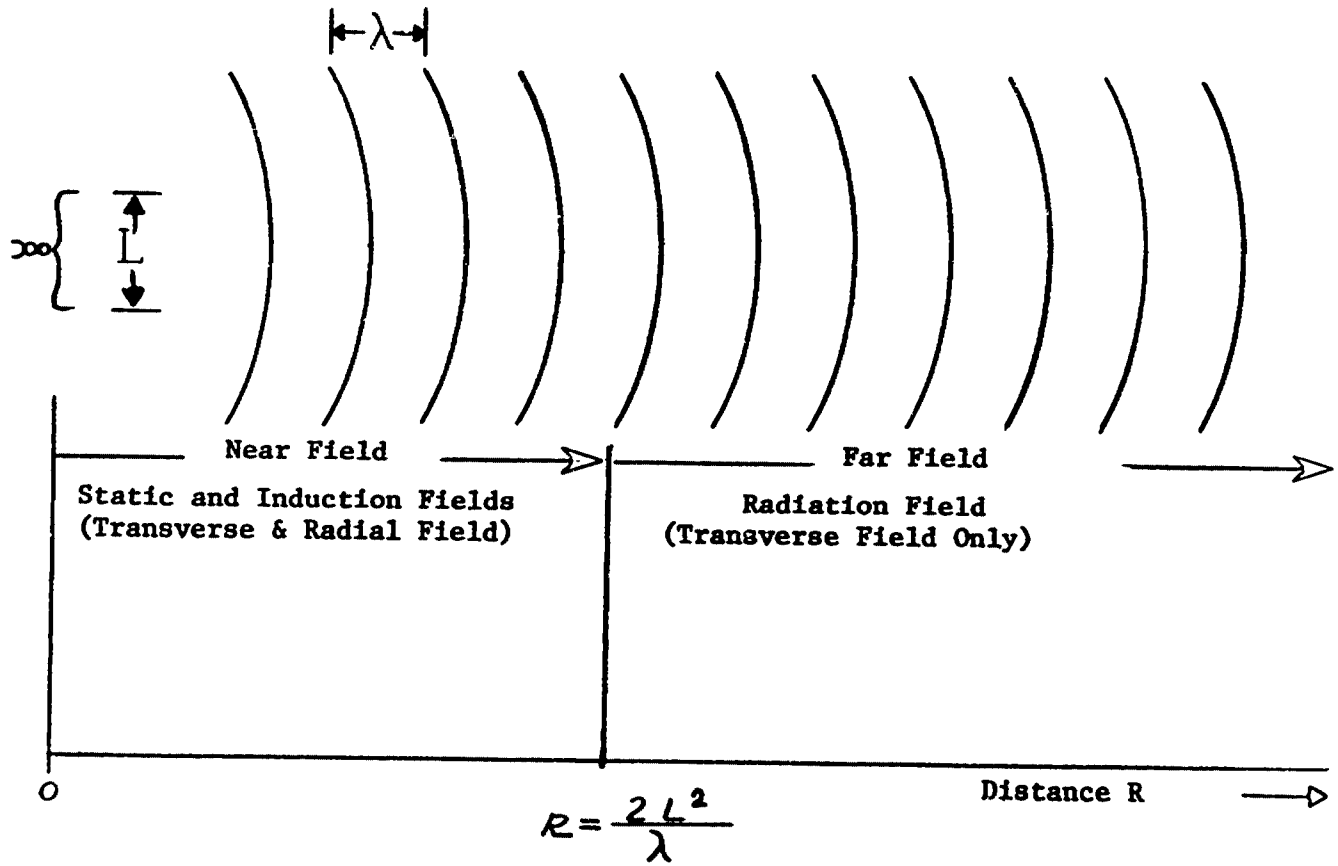
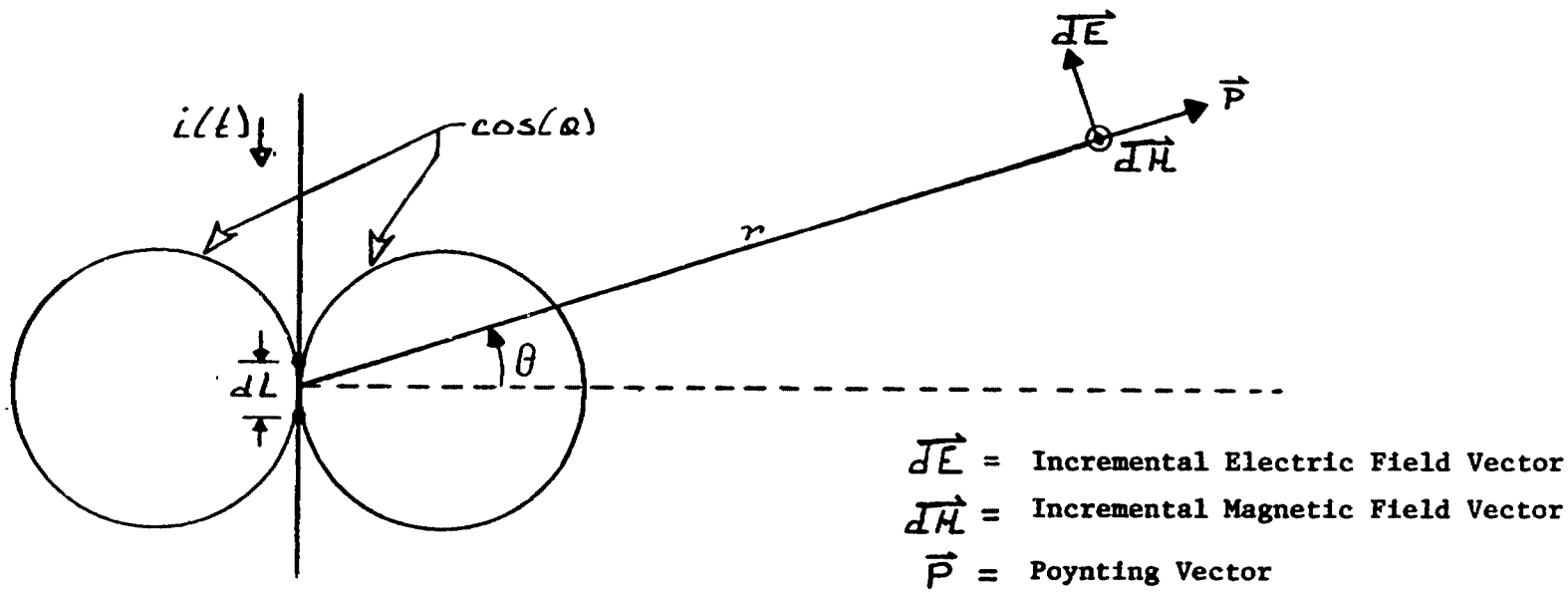


Figure 2.6.2.2 -- Near-and Far-Field Regions of an Antenna



$\vec{dE}$ = Incremental Electric Field Vector

$\vec{dH}$ = Incremental Magnetic Field Vector

$\vec{P}$ = Poynting Vector

$$i(t) = I_0 \sin(\omega t)$$

$$\vec{dE} = \left(\frac{60\pi I_0}{r\lambda}\right) \cos(\Omega) \cos[\omega(t - r/c)]\, dL$$

$$\vec{dH} = \left(\frac{I_0}{2r\lambda}\right) \cos(\Omega) \cos[\omega(t - r/c)]\, dL$$

Figure 2.6.2.3 -- Incremental Radiative Fields Produced by an Elemental Antenna

elements of an extended antenna will vary. The expressions for the incremental fields must be appropriately modified to reflect these variations.

2.6.4 Antenna Impedance -- When an antenna is excited by an alternating current, portions of the electric and magnetic fields alternately emanate from, and collapse back upon, the antenna. The antenna thus alternately stores energy in, and absorbs energy from, the fields; thereby exhibiting the characteristics of a device with a complex impedance (i.e. resistance, inductance, and capacitance). In general, the input impedance of an antenna, such as the dipole antenna shown in Figure 2.6.4.1, can be represented by a series R-L-C circuit as shown in the figure. The equation for the input impedance is, then:

$$Z_A = R_A + j\left(\omega L_A - \frac{1}{\omega C_A}\right) \text{ (Ohms)} \qquad [2.6.4.1]$$

where R_A, L_A, and C_A are the effective values of resistance, inductance, and capacitance for the antenna. It should be emphasized that these are effective values and, in general, depend upon the frequency of the driving current.

As can be seen from Equation [2.6.4.1], an antenna possesses a resonant frequency, given by:

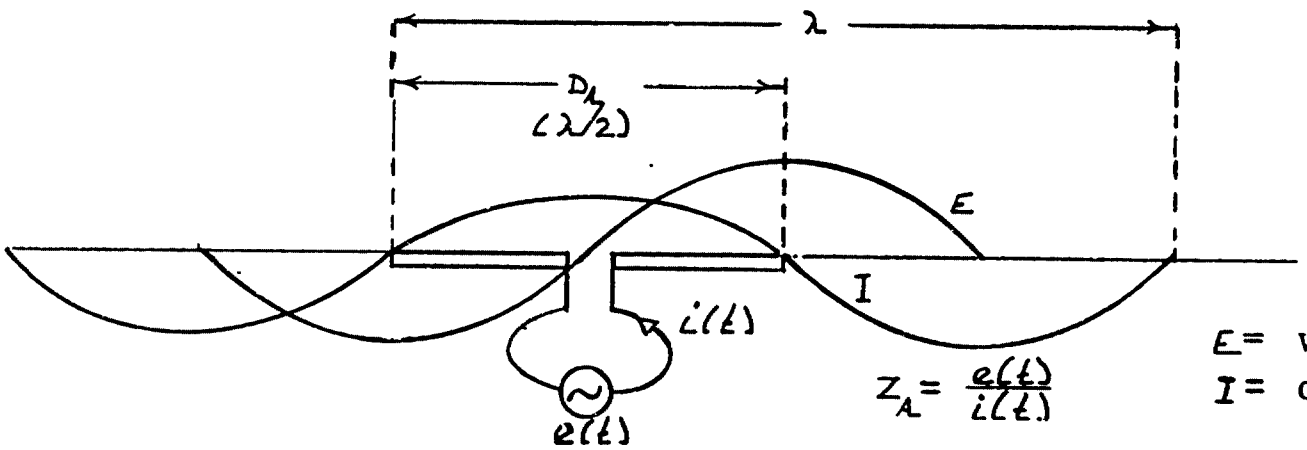
$$\omega_R = \frac{1}{\sqrt{L_A C_A}} \qquad \text{(Radians/Second)} \qquad [2.6.4.2]$$

At that frequency, the reactive portion of the input impedance vanishes and the input impedance becomes purely resistive. That is:

$$Z_A = R_A \quad \text{(Ohms)}$$

The resistance, R_A, is not entirely ohmic resistance in the antenna elements.

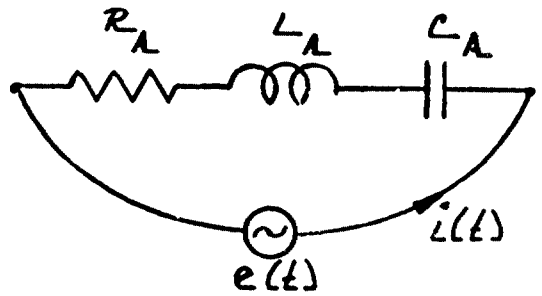
(a) Half-Wave Dipole Antenna



E = Voltage Amplitude

I = Current Amplitude

(b) Equivalent Circuit (Input Impedance)



$$Z_A = R_A + j\left(\omega L_A - \frac{1}{\omega C_A}\right)$$

$$\omega_R = \frac{1}{\sqrt{L_A C_A}}$$

Figure 2.6.4.1 --- Half-Wave Dipole Antenna Characteristics

It is best considered as the ratio of the total input power to the square of the input current. The input power consists of two parts; that part dissipated by ohmic losses in the antenna elements, and that part radiated into the surrounding medium by the antenna. Thus, the total "energy dissipating resistance" is given by:

$$R_A = R_o + R_r \text{ (Ohms)} \qquad [2.6.4.3]$$

where R_o is the ohmic resistance and R_r is the radiation resistance of the antenna.

At the resonant frequency, the input impedance is at a minimum. The input current, and, hence, the radiated power, are at a maximum. At other frequencies, the radiated power, for a given driving voltage, is reduced. An antenna thus has a natural bandwidth, or limited range of frequencies, over which it performs well. The bandwidth is, as usual, bounded by the half-power points.

The effective values for the input resistance, inductance, and capacitance of an antenna, and hence its resonant frequency, are determined by its physical size. To better understand the relationship between the dimensions of an antenna and its electrical characteristics, consider the case of the center-fed dipole shown in Figure 2.6.4.1. In that figure are shown the current and voltage amplitude distributions for the special case when the dipole is one-half wavelength long (two quarter-wavelength elements). The voltage and current distributions shown represent standing waves generated by interference between the incoming waves and the reflections from the open end of the antenna elements. By definition,

the input impedance is the ratio of the voltage to the current, at the driving point (center). That is:

$$Z_A = \frac{e(t)}{i(t)} \quad \text{(Ohms)} \qquad [2.6.4.4]$$

This ratio is minimized, as shown in the figure, when the length of the dipole is exactly one-half wavelength long, as shown. Thus, a one-half wavelength dipole has a natural resonant frequency just equal to the frequency corresponding to that wavelength. Similar considerations can be applied to demonstrate that dipoles of lengths equal to integral multiples of one-half wavelength are also resonant at that frequency.

There are five major factors that contribute to the wavelength limitations of an antenna. They are:

(1) The input impedance of an antenna increases at frequencies other than the resonant value.

(2) The impedance match between the antenna and the feeder (transmission line) is disrupted by a change in frequency from the design value.

(3) The impedance match between the antenna and the propagation medium is disrupted by a change in frequency from the design value.

(4) The directive gain decreases with increasing wavelength (decreasing frequency).

(5) The phase relationships in an array are disrupted by a change in frequency from the design value.

As with any power transmission system, the characteristic impedances of all segments of the system must be matched to avoid reflections at the interfaces. Thus, the transmitter, the transmission line (antenna feeder), the antenna, and the transmission medium must all have the same characteristic impedance. The impedances of the various portions of the system are sufficiently matched over only a limited range of frequencies. The directive gain (radiation pattern) of

an antenna, or antenna array, is also frequency dependent. Directive gain is discussed in Section 2.6.7 of this text.

2.6.5 Antenna Feeders -- The feeder, or transmission line, for the antenna can seriously affect the performance of a communications system. For example, even a small dent in a waveguide can significantly degrade performance. When the intrinsic impedances of the various segments of a transmission system do not match, the arrangement shown in Figure 2.6.5.1 can be employed. The impedance-matching transformers can be conventional inductive transformers, (at low frequencies), capacitors, (at somewhat higher frequencies), or short sections of transmission lines, (at high frequencies). Devices called baluns (BALanced/UNbalanced) are used when matching segments balanced with respect to ground to segments unbalanced with respect to ground, (i.e. with a "high" side and a "ground" side).

The four types of transmission line commonly employed are:

(1) Parallel wires

(2) Coaxial cable

(3) Stripline

(4) Waveguide

The parallel-wire transmission line, shown in Figure 2.6.5.2(a), is generally employed in the frequency range below 30 megahertz. The characteristic impedance of this type of line ranges from about 200 ohms to 800 ohms, and the line is balanced; that is, neither side of the line is at ground or earth potential. Typical of this type of transmission line is the 300 ohm twin-lead often used for television lead-in wire. Parallel-wire line may or may not have a surrounding shield.

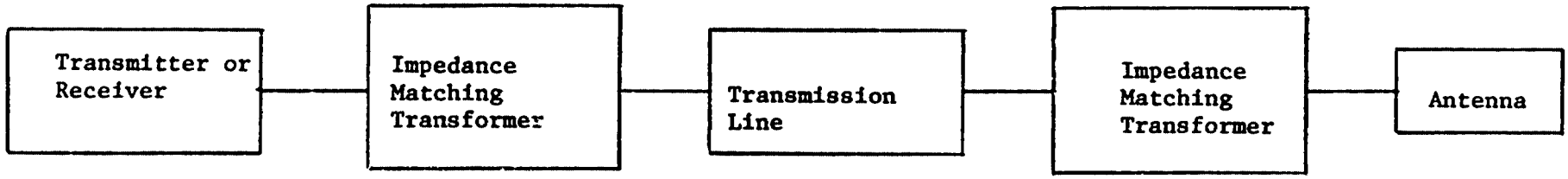
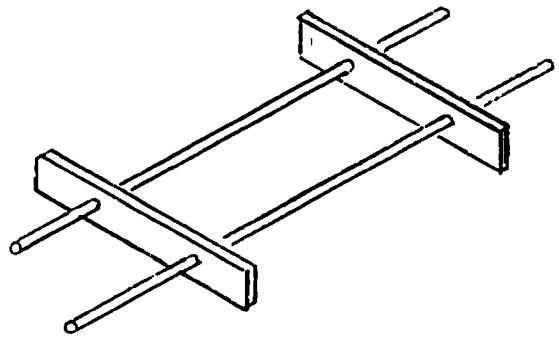
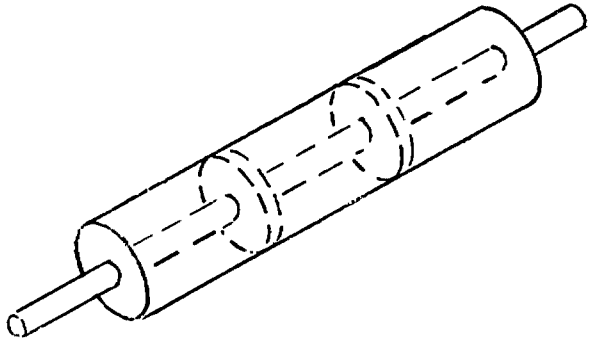


Figure 2.6.5.1 -- Antenna Feeder System

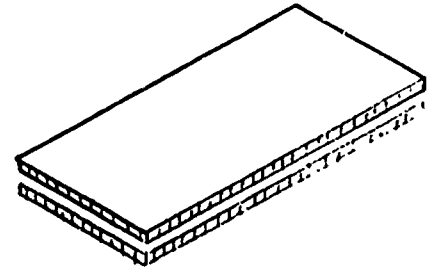
(a) Parallel Wire



(b) Coaxial Cable



(c) Strip Line



(d) Waveguide

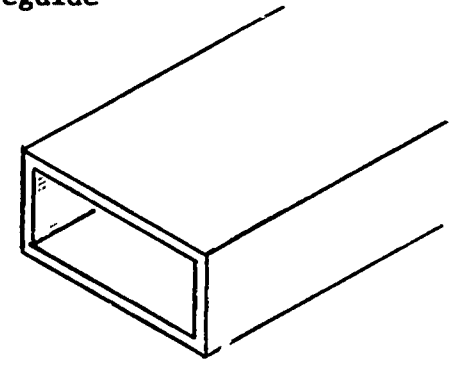


Figure 2.6.5.2 -- Transmission Lines

At frequencies above about 30 megahertz, unacceptable losses occur in the fields surrounding a parallel wire transmission line. In order to prevent these losses, coaxial cable is employed in the frequency range from about 30 megahertz to 3000 megahertz. Coaxial cable can be considered a two-wire line with one of the wires wrapped around the other, as shown in Figure 2.6.5.2(b), in order to constrain the fields between them. Coaxial cables have characteristic impedances in the range between 20 and 100 ohms and are unbalanced lines with the outer "shield" normally at ground potential.

Striplines, like coaxial cables, are designed to constrain the fields between the conductors. As can be seen in Figure 2.6.5.2(c), the narrow gap between the conductors accomplishes that purpose, though not as well as does the coaxial cable. Striplines are often used in printed-circuit board construction.

At frequencies above 3000 megahertz, the losses in the conductors and leakage between the inner and outer conductors of coaxial cable become excessive. These losses can be greatly reduced by eliminating the inner conductor. When the outer conductor is properly dimensioned, the propagating energy in a coaxial line is almost entirely in the fields and the current in the inner conductor vanishes. Under these conditions, the inner conductor can be eliminated, thus producing a hollow "pipe", called a waveguide, shown in Figure 2.6.5.2(d). A waveguide exhibits two unusual characteristics. The first is that it possesses a cutoff frequency below which it cannot propagate wave energy. The cutoff frequency is determined by the dimensions of the waveguide. The second unusual characteristic is that there are two wave "velocities" for a wave traveling in a waveguide. One, called the group velocity, is the velocity with which energy, and also information, is propagated. The other, called the phase velocity, is the one associated with

the phase of the propagating wave and for which $v = f \lambda$ where f is the frequency and λ is the wavelength in the waveguide. Because the wave energy is contrained in the waveguide, the wavelength there is greater than the free-space wavelength and, hence, the (phase) velocity is greater than the free-space velocity, c. The phase velocity, v_{ph}, and group velocity, v_g, are related by the equation:

$$v_{ph}\ v_g = c^2 \quad (\text{Meters}^2/\text{Second}^2) \qquad [2.6.5.1]$$

where c is the free-space velocity. When the frequency of the propagating wave is much greater than the cut-off frequency, the phase and group velocities are nearly identical and equal to the free-space velocity. Under those conditions, the wave propagation approaches that of free-space and the characteristic impedance of the waveguide approaches 377 ohms, the free-space intrinsic impedance. The power-handling capacity and voltage breakdown (arcing) characteristics of waveguides are superior to those for the other types of transmission lines.

Single antennas and/or transmission lines are often used for both the transmitting and receiving functions. Under such circumstances, a device, called a duplexer, must be used to prevent the high-power transmitted signal from entering the receiver and to prevent the low-power received signal from being "shorted-out" by the transmitter. Two basic approaches are commonly employed. One is the transmit/receive (T/R) switch which automatically "switches" to route the outgoing signal to the antenna, (and short the receiver input), during transmission; and to route the incoming signal to the receiver during reception. The alternative approach is to use a device, called a circulator, which does not "switch" between transmit and receive states, but simultaneously routes the transmitted and received signals to their appropriate destinations. The transmit/receive switch

and circulator duplexers are usually represented by the symbols shown in Figure 2.6.5.3.

2.6.6 Antenna Reciprocity -- It can be shown that, generally, the transmitting and receiving characteristics of an antenna are the same. This fundamental property, called reciprocity, implies that:

(1) The radiation and absorption patterns of an antenna are the same.

(2) All other antenna characteristics are the same; including gain, beam width, effective aperture, bandwidth, input impedance, and efficiency.

(3) If a given power into antenna A results in a certain signal level from antenna B, then an identical power into antenna B will produce the same signal level from antenna A.

There are three circumstances in which the reciprocity theorem does not apply:

(1) in active antennas, in which unidirectional signal processing elements are incorporated in the antenna.

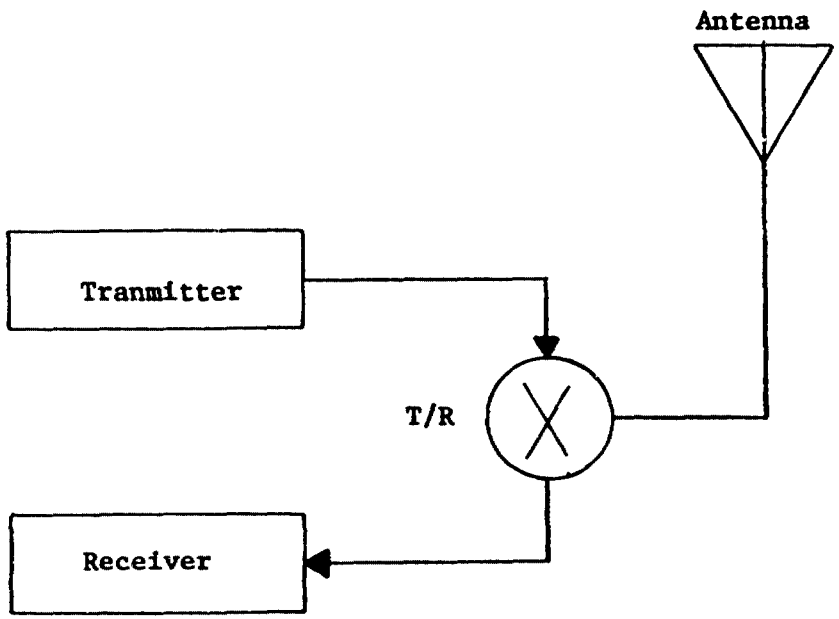
(2) in communications links involving tropospheric propagation, in which the earth's magnetic field affects propagation differently in the two directions.

(3) in considerations of power dissipation, in which the power capacity of an antenna might be adequate for reception but not for transmission.

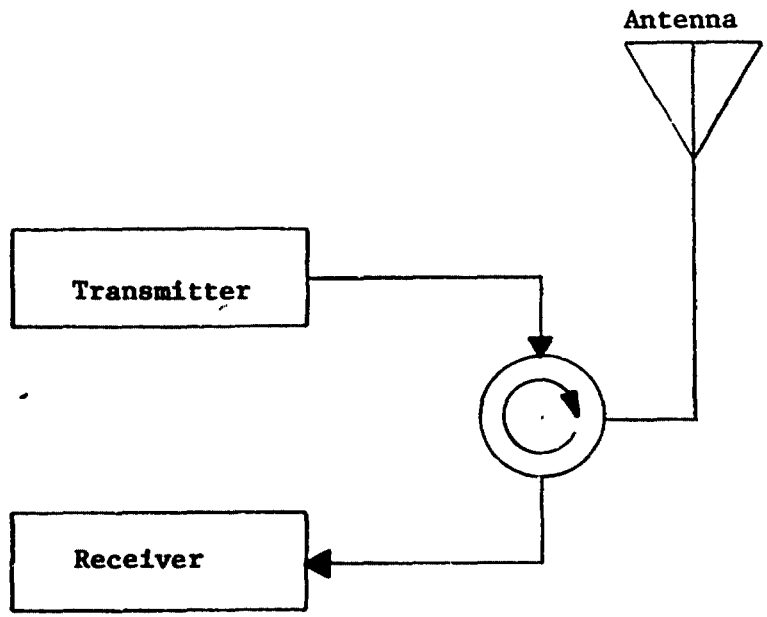
Antenna reciprocity is of great importance in the field of antenna test and evaluation. Due to space and/or instrumentation requirements, it is often desirable to test an antenna intended only for reception (such as some navigation antennas), by incorporating it into an arrangement that evaluates its performance in the transmitting mode.

2.6.7 Antenna Radiation Patterns -- An antenna pattern is a graphical representation of radiated field power density, (or field strength), as a function of angular offset from a reference axis, usually the antenna bore sight axis (direction of maximum radiation). Two presentations are commonly employed. One

(a) Transmit/Receive Switch



(b) Circulator



2.132a

Figure 2.6.5.3 -- Duplex Diagrams

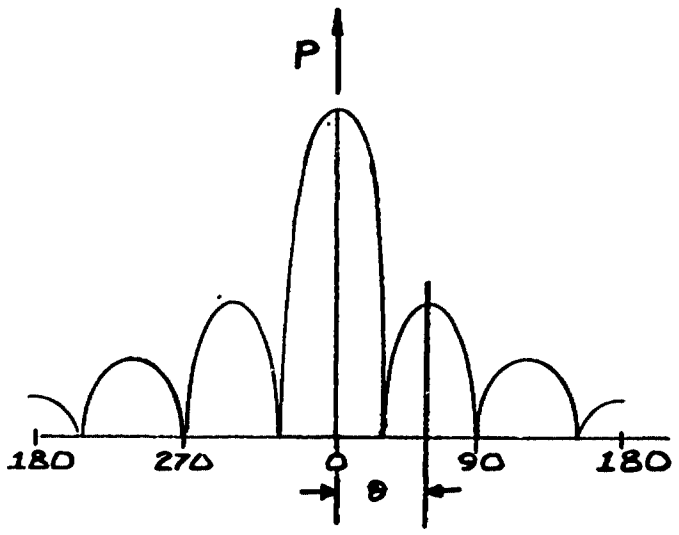
is a plot in cartesian (rectangular) coordinates, with the power density as the ordinate (y-axis) and the offset angle as the abscissa (x-axis). The other is a plot in polar coordinates, with the power density as radial distance and the offset as the polar angle. Both presentations are shown in Figure 2.6.7.1. In actuality, these patterns are cross-sections of three-dimensional field patterns. An antenna pattern can be displayed in three dimensions by utilizing several cross-sectional views. Fortunately, many antennas have axes (or planes) of symmetry which produce corresponding symmetries in their radiation patterns. Thus, a planar (two-dimensional) plot often suffices.

Isotropic Antenna -- An isotropic antenna is a theoretic antenna that radiates in a perfectly spherical pattern; that is, equally in all directions. Its radiation pattern, in any plane through the source, is a circle. Although a truly isotropic antenna cannot be built, antennas with circular (omnidirectional) patterns in one plane are common.

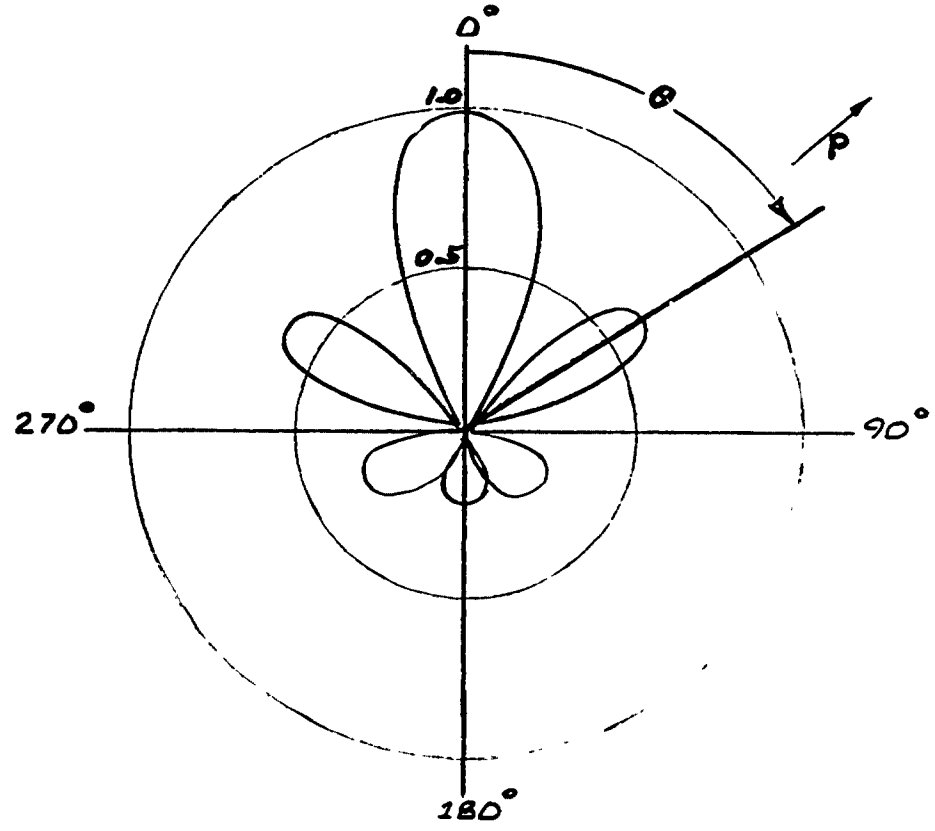
Directive Antennas -- Any non-isotropic antenna is, by definition, directive. That is, it radiates more power in some directions than in others. The patterns shown in Figure 2.6.7.1 are typical of a directional antenna.

Directive Gain -- The directive gain, (sometimes called directivity), of an antenna is a measure of the degree to which it radiates in a preferred direction. No antenna possesses gain in the sense that more total power is radiated than is input. A directive antenna merely re-directs the radiation, reducing the flow of energy in some directions in order to increase it in the preferred directions.

(a) Cartesian Plot



(b) Polar Plot



P = Radiant Power Density

θ = Directional (Offset) Angle

Figure 2.6.7.1 -- Antenna Radiation Patterns

The directive gain has two equivalent definitions:

(1) Directive gain is the ratio of the power density in the preferred direction to the average power density for all directions.

(2) Directive gain is the ratio of the power density in the preferred direction to that which would be produced by an isotropic antenna.

An isotropic antenna has no directive gain. If a total power, P_T, is radiated by an isotropic antenna, the power density at a distance R from the antenna is given by:

$$P_I(R) = \frac{P_T}{4\pi R^2} \quad \text{(Watts/Meter}^2\text{)} \qquad [2.6.7.1]$$

That is, the power P_T is distributed uniformly over the spherical area $4\pi R^2$. If the same total power, P_T, is radiated by an antenna with directive gain, G_D, the power density <u>in the preferred direction</u> is given by:

$$P_D(R) = \frac{P_T G_D}{4\pi R^2} \quad \text{(Watts/Meter}^2\text{)} \qquad [2.6.7.2]$$

That is:

$$G_D = \frac{P_D}{P_I} \quad \text{(Non-Dimensional)} \qquad [2.6.7.3]$$

assuming the same (or no) internal power losses in either antenna.

The directive gain of an antenna that radiates uniformly into a solid angle of Ω stearadians is given by:

$$G_D = \frac{4\pi}{\Omega} \quad \text{(N.D.)} \qquad [2.6.7.4]$$

(The stearadian measure of a solid angle is equal to the area subtended by that

solid angle, on a sphere of radius R, divided by the square of the radius). Although Equation [2.6.7.4] applies strictly only for the case of uniform radiation in a given solid angle, it often finds application in estimating the gain of an antenna from its radiation pattern.

Power Gain -- The power gain of an antenna is defined to include the effects of power losses internal to the antenna. The power gain, G, is related to the directive gain, G_D, by the equation:

$$G = \rho_r G_D \quad \text{(N.D.)} \qquad [2.6.7.5]$$

where ρ_r is the radiation efficiency. Although the radiation efficiency of some ground installations may be as low as 0.05, many airborne antennas have efficiencies approaching unity. For these high-efficiency antennas, the terms power gain and directive gain are often used interchangeably.

Beam Width -- The beam width of a directive antenna is a measure of the polar angle subtended by the main lobe of the antenna pattern. It is not, however, measured to the minima on either side of the main lobe. By convention, it is measured to the points, on either side of the peak, where the power density has declined to one-half the value at the peak. In terms of the decibel, the beam width is defined as the angle between two radial lines intercepting the main lobe pattern at the points 3 dB down from the maximum. The beam width is illustrated in Figure 2.6.7.2. Also labeled in the figure are the main and side lobes of the radiation pattern. Note that a receiver just outside the "beam width" of an antenna receives a signal almost as large as one just inside the beam width.

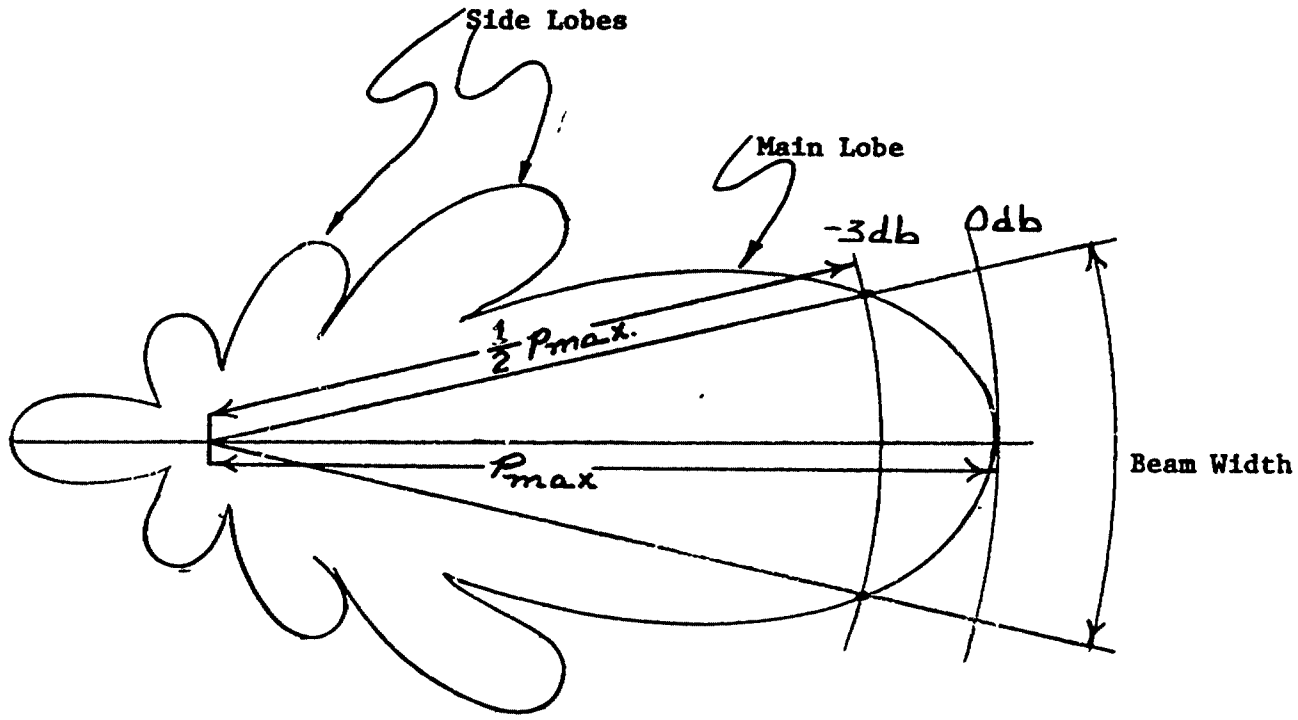


Figure 2.6.7.2 -- Antenna Beam Width

Diffraction effects, as discussed in Section 2.4.2 of this text, make antenna performance dependent upon the size of the antenna, in comparison to a wavelength. A rule of thumb relating antenna beam width to antenna aperture size is given by:

$$[\text{Beam Width}] = \left(\frac{3.2}{\pi}\right)\frac{\lambda}{D_A} \quad \text{(Radians)} \qquad [2.6.7.6]$$

where D_A is the dimension of the antenna aperture in a plane perpendicular to the plane in which beam width is to be determined.

Antenna Aperture -- The aperture of an antenna is its effective capture area for electromagnetic radiation. That is, if a radiation power density, P, is incident upon a receiving antenna with aperture of effective area, A_E, the power absorbed by the antenna, (and delivered to the receiver), is given by:

$$P_R = P A_E \quad \text{(Watts)} \qquad [2.6.7.7]$$

The effective area of an antenna, though related to its physical size, is rarely equal to its physical (profile) area. It can be either smaller or larger, depending upon antenna design. The effective area, A_E, and physical area, A, are related by the equation:

$$A_E = \rho_a A \quad \text{(Meter}^2\text{)} \qquad [2.6.7.8]$$

where ρ_a is the aperture efficiency. The effective area and directive gain of an antenna are related, through reciprocity, by the equation:

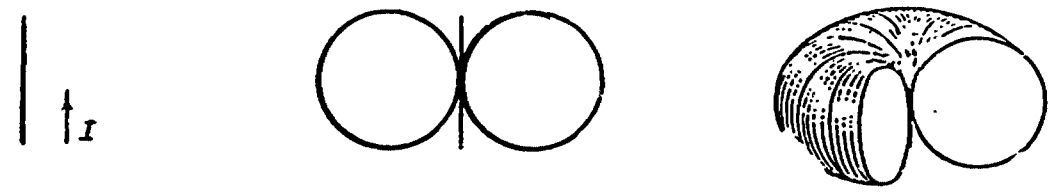
$$A_E = \frac{\lambda^2 G_D}{4\pi} \quad \text{(Meter}^2\text{)} \qquad [2.6.7.9]$$

where λ is the wavelength of the radiation. Thus, an increase in antenna aperture (size) generally produces an increase in directive gain and a decrease in beamwidth. Aperture, or effective area, can be increased by increasing antenna size, utilizing multiple antennas (an antenna array), or by employing a reflector or lens.

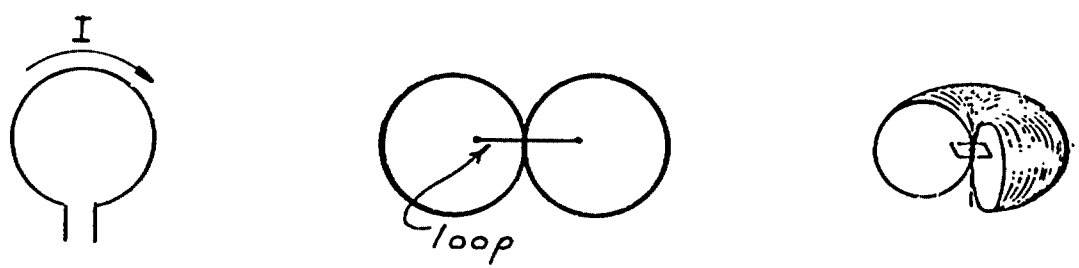
2.6.8 Basic Radiating Elements -- Fundamentally, there are three electromagnetic-wave-radiating elements (antennas small in comparison with the radiation wavelength). These are the dipole (a short, straight wire), the loop (a small closed loop of wire), and the slot (a small hole in a waveguide). The (far-field) directional patterns of the basic antenna elements are all $\cos(\theta)$ patterns as shown in Figure 2.6.8.1. The fields for the loop and the slot differ from those for the dipole, however, in that the directions of the electric and magnetic fields are interchanged. For this reason, a loop is sometimes called a "magnetic doublet" and the slot is considered the "dual" of the dipole.

2.6.9 Polarization -- As discussed in Section 2.4.1 of this text, the direction of polarization of an electromagnetic wave is defined as the direction of the electric field vector, $\vec{\mathcal{E}}$. All physical antennas produce polarized radiation. The basic radiating elements described in Section 2.6.8 all produce plane polarized radiation. The direction of the electric field vector for a dipole or doublet is in a plane containing the doublet and is perpendicular to the direction of propagation. The direction of the electric field vector for a loop is in a plane perpendicular to the central axis of the loop and is perpendicular to the direction of propagation. The direction of the electric field vector for a slot is in a plane perpendicular to the long axis of the slot and is perpendicular to the direction of propagation.

(a) Dipole (Short Doublet)



(b) Loop (small wrt λ)



(c) Slot (Complement of Dipole)

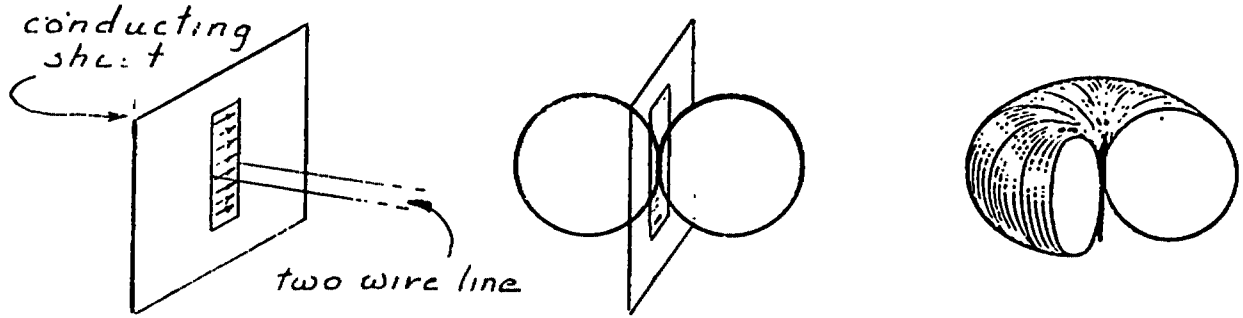


Figure 2.6.8.1 -- Basic Radiating Elements

Radiation containing components with plane polarization of various orientations can be produced by extended lengths, or arrays, of elements. When two such components, with perpendicular directions of polarization (in space quadrature), are ninety degrees out of phase electrically (in phase quadrature), elliptically polarized radiation is produced. The turnstile antenna, shown in Figure 2.6.9.1(a) employs that principle to produce elliptically or circularly polarized radiation in the direction perpendicular to the plane of the antenna. Airborne systems often employ the helical antenna shown in Figure 2.6.9.1(b) for the same purpose.

2.6.10 Highly Directive Antennas -- As indicated in Section 2.6.7 of this text, all physical antennas are directive to some extent. When a high degree of directivity is required, however, one of four basic methods is generally employed. These methods are:

(1) the use of an array

(2) the use of a reflector

(3) the use of a lens

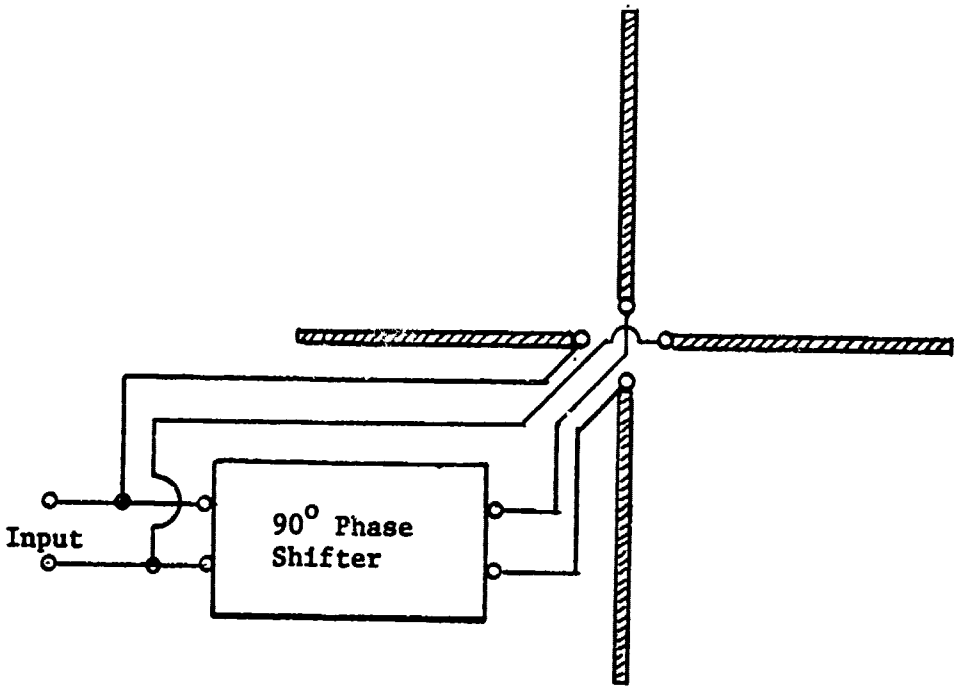
(4) the use of a horn

An antenna array is an arrangement of radiating elements situated in such a way that the radiation from the individual elements combines, via wave interference, to produce a directional pattern. Any antenna arrangement more extensive than one of the three basic radiating elements described in Section 2.6.8 can be considered an antenna array. Even a single doublet (length of wire) can be considered an end-to-end arrangement of elementary dipoles. The radiation pattern produced by an array depends on three factors:

(1) the radiation patterns of the individual elements

(2) The ralative position and orientation of the elements

(a) Turnstile Antenna



(b) Helical Antenna

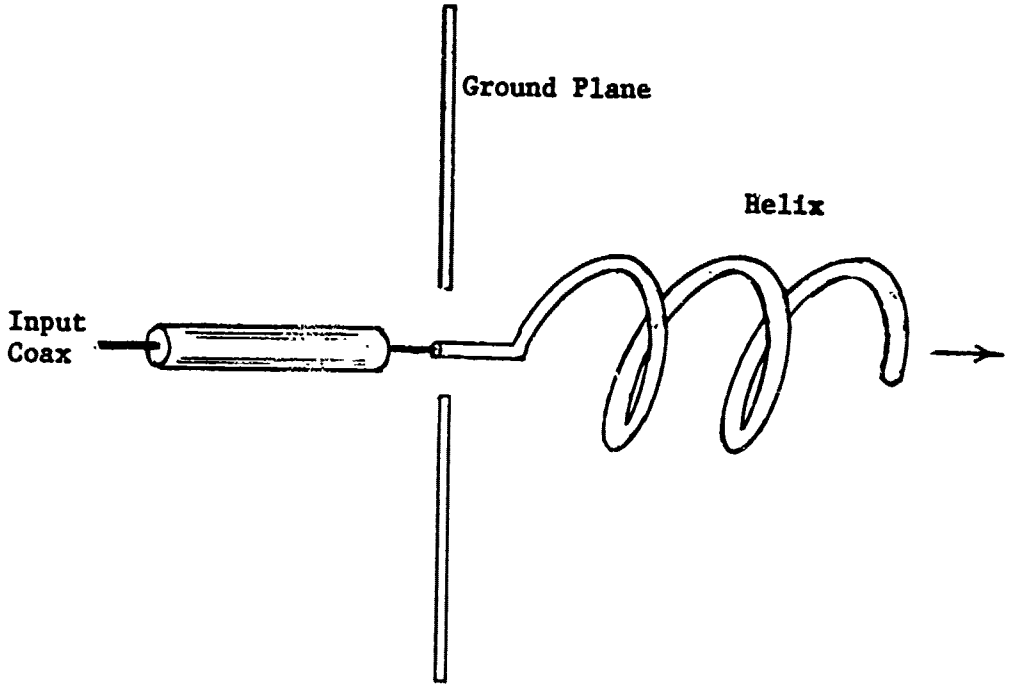


Figure 2.6.9.1 -- Circularly Polarized Antennas

(3) the relative magnitudes and phases of the currents in the elements (or fields in waveguides).

The radiation pattern of the array can be determined by: (1) determining the pattern (field intensity) of the individual elements, $E_e(\theta)$; (2) determining the pattern of the array if composed of isotropic radiators, $E_i(\theta)$; and (3) taking the product of the two patterns. That is, the radiation pattern of the array, $E_A(\theta)$, is given by:

$$E_A(\theta) = E_e(\theta)\, E_i(\theta) \quad \text{(Watts/Meter}^2\text{)} \qquad [2.6.10.1]$$

where it is assumed that the individual elements are identical, in-phase electrically, and identically oriented to produce maximum radiation in the preferred direction. It should be noted that, in general, mutual impedance (coupling) between the elements of an array affects the currents in the elements and must be taken into account. This process is illustrated in Figure 2.6.10.1 for the case of a linear array of dipoles spaced one wavelength apart. The beamwidth of this type of array, called a broadside array, is given approximately by:

$$[\text{Beam width}] = \frac{1}{n}\left(\frac{\lambda}{d}\right) \quad \text{(Radians)} \qquad [2.6.10.2]$$

where n is the number of elements in the array, (assumed large), and d is the spacing of the elements. (Note that (λ/d) is the approximate beamwidth of an antenna of dimension d, as given by Equation [2.6.7.6.]. The directive gain, G_D, of a broadside array is given approximately by:

$$G_{DA} = \frac{4\pi A}{\lambda^2} \quad \text{(N.D.)} \qquad [2.6.10.3]$$

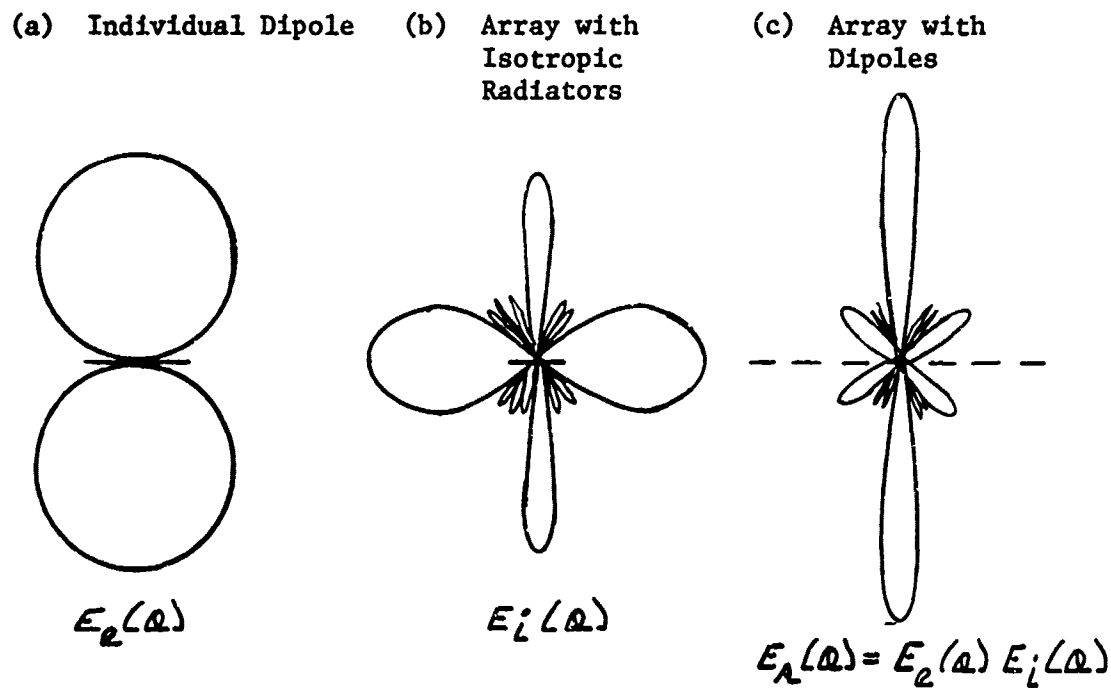


Figure 2.6.10.1 -- Radiation Pattern for Linear Array of Dipoles

where A is the frontal effective area of the array.

Reflectors are most effective when their dimensions are much larger than the wavelength of the radiation. For this reason, reflectors find greatest use at frequencies above one gigahertz (wavelength about one foot). Under these conditions, the methods of geometrical optics (wavefront and ray tracing) apply.

The most common radiation reflector is the parabolic reflector. The parabolic shape has the property that rays emanating from a source at its focal point will be reflected in a collimated beam, (parallel rays), as shown in Figure 2.6.10.2. It can be shown that, theoretically, the electromagnetic waves emanating from the aperture of a parabolic transmitting antenna are phase coherent and constitute a planar wave front. In accordance with reciprocity, rays incident parallel to the axis of a parabolic antenna are focused to a single point at the focus. Because of its compact construction, the cassagrainian antenna shown in Figure 2.6.10.3 is frequently used in airborne applications. Also frequently employed is the offset parabolic antenna shown in Figure 2.6.10.4. In practice, even a perfectly constructed antenna is limited by diffraction to a finite beam width. As for the case of antenna arrays, the beamwidth and directive gain of reflector antennas are determined by the size of the antenna aperture, according to the equations:

$$[\text{Beam Width}] \cong \frac{\lambda}{D} \quad \text{(Radians)} \qquad [2.6.10.4]$$

$$G_D = \frac{4\pi A}{\lambda^2} \quad \text{(N.D.)} \qquad [2.6.10.5]$$

where D is the linear dimension of the aperture and A is the effective area.

Parabolic Reflector Antenna

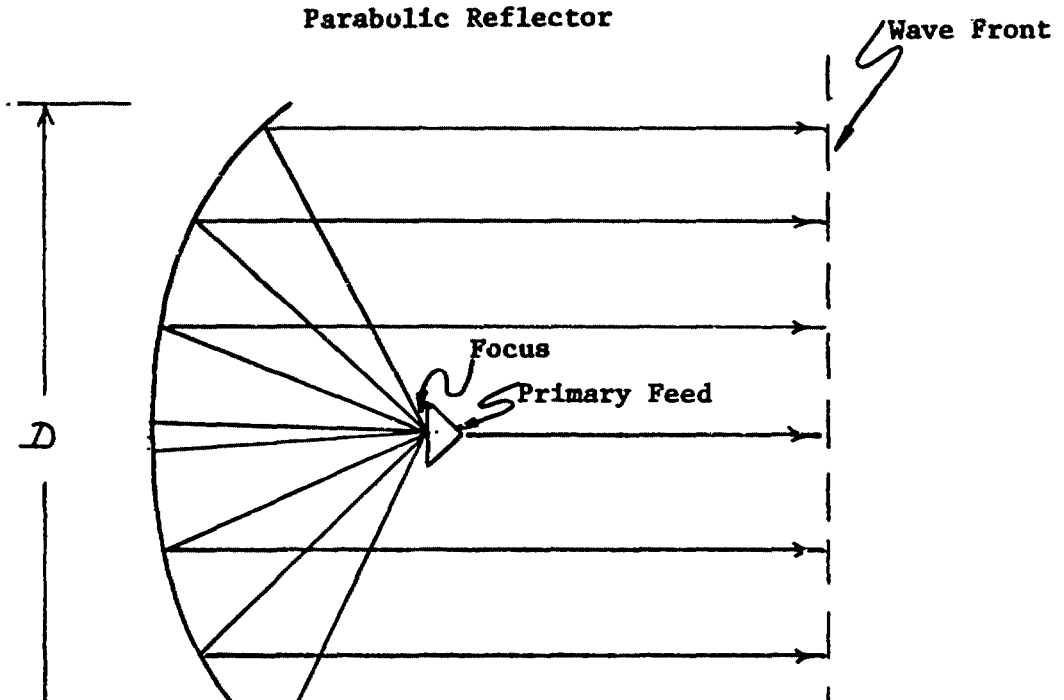


Figure 2.6.10.2 -- Parabolic Reflector Antenna

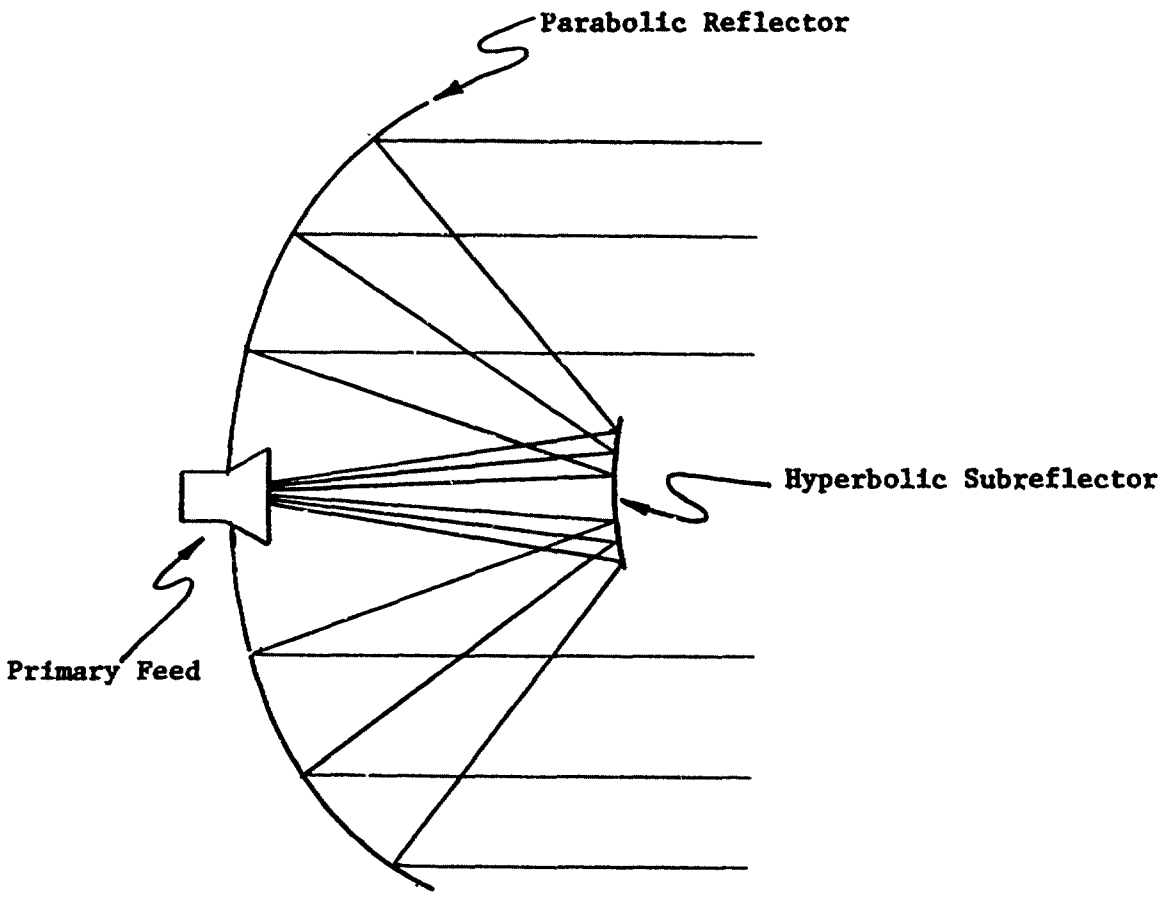


Figure 2.6.10.3 -- Cassegrainian Reflector Antenna

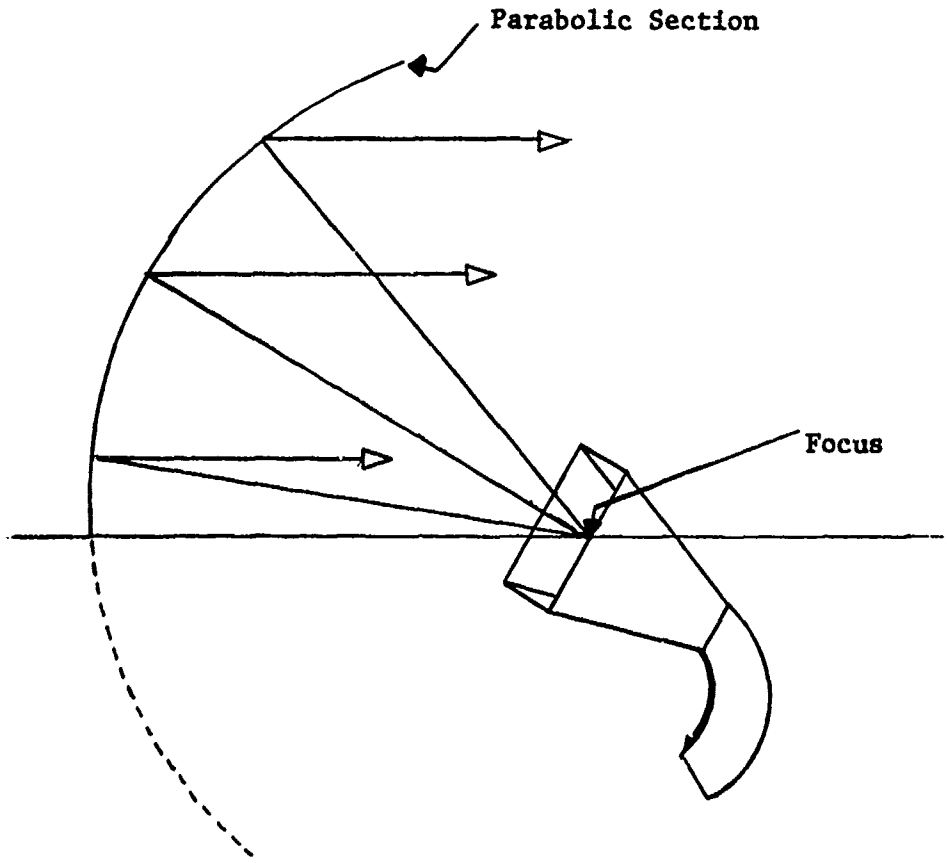


Figure 2.6.10.4 -- Offset Parabolic Reflector Antenna

At frequencies above about three gigahertz, lens antennas are often employed. Lens antennas utilize the phenomenon of refraction to accomplish the same functions for which reflector antennas utilize reflection. When a radiation source is placed at the focal point of the lens, the rays emanate from the lens parallel to the axis as shown in Figure 2.6.10.5. Rays entering the lens parallel to the axis are, of course, refracted to a point at the focus.

A horn antenna is best considered an impedance matching device between a waveguide and free space. The characteristic impedance of a waveguide is determined by its internal dimensions in comparison with the wavelength of the radiation. In order to provide the necessary impedance transformation from that of the waveguide to that of free space, the waveguide is terminated in a flared section (horn). Ideally, an infinitely long section would be required for a perfect match to free space. In practice, a truncated section is used, similar to one of the horns shown in Figure 2.6.10.6.

2.6.11 Effect of Ground Reflections -- As discussed in Section 2.4.2 of this text, electromagnetic waves are reflected by the air-to-earth interface. For that reason, the radiation patterns of antennas in proximity to the ground depend upon the orientation of the antenna and its height above the ground. The determination of such antenna patterns is greatly simplified by the application of a technique known as the Method of Images. The Method of Images makes use of the fact that the radiation reflected by the earth is identical (above the ground) to that which would emanate from an antenna, below ground level, that is the image, reflected in the air-to-ground interface, of the original antenna. The configuration, orientation, and current distribution of the image antenna

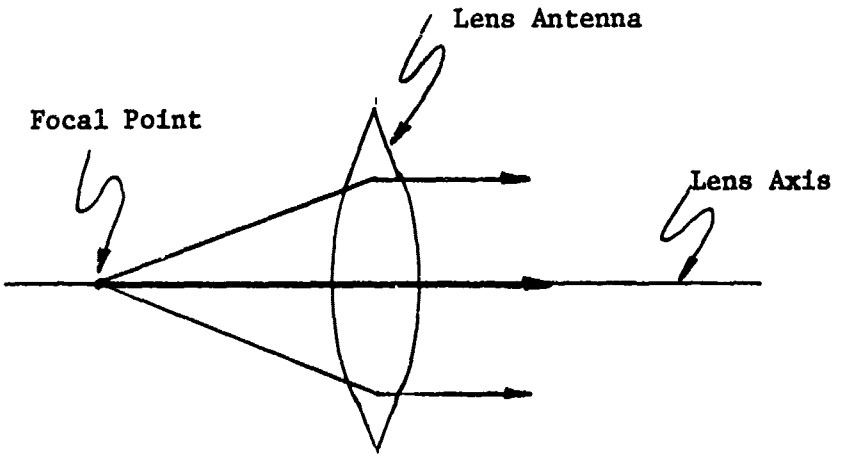
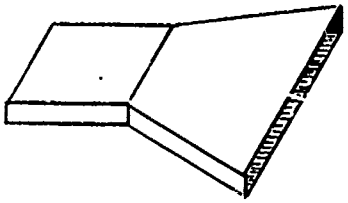
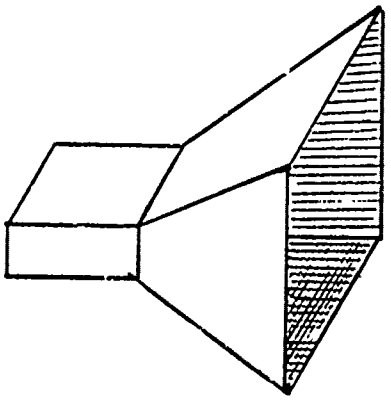


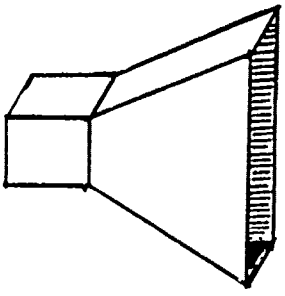
Figure 2.6.10.5 -- Lens Antenna



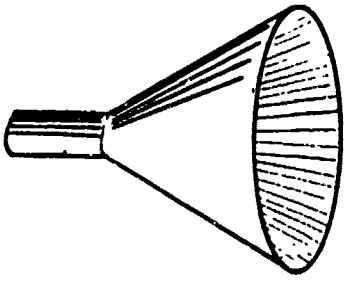
H - Plane Horn



Pyramidal Horn



E - Plane Horn



Conical Horn

Figure 2.6.10.6 -- Horn Antennas

are those of a mirror image as shown in Figure 2.6.11.1. The vertical component of current in the image antenna flows in the same direction as that of the original antenna. The horizontal component flows in the opposite direction, as shown in the figure. The radiation from the real and image antennas, being coherent, interferes, constructively or destructively, depending on the phase and relative path lengths of the direct and reflected (imaged) rays. The composite pattern of a vertical, quarter-wavelength doublet, shown in Figure 2.6.11.2(a), is shown in Figure 2.6.11.2(b). Note that, above ground, it is the field of a half-wave dipole, as required. Since the currents in the image of a horizontally-polarized (oriented) antenna are reversed from those of the real antenna, destructive interference occurs at zero elevation angle. For example, the original and composite radiation patterns, for a horizontally-polarized, isotropic antenna one half wavelength above the ground, are shown in Figure 2.6.11.3. The radiation patterns shown in Figures 2.6.11.2 and 2.6.11.3 assume a perfectly reflecting earth. The patterns for a non-perfectly reflecting earth would be very similar with the maxima slightly reduced and the minima somewhat filled in.

2.6.12 Antenna Directional Scanning -- For many applications, it is necessary to change the direction of maximum (or minimum) gain (boresight) of an antenna. The principal methods of effecting the change are:

(1) Mechanically moving (rotating or translating) the entire antenna.

(2) Mechanically moving antenna elements, reflectors, lenses, or feeders.

(3) Switching between fixed antennas or horns.

(4) Electrically changing the phase relationships in an antenna array.

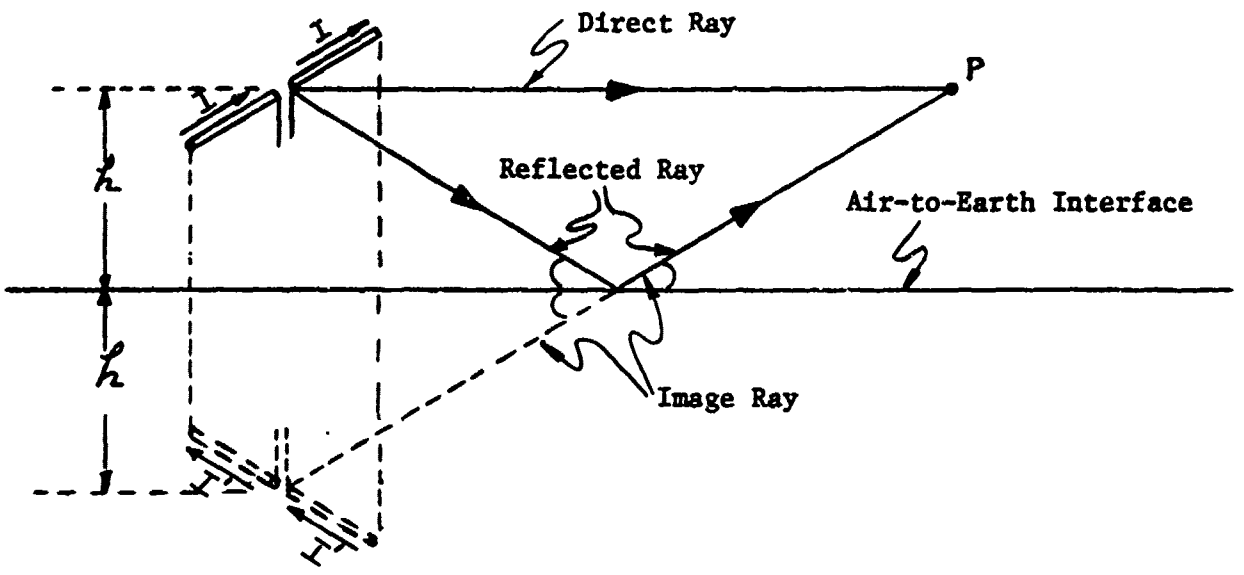
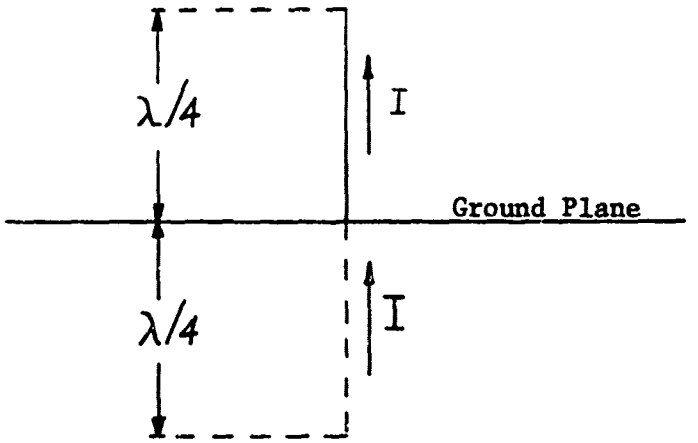


Figure 2.6.11.1 -- Image Antenna

(a) Vertical Quarter - Wave Dipole and Its Image



(b) Radiation Pattern of Vertical Quarter-Wave Dipole (and Its Image)

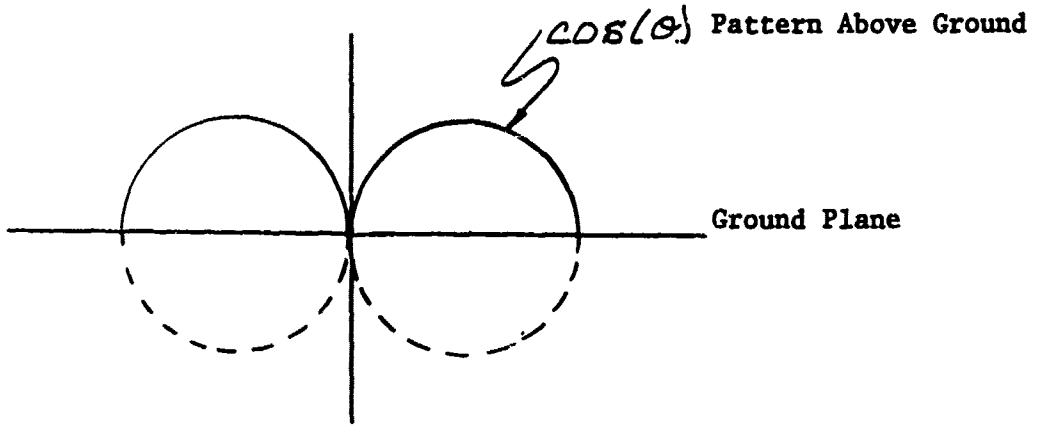


Figure 2.6.11.2 -- Vertical Quarter-Wave Doublet and Its Radia

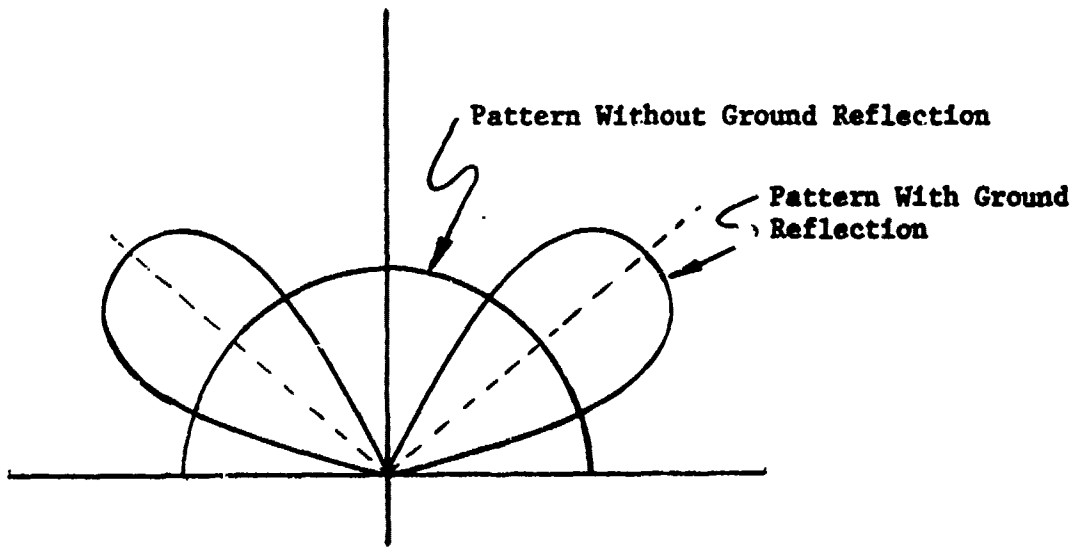


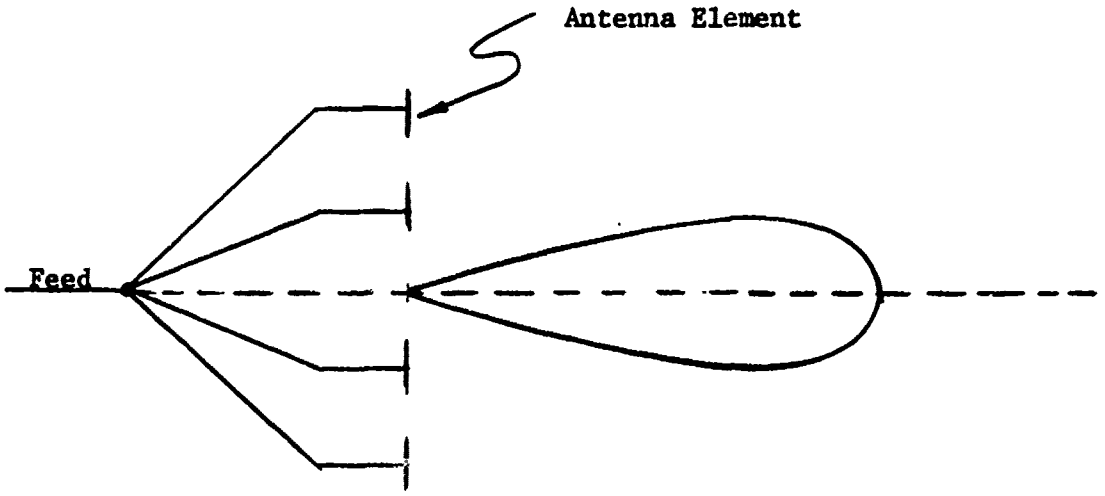
Figure 2.6.11.3 -- Radiation Patterns of a Horizontally Oriented, Isotropic Antenna Mounted One-Half Wavelength Above Ground

Typical examples of airborne moving antennas are the airborne early warning radar antenna that scans 360° in azimuth (rotation) and the radar ground mapping antenna that utilizes the forward motion of the aircraft to scan parallel to the flight path (translation). An example of an airborne moving-feeder antenna is the conical-scan arrangement, often used for air-to-air tracking, in which a conically moving horn illuminates a fixed parabolic reflector. An example of a moving element antenna is the scan antenna in which a cylindrical array of vertical, passive elements rotates about the central active radiator. Not all directional scanning is continuous. The sequential lobing radar antenna generally switches between multiple horns, irradiating a fixed parabolic reflector, to move the beam successively to several discrete positions. A monopulse system employs multiple horns or antennas to simultaneously direct beams (or receiving lobes) to several positions. The "scanning" is performed electronically by mixing the signals received from the several positions. This method is used in airborne radar trackers and also in missile warning systems. An unscanned linear antenna array is shown in Figure 2.6.12.1(a). The elements in the array are all excited in-phase and produce a "beam" (direction of maximum gain) normal to the array. If the phases of the individual element excitations are progressively shifted as shown in Figure 2.6.12.1(b), the wavefront is canted and the direction of the beam is offset from the normal. The scanning can be performed by changing either the phase shift, ϕ, or the wavelength (λ). Because of their reduced mechanical size, weight, and complexity, electronically scanned arrays are finding increasing airborne application.

2.6.13 Synthetic Array Antenna -- In Section 2.6.10, the minimum attainable beam width of an antenna array is given by:

$$[\text{Beam Width}] = \frac{\lambda}{D} \text{ (Radians)} \qquad [2.6.13.1]$$

(a) Unscanned Linear Antenna Array



(b) Scanned Linear Antenna Array

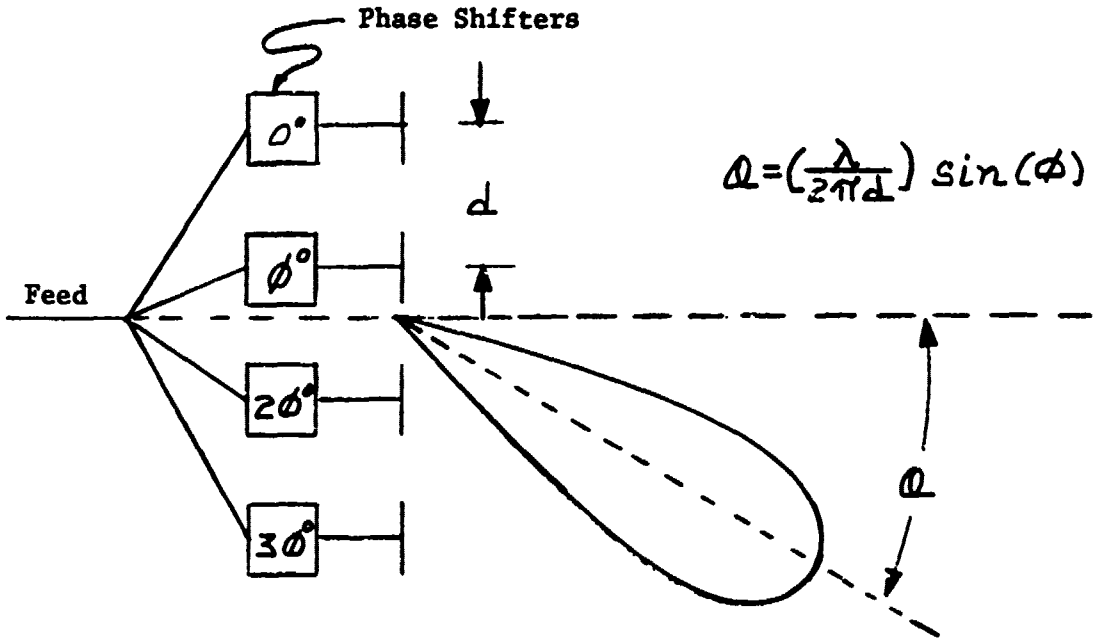


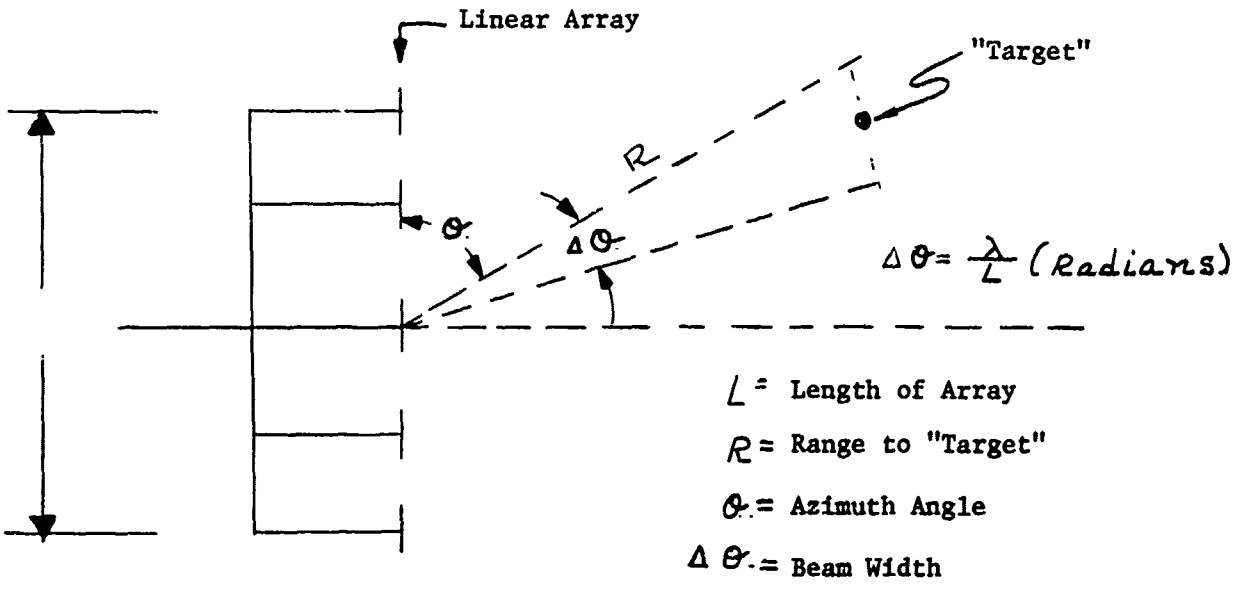
Figure 2.6.12.1 -- Scanned and Unscanned Linear Antenna Arrays

where λ is the radiation wavelength and D is the pertinent effective aperture dimension of the array. Thus, by employing a large array, a narrow beamwidth can be attained. A linear antenna array is shown in Figure 2.6.13.1(a). Similar results can be realized by employing what is know as a synthetic array. A synthetic array is formed by positioning a single antenna at successive points in space as shown in Figure 2.6.13.1(b). These points, for example, can be the positions successively occupied by a moving vehicle. In order to achieve the effect of an array of antennas, the signals from the successive points must be "stored" and "played back", simultaneously, with the proper phasing. For instance, the signal processing must account for the phase shift due to the change in distance from the successive positions of the antenna to the intended "target". The correction of these phase differences is known as focusing. For the case of a moving vehicle, the fact that the Doppler shifts of the signals will differ with position of the antenna also must be accounted for in the signal processing in order to achieve the coherence necessary for proper combination of the signals. (Note that coherent processing also assumes a coherent transmitted signal and a stationary "target". As with any antenna "array", the direction of the array beam (direction of maximum gain) can be scanned by introducing proportional phase shifts into the signals from successive positions. As for the real antenna array, the minimum attainable beamwidth of a synthetic array is given by:

$$[\text{Beam Width}] = \frac{\lambda}{D} \quad \text{(Radians)} \qquad [2.6.13.2]$$

where D is the distance between the first and last antenna positions in the "array". Synthetic arrays have found application in ground mapping radar and in the location of sources of radiation (electronic surveillance).

(a) Real Antenna Array



(b) Synthetic Array

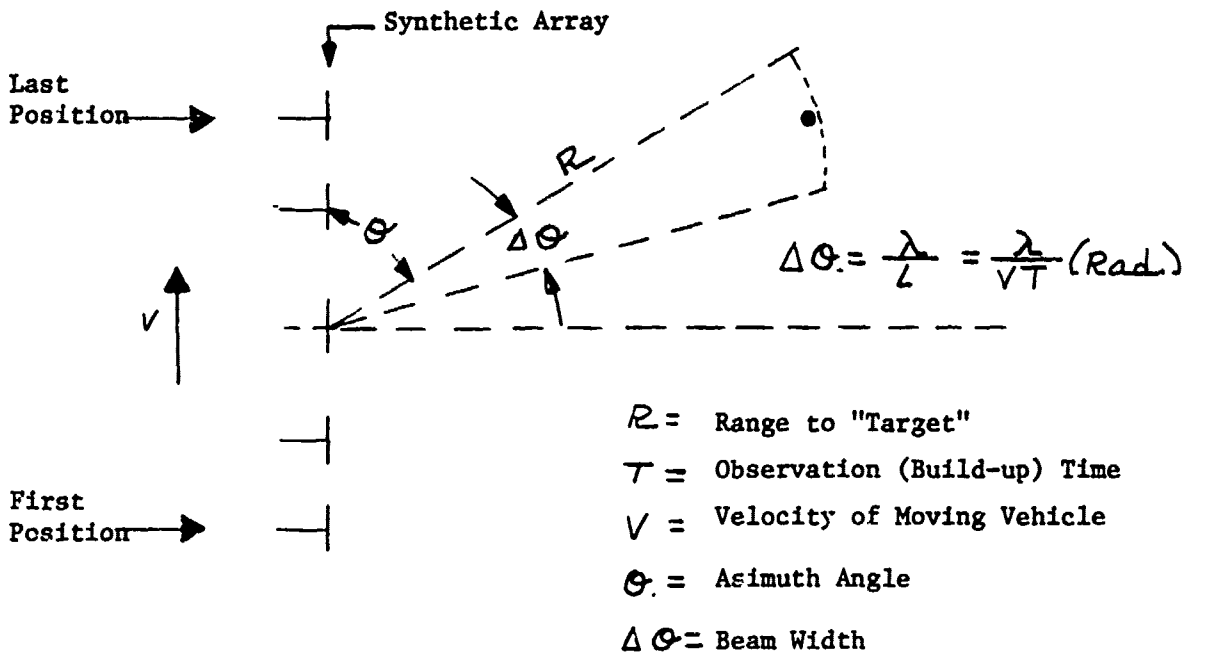


Figure 2.6.13.1 -- Real and Synthetic Antenna Arrays

3.0 Communications System Characteristics

3.1 Generic Communications System

3.1.1 General Description -- Under the broad definition we have assumed for this text, "communications systems" include all systems which convey information from one place to another. The block diagram of a generic communication system is shown in Figure 3.1.1.1. The information signal, as it comes from the source, is rarely in a form suitable for efficient transmission to a distant receiver. For that reason it is modified by a signal processor, (discussed later), and impressed upon a carrier with properties suitable for transmission. (Refer to Section 2.5.5 of this text for a detailed discussion of the modulation process.) The modulated carrier is then power amplified and transmitted through an antenna into the medium of propagation. Although Figure 3.1.1.1 indicates a radio-frequency communications link, the concepts of interest here apply equally well to other forms of signal propagation. As indicated in Section 2.6 of this text, an antenna is primarily an impedance-matching energy transducer. When the transmitted energy is in the form of electromagnetic radiation, the antenna takes one of the forms discussed in Section 2.6. When the transmitted energy is acoustic, as for a sonar system, the "antenna" takes the form of an electromechanical transducer that converts electrical current to sound waves and vice versa. The transmitted signal energy is absorbed from the medium by a receiving antenna (transducer) and is selectively filtered, on the basis of frequency, and amplified, by a receiver. The information signal is then recovered from the carrier by the demodulator. (See Section 2.5.5 of this text for a detailed discussion of demodulation). It is, then, further processed, (as discussed later), and presented, in

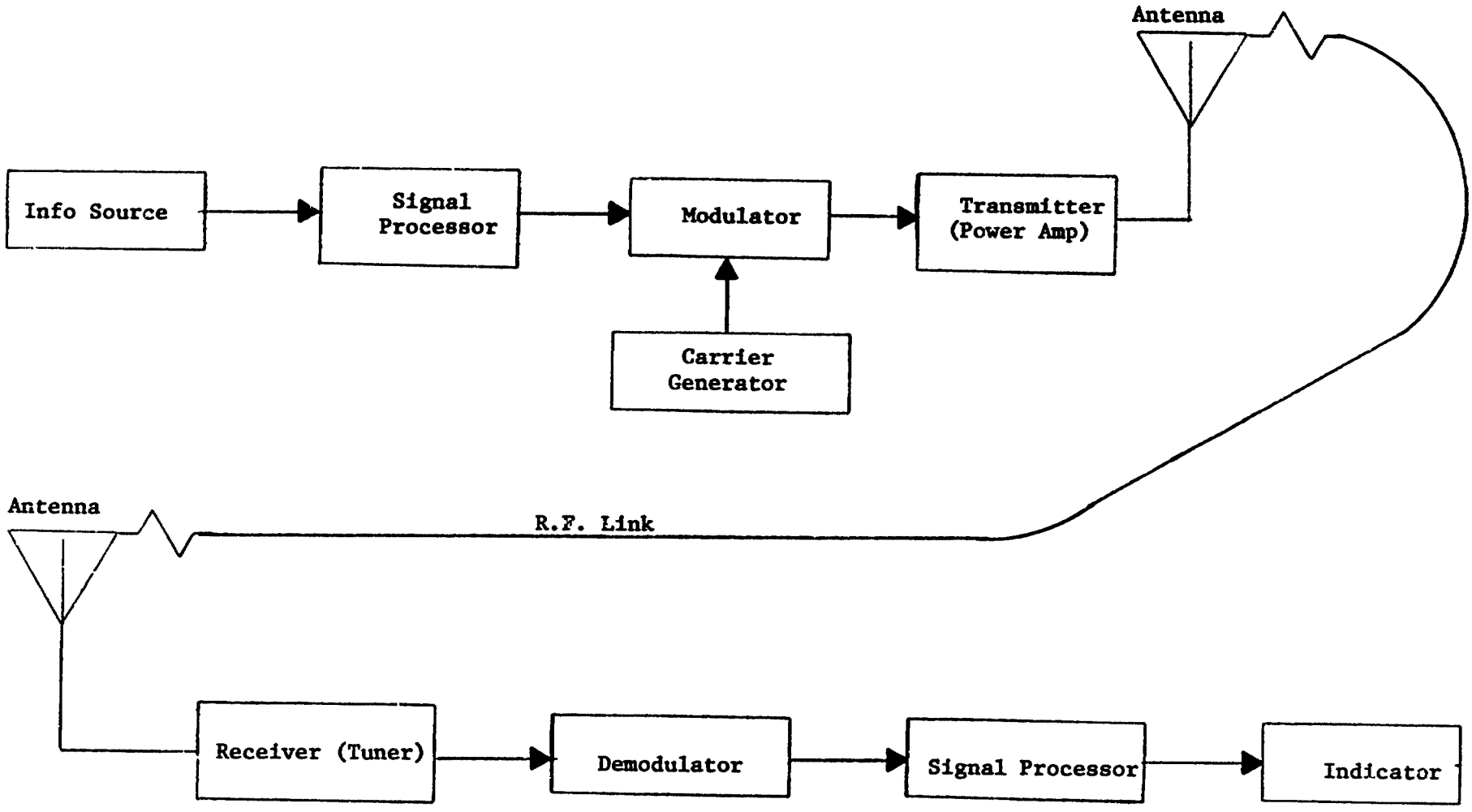
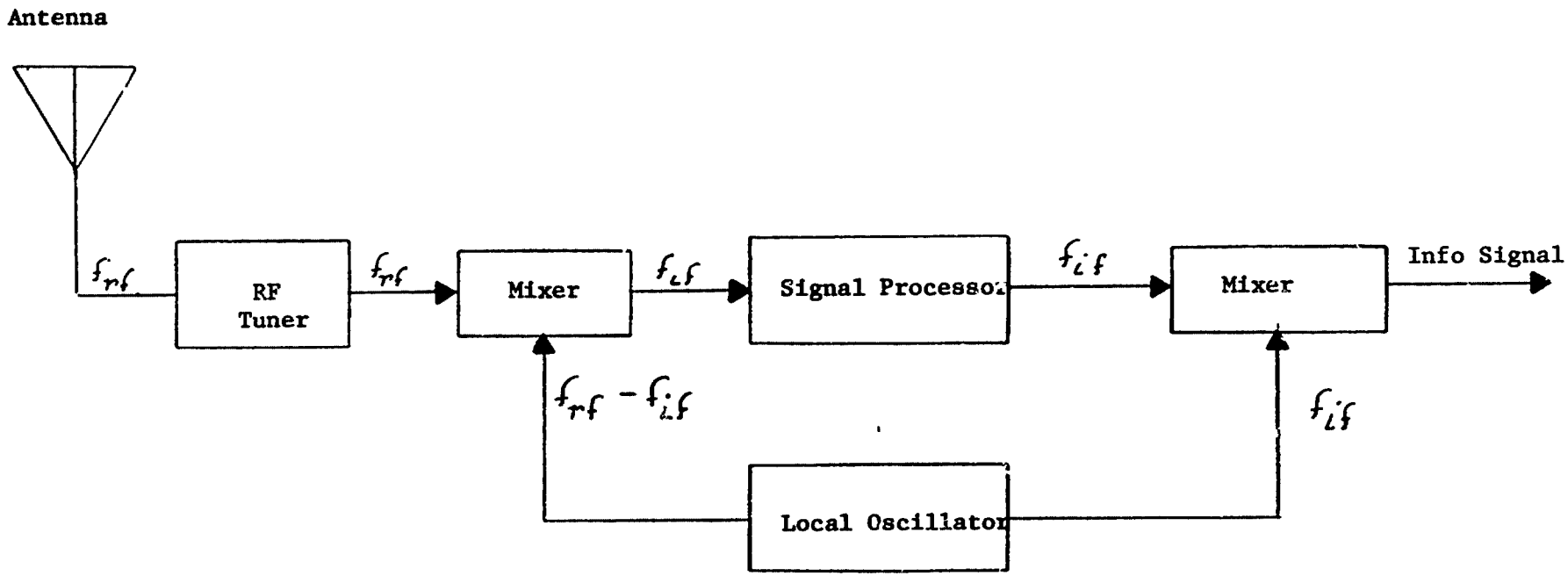


Figure 3.1.1.1 -- Generic Communications System

suitable form, to the intended recipient. The signal processors can employ any of the processes described in Sections 2.5.3 through 2.5.8 of this text, including amplification, attenuation, limiting, frequency filtering, modulation, demodulation, time sampling (commutation), decommutation, multiplexing, demultiplexing, digital encoding, decoding, encryption, decryption, correlation, and spread spectrum processing.

In the foregoing discussion, the process of recovering the information signal from the modulated carrier was treated as a single frequency conversion. Almost all modern receivers employ at least two conversions. For example, the modulated radio frequency (RF) carrier signal is first converted to a signal with the same information content (modulation) but with an intermediate frequency (IF) carrier. After signal processing, the modulated IF signal is demodulated, (converted to a zero-frequency carrier), thus recovering the information signal (audio or video signal). The frequency conversion process, called heterodyning, is identical to the modulation process, as discussed in Section 2.5.5 of this text. The signal to be converted is multiplied by the local oscillator signal in a device called a mixer. The output of the mixer is then filtered to remove the unwanted sidebands, leaving the frequency-converted carrier, still modulated by the information signal. The dual-conversion process described above is called superheterodyning. The reason for employing dual conversion is that most signal processing is more easily performed at intermediate frequencies than at either higher or lower frequencies. The block diagram for a superheterodyne receiver is shown in Figure 3.1.1.2.



f_{rf} = Carrier frequency of transmitted (RF) Signal

f_{if} = Carrier frequency of intermediate (IF) Signal

Figure 3.1.1.2 -- Superheterodyne Receiver

3.1.2 Communication System Requirements -- Signal processing and other aspects of system design are determined by the functional and operational requirements placed upon the communications system. The major requirements are listed below and briefly discussed in the following paragraphs.

Range (Distance)
Transmitting Media
Carrier Constraints
Modulation Constraints
Information Signal Content
Information (Data) Rate
Interference Rejection
Fidelity
Acceptable Error Rate
Operational Constraints
Electronic Counter-Countermeasures
Secure Operation
Covert Operation
Reliability and Maintainability
Electromagnetic Compatibility
Environmental Tolerance
Hardware Constraints
Development, Production, and Operation Costs

Range -- The required range may vary from a few feet (for a cockpit intercommunications system) to millions of miles (for an interplanetary probe data link).

Transmitting Media -- The transmitting media involved may be electrical conductors, radio or optical frequency waveguides, the land or sea masses, the atmosphere, and the vacuum of outer space. More than one medium may be involved.

Carrier Contraints -- Constraints on the carrier to be employed may arise from functional or operational requirements or they may be imposed by an outside authority such as the Federal Communications Commission. They may include constraints as to type, polarization, frequency, bandwidth, power, and time or location of use.

Modulation Constraints -- Constraints on the modulation to be employed are a result of the same factors involved in constraints on the carrier. They may include constraints as to type (AM, FM, PM, PCM), bandwidth, and modulation index.

Information Signal Content -- The content of the information signal to be conveyed impacts nearly every aspect of system design. Equally important are the time domain characteristics (timing, duration, repetition interval), frequency domain characteristics (spectral content, bandwidth), information content (statistical information content or "entropy"), and dynamic range (ratio of largest to smallest values). Fundamental system characteristics depend upon whether the information signal is continuous or discrete, analog or digital, amplitude or frequency modulated, and encoded or unencoded.

Information Rate -- The information rate or data rate is defined as the rate at which basic message symbols or units occur. For a binary digital signal, the information rate is simply the bit rate. For an analog signal, there is an equivalent data rate that depends upon the spectrum of the signal.

Interference Rejection -- As previously indicated, the ultimate determinant of system performance is signal-to-noise ratio. For many communications applications, the prevalent atmospheric or other noise dictates the essential characteristics of the system. The interference may be natural or man-made, and incidental or intentional (jamming).

Fidelity -- Fidelity is the ability of the communications system to exactly reproduce the information signal. Measures of fidelity are accuracy (ability

to reproduce the input), repeatability (ability to repeat the same output for the same input), and resolution (ability to distinguish between two, nearly identical, inputs). In voice communication systems, fidelity is usually measured in terms of distortion. (See Section 3.2.2 of this text for a brief discussion of distortion.)

Acceptable Error Rate -- The acceptable error rate is the ultimate quantitative specification of system performance. Measured in the same terms used for information rate, it is the rate at which information symbol errors occur.

Operational Constraints -- Operational constraints are those which require (or preclude) operation under certain conditions -- for example, the requirement that the system be operable by a one-man crew.

Electronic Counter-Countermeasures -- Electronic counter-countermeasures requirements are imposed to prevent an adversary from interfering with the operation of the system. Electronic countermeasures are discussed in the text on electronic warfare.

Secure Operation -- Secure operation requires design features that prevent unintended recipients from recovering the information contained in the signal.

Covert Operation -- Covert operation requires that the probability of intercept or detection of the transmission be acceptably low. Such a requirement is often independent of a requirement for secure operation.

Reliability and Maintainability -- Requirements for reliability and maintainability are intended to enhance operational readiness. As systems become more complex, reliability and maintainability can be ensured only by measures taken early in the design process.

Electromagnetic Compatibility -- Electromagnetic compatibility requires that no electromagnetic effect produced by the equipment in question adversely affect the operation of other equipment or vice versa.

Environmental Tolerance -- Environmental tolerance is the ability to operate satisfactorily under specified environmental stresses including temperature, pressure (altitude), vibration, acceleration, shock, moisture, and radiation.

Hardware Constraints -- Hardware constraints place restrictions on the size, weight, power consumption, and other physical characteristics of system equipment.

Development, Production, and Operation Costs -- To an increasing extent, the life-cycle costs of airborne systems are affecting system design at all stages of development.

3.1.3 Communications System Design Features -- The principal design features available to the designer of a communications system are listed below.

Carrier Type
Carrier Frequency and Bandwidth
Carrier Power
Propagation Mode
Antenna Design
Number of Channels
Multiplexing
Redundancy
Modulation
Spectral (Frequency-Dependent) Filtering
Nonlinear (Amplitude-Dependent) Filtering
Digital Encoding
Signal Correlation
Statistical Signal Processing
Encryption
Spread-Spectrum Signal Processing
Operational Techniques

Carrier Type -- Available carrier types are radio frequency waves, optical frequency waves, acoustic waves, electrical currents, mechanical displacements, and fluid pressures. The choice of carrier depends primarily upon the range required, media involved, interference anticipated, need for covert communications, and hardware limitations.

Carrier Frequency -- The optimum carrier frequency is a function of range required, media involved, interference anticipated, electromagnetic compatibility, external constraints, and hardware limitations. A table of electromagnetic spectrum band designations is presented in Figure 3.1.3.1.

Carrier Bandwidth -- The carrier bandwidth of the system depends primarily upon the carrier frequency, information signal bandwidth, interference spectrum, electromagnetic compatibility requirements, need for spread-spectrum processing, and external constraints.

Microwave Frequency Band Letter Designations	
P-band	225 to 390 MHz (133.3 to 76.9 cm)
L-band	390 to 1550 MHz (76.9 to 19.3 cm)
S-band	1550 to 5200 MHz (19.3 to 5.78 cm)
C-band	5200 to 8500 MHz (5.78 to 3.53 cm)
X-band	8500 to 10900 MHz (3.53 to 2.75 cm)
K-band	10.90 to 36.00 KMHz (2.75 to 0.834 cm)
Q-band	36.00 to 46.00 KMHz (8.34 to 6.53 mm)
V-band	46.00 to 56.00 KMHz (6.53 to 5.36 mm)

Frequency Band Nomenclature	
ELF	30 to 300 Hz
ULF	300 to 3000 Hz
VLF	3 to 30 KHz
LF	30 to 300 KHz
MF	300 to 3000 KHz
HF	3 to 30 MHz
VHF	30 to 300 MHz
UHF	300 to 3000 MHz
SHF	3 KMHz to 30 KMHz
EHF	30 KMHz to 300 KMHz

Designations Per Air Force Reg. No. 5544	
BAND	FREQUENCY, MHz
A	0-250
B	250-500
C	500-1,000
D	1,000-2,000
E	2,000-3,000
F	3,000-4,000
G	4,000-6,000
H	6,000-8,000
I	8,000-10,000
J	10,000-20,000
K	20,000-40,000
L	40,000-60,000
M	60,000-1000,000

Figure 3.1.3.1 -- Electromagnetic Spectrum Band Designations

Carrier Power -- Carrier power is determined by range required, media involved, interference anticipated, electromagnetic compatibility requirements, operational requirements, hardware limitations, and external constraints.

Propagation Mode -- The mode of propagation is selected primarily on the basis of the range required. Electrical conductors and waveguides provide many advantages but have obvious limitations with respect to range. As discussed in Section 2.4.4 of this text, the basic modes of unconfined electromagnetic wave propagation are space wave, ground wave, sky wave, and ducted wave. The relative advantages and disadvantages of these propagation modes are discussed in that section.

Antenna Design -- The design of the optimum antenna is a function of carrier type, frequency, bandwidth, and power, range and directional coverage requirements, interference anticipated, covert communications requirements, and hardware limitations. The impact of all of these factors is discussed in Section 2.6 of this text. In the selection of an antenna design, three fundamental questions must be posed:

(1) Is omnidirectional coverage required, or can an antenna with directive gain be employed.

(2) Is circular polarization required or can a simpler, plane polarized antenna be employed.

(3) Is a broad-band antenna required or can a narrow-band antenna be employed.

The answers to these three questions lead to a selection of antenna design as indicated in Figure 3.1.3.2. Further selection of antenna design can be made

System Requirements			
Pattern	Polarization	Bandwidth	Antenna Type
Omnidirectional Antenna	Linearly Polarized	Narrow Band	Dipole
			Loop
		Broad Band	Biconical
			Swastika
	Circularly Polarized	Narrow Band	Normal Mode Helix
		Broad Band	Biconical w/ Polarizer
			Lindenblad
			4-Arm Conical Spiral
Directional Antenna	Linearly Polarized	Narrow Band	Yagi
			Dipole Array
		Broad Band	Log Periodic
			Horn
	Circulatory Polarized	Narrow Band	Axial Mode Helix
			Horns w/ Polarizers
		Broad Band	Cavity Spirals
			Conical Spirals

Figure 3.1.3.2 -- Antenna Selection Criteria

on the basis of the antenna characteristics presented in Figure 3.1.3.3. In Figure 3.1.3.4 are shown typical dipole array, slot array, and parabolic reflector antennas.

Number of Channels -- Multiple channels are employed to increase the information-carrying capacity of a communications system.

Multiplexing -- The information rate capability of a communications system can be increased by using multiple channels (carriers) or by time multiplexing or frequency multiplexing the information signals as discussed in Section 2.5.8 of this text. With time multiplexing, two or more information signals time share a single carrier. With frequency multiplexing, two or more information signals are impressed upon individual carriers (called sub-carriers). The sub-carriers are then impressed upon a single carrier. The information signals thus occupy different portions of the spectrum.

Redundancy -- Redundancy is employed for reliability and for noise rejection.

Modulation -- The choice of Modulation depends upon allowable carrier bandwidth, information signal content, information rate, interference anticipated, acceptable error rate, and encryption requirements. The relative advantages of AM, FM, PM, and PCM are discussed in Section 2.5.5 of this text.

Spectral Filtering -- Frequency-dependent amplification or attenuation is used to discriminate between the information signal and noise when the two signals differ in spectral content. It is also used to demodulate a frequency-modulated

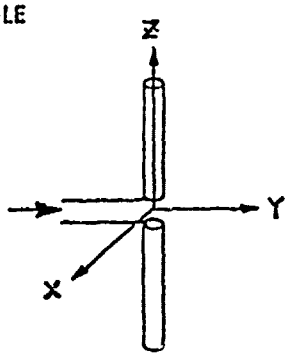
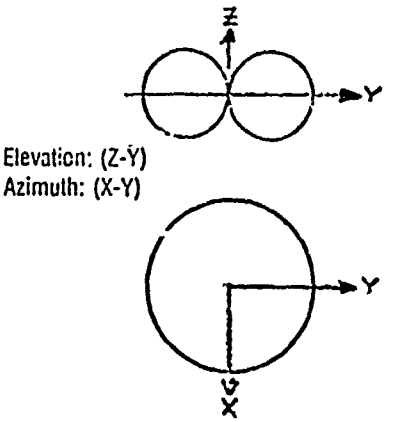
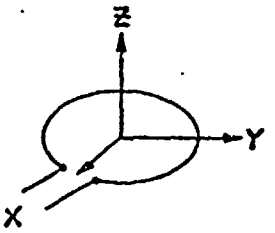
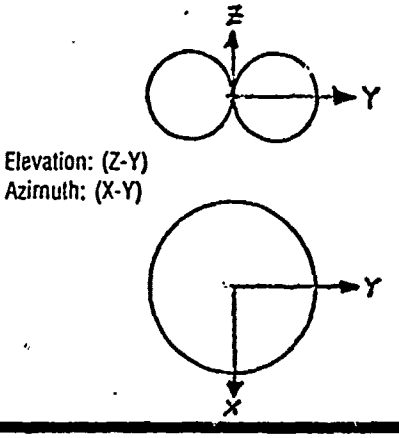
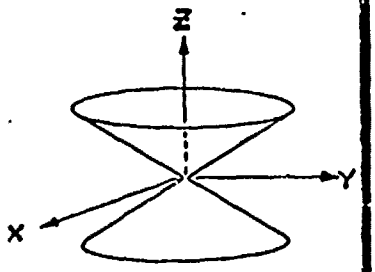
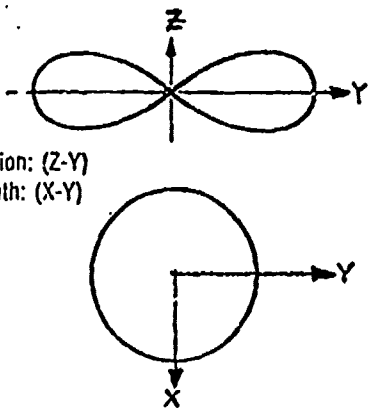
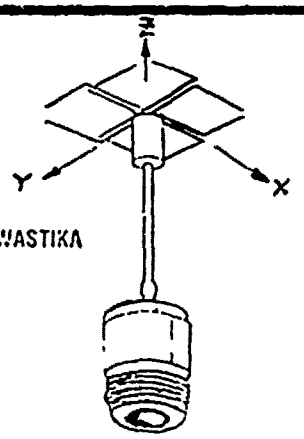
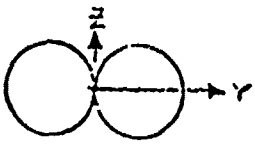
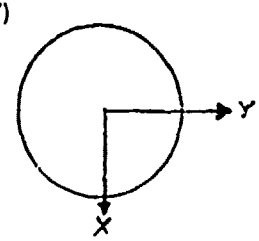
Antenna Type	Radiation Pattern	Specifications
DIPOLE 	 Elevation: (Z-Y) Azimuth: (X-Y)	Polarization Vertical Typical Half-Power Beamwidth 80° X 360° Typical Gain 2 dB Bandwidth 10% Frequency Limit Lower: None Upper: 8GHz
LOOP 	 Elevation: (Z-Y) Azimuth: (X-Y)	Polarization Horizontal Typical Half-Power Beamwidth 80° X 360° Typical Gain -2dB Bandwidth 10% Frequency Limit Lower: 50MHz Upper: 1 GHz
BICONICAL 	 Elevation: (Z-Y) Azimuth: (X-Y)	Polarization Vertical Typical Half-Power Beamwidth 20° - 100° X 360° Typical Gain 0 to 4 dB Bandwidth 4:1 Frequency Limit Lower: 500 MHz Upper: 40 GHz
SWASTIKA 	 Elevation: (Z-Y) Azimuth: (X-Y) 	Polarization Horizontal Typical Half-Power Beamwidth 80° X 360° Typical Gain -1 dB Bandwidth 2:1 Frequency Limit Lower: 100 MHz Upper: 12 GHz

Figure 3.1.3.3 -- Antenna Characteristics

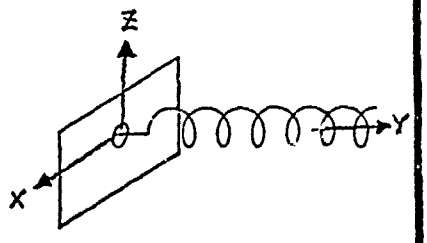
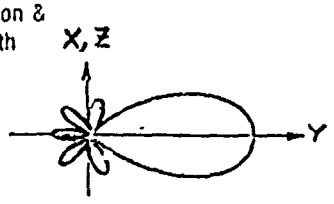
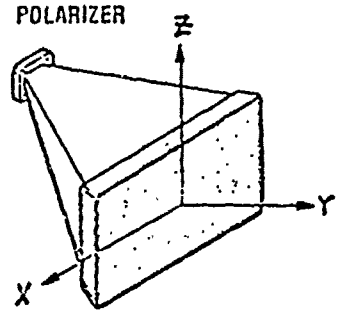
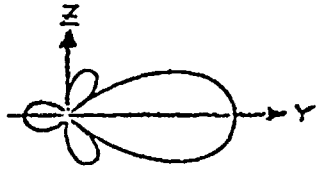
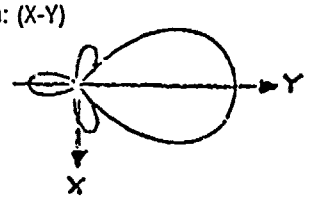
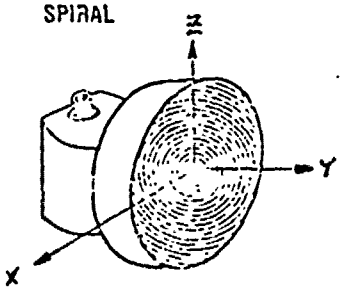
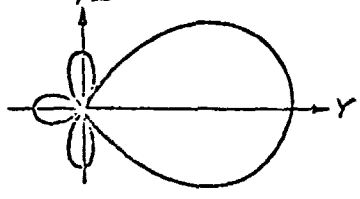
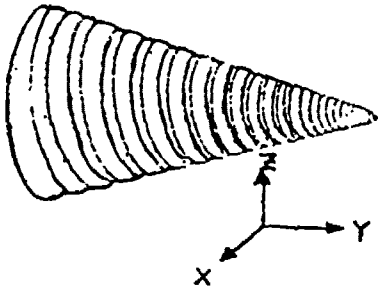
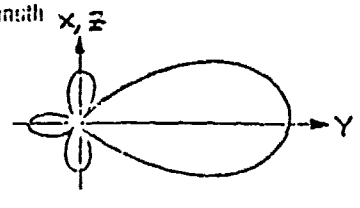
Antenna Type	Radiation Pattern	Specifications
AXIAL MODE HELIX 	Elevation & Azimuth X, Z 	Polarization Circ: Typ ca: alf-Power Be; mw. 50° X Typical Gain 10 dB Bandwidt.. 1.7:1 Frequency Limit Lower: 100 MH_z Upper: 3 GH_z
HORNS WITH POLARIZER 	 Elevation: (Z-Y) Azimuth: (X-Y) 	Polarization Circular Typical Half-Power Beamwidth 40° X 40° Typical Gain 5 to 10 dB Bandwidth 3:1 Frequency Limit Lower: 2 GH_z Upper: 18 GH_z
CAVITY BACKED SPIRAL 	Elevation & Azimuth X, Z 	Polarization Circular Typical Half-Power Beamwidth 60° X 90° Typical Gain 2 to 4 dB Bandwidth 9:1 Frequency Limit Lower: 500 MH_z Upper: 18 GH_z
CONICAL SPIRAL 	Elevation & Azimuth X, Z 	Polarization Circular Typical Half-Power Beamwidth 60° X 60° Typical Gain 5 to 8 dB Bandwidth 4:1 Frequency Limit Lower: 50 MH_z Upper: 18 GH_z

Figure 3.1.3.3 -- Antenna Characteristics (cont'd)

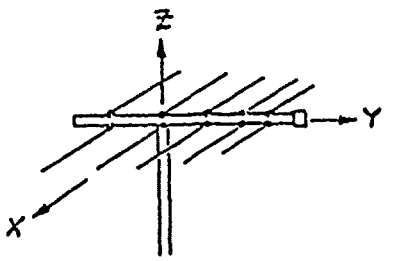
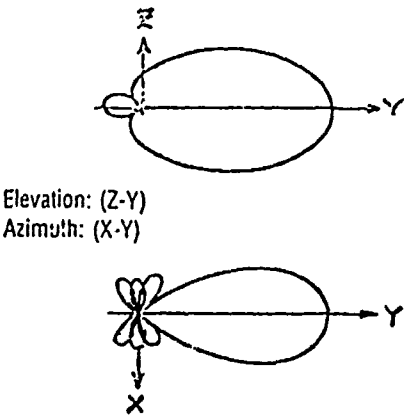
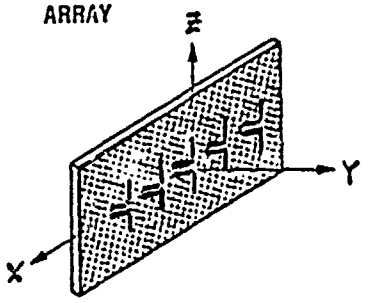
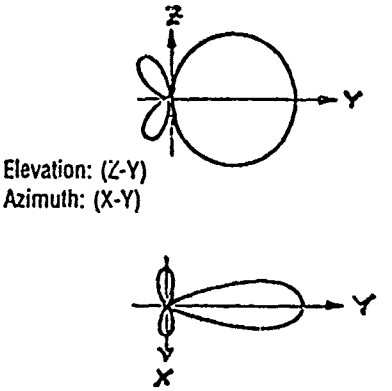
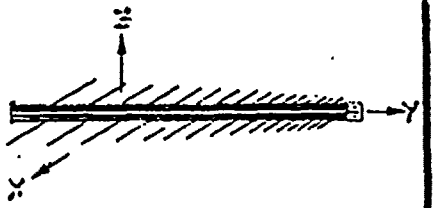
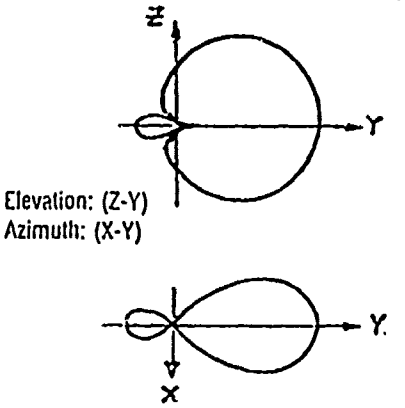
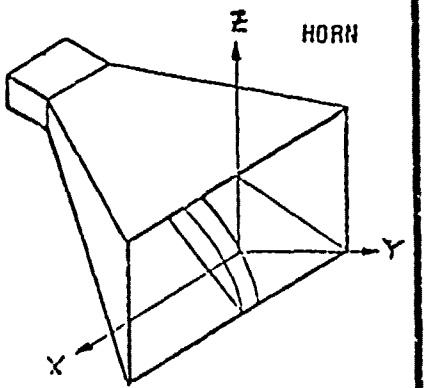
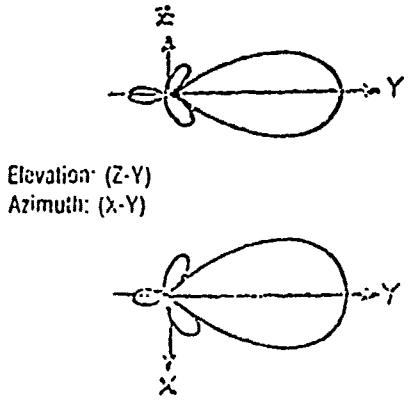
Antenna Type	Radiation Pattern	Specifications
YAGI 	 Elevation: (Z-Y) Azimuth: (X-Y)	Polarization Linear Typical Half-Power Beamwidth 50° X 50° Typical Gain 5 to 15 dB Bandwidth 5% Frequency Limit Lower: 50 MHz Upper: 2 GHz
DIPOLE ARRAY 	 Elevation: (Z-Y) Azimuth: (X-Y)	Polarization Linear Typical Half-Power Beamwidth 20° X 20° Typical Gain 5 to 30 dB Bandwidth 10% Frequency Limit Lower: 50 MHz Upper: 4 GHz
LOG PERIODIC 	 Elevation: (Z-Y) Azimuth: (X-Y)	Polarization Linear Typical Half-Power Beamwidth 60° X 80° Typical Gain 6 to 8 dB Bandwidth 10.1 Frequency Limit Lower: 3 MHz Upper: 18 GHz
HORN 	 Elevation: (Z-Y) Azimuth: (X-Y)	Polarization Linear Typical Half-Power Beamwidth 40° X 40° Typical Gain 5 to 20 dB Bandwidth 4:1 Frequency Limit Lower: 50 MHz Upper: 40 GHz

Figure 3.1.3.3 -- Antenna Characteristics (cont'd)

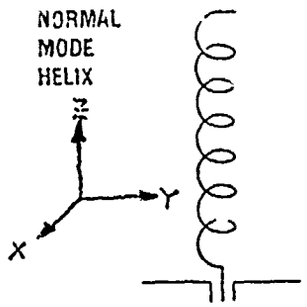
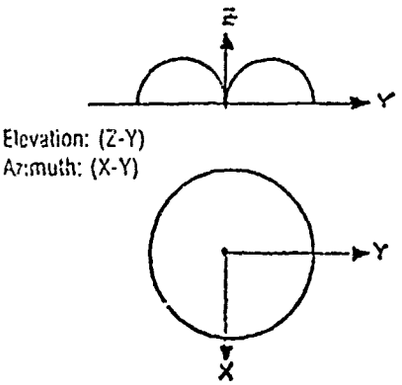
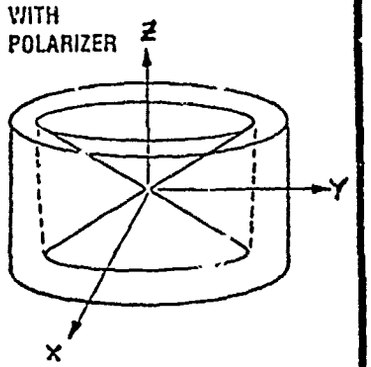
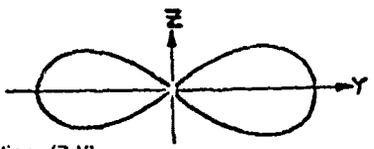
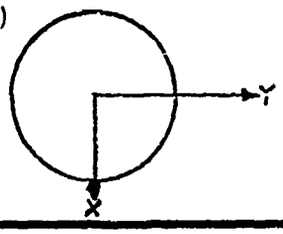
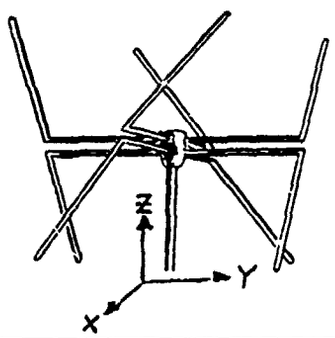
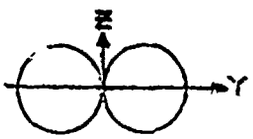
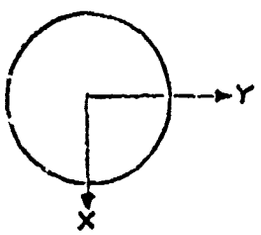
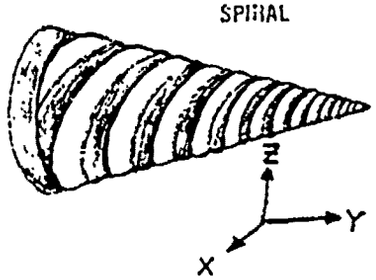
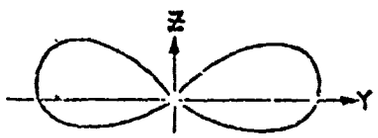
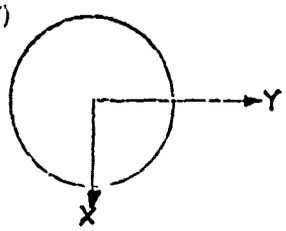
Antenna Type	Radiation Pattern	Specifications
NORMAL MODE HELIX 	 Elevation: (Z-Y) Azimuth: (X-Y)	Polarization Circular Typical Half-Power Beamwidth 60° X 360° Typical Gain 0 dB Bandwidth 5% Frequency Limit Lower: 100 MHz Upper: 3 GHz
BICONICAL WITH POLARIZER 	 Elevation: (Z-Y) Azimuth: (X-Y) 	Polarization Circular Typical Half-Power Beamwidth 20° -100° X 360° Typical Gain 0 to 4 dB Bandwidth 3:1 Frequency Limit Lower: 2 GHz Upper: 18 GHz
LINDENBLAD 	 Elevation: (Z-Y) Azimuth: (X-Y) 	Polarization Circular Typical Half-Power Beamwidth 80° X 360° Typical Gain - 1 dB Bandwidth 2:1 Frequency Limit Lower: 100 MHz Upper: 12 GHz
4-ARM CONICAL SPIRAL 	 Elevation (Z-Y) Azimuth: (X-Y) 	Polarization Circular Typical Half-Powr Beamwidth 50° X 360° Typical Gain 0 dB Bandwidth 4:1 Frequency Limit Lower: 500 MHZ Upper: 18 GHz

Figure 3.1.3.3 -- Antenna Characteristics (cont'd)

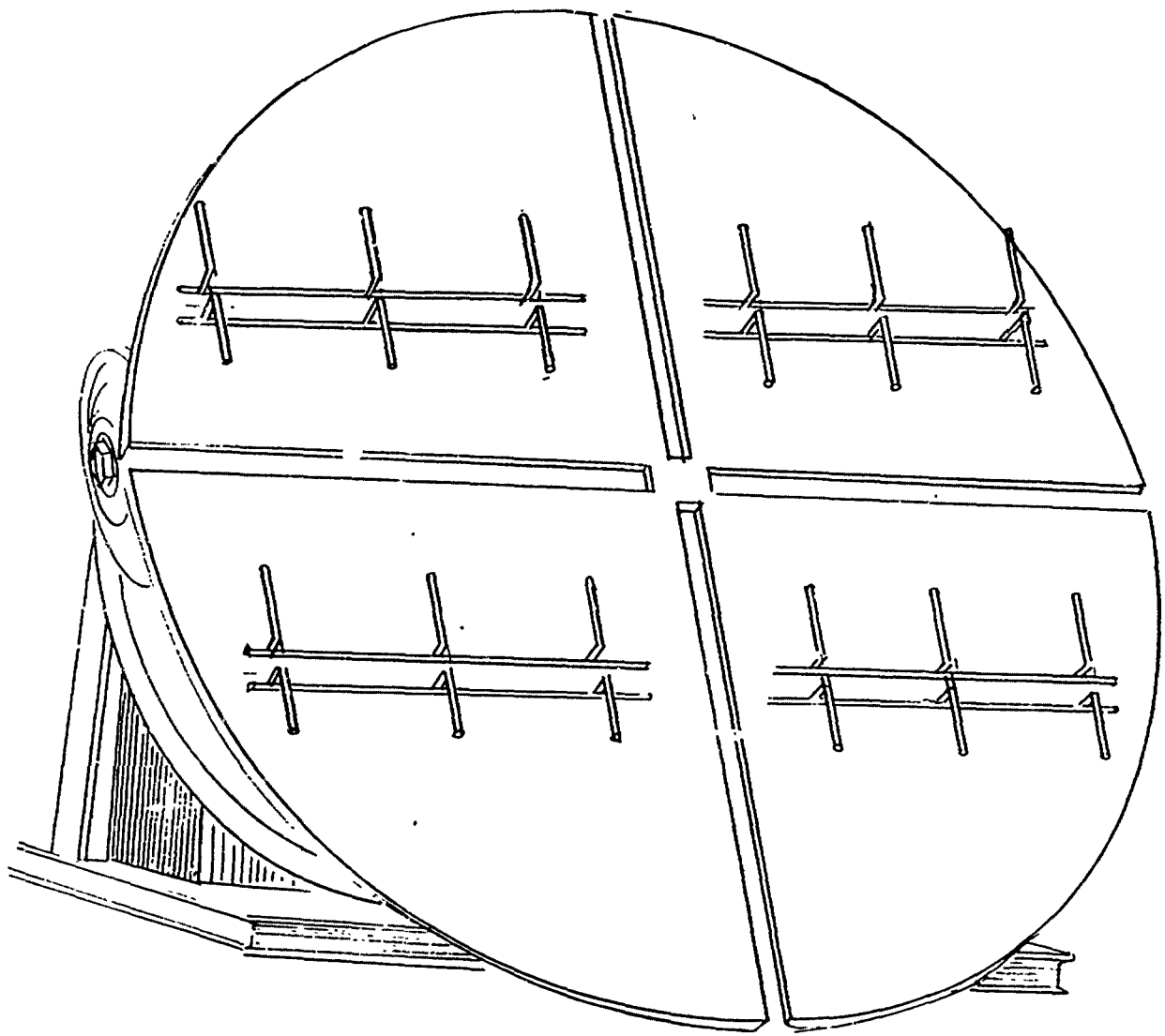


Figure 3.1.3.4 (a) -- Dipole Planar Array Antenna

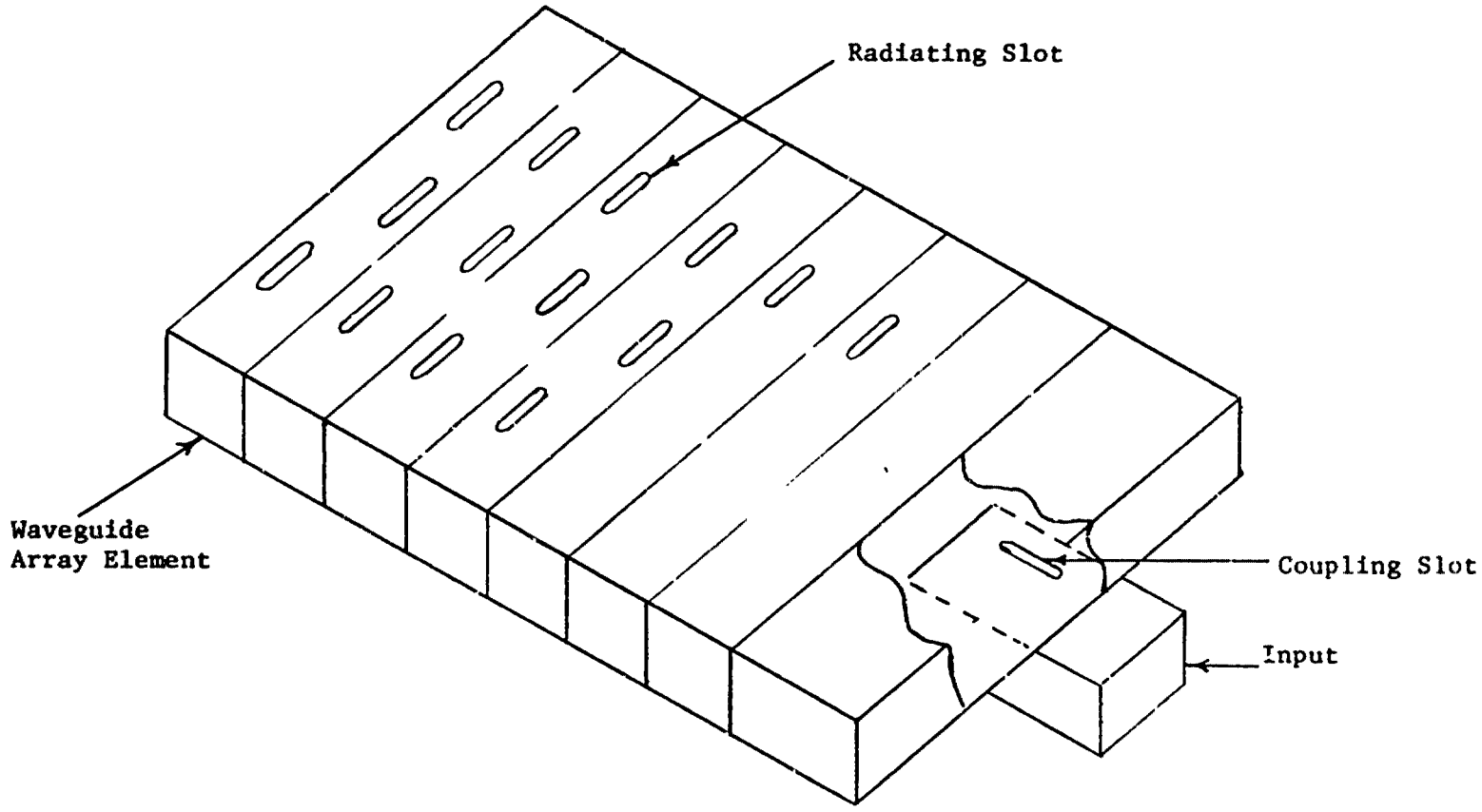


Figure 3.1.3.4 (b) -- Planar Slot Array Antenna

Parabolic Reflector Antenna

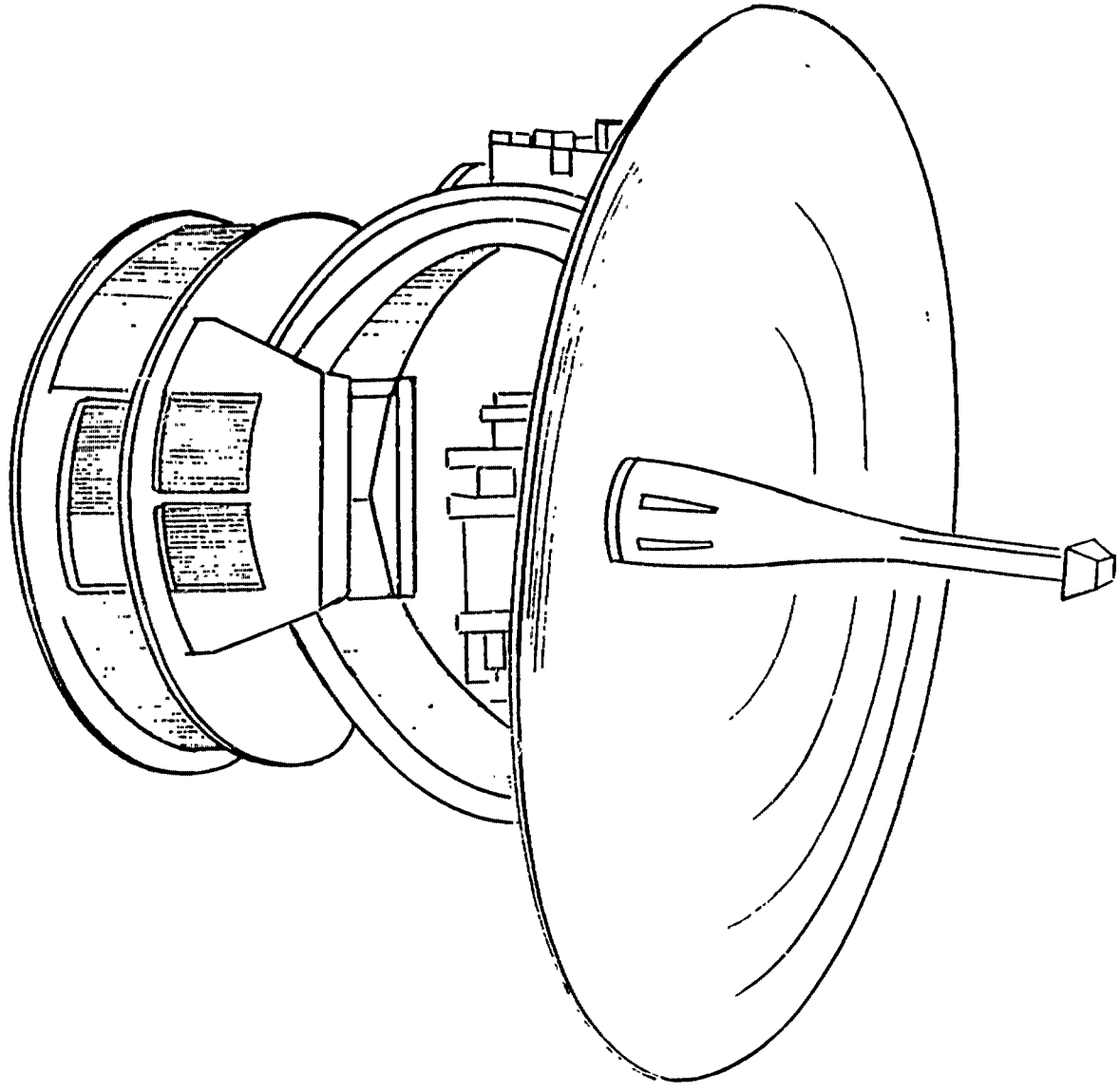


Figure 3.1.3.4 (c) -- Parabolic Reflector Antenna

carrier. Spectral filtering is discussed in Section 2.5.3 of this text.

Amplitude Filtering -- Amplitude-dependent amplification or attenuation is used to discriminate against, or in favor of, large variations in the amplitude of the information-bearing signal. The effects of nonlinear signal processing are discussed in Section 2.5.2 of this text.

Digital Encoding -- The use of digital encoding depends upon the factors concerned with the use of PCM as mentioned above. The characteristics of digital systems are discussed in detail in Section 2.2 of this text.

Encryption -- Encryption of the information signal is employed when secure (covered) communication is required.

Signal Correlation -- Time correlation and frequency or phase correlation are employed in order to discriminate in favor of signals similar, in time history or frequency content, to a reference signal. Signal correlation is discussed in Section 2.5.7 of this text.

Statistical Signal Processing -- Statistical processing is the processing of one or more signals to derive some statistic of the signals. The most common statistical process is the time averaging of a signal to remove random noise. Another example is the use of the Kalman Filter to obtain an optimal estimate of the signal from noisy measurements. In all such schemes, the purpose is to improve the signal-to-noise ratio.

Spread Spectrum Signal Processing -- Spread spectrum processing is employed for encryption, covert communication (low probability of intercept), selective addressing, code multiplexing, noise rejection, anti-jamming, and high resolution radar. These applications, and the general advantages of spread-spectrum signal processing, are discussed in detail in Section 2.5.8 of this text.

Operational Techniques -- The communications system designer sometimes overcomes design deficiencies by requiring (or prohibiting) certain operational practices. Such operational limitations and their consequences should be thoroughly examined in test and evaluation.

3.2 Communication System Characteristics

3.2.1 Types of Airborne Communications Systems - The major types of airborne "communications" systems are those listed below.

Voice Communication
Data Link
Radio Navigation
Sensor
Identification, Friend or Foe (IFF)
Weapon Control
Electronic Warfare

The characteristics of airborne voice communication and data link systems will be discussed in this section. Radio navigation and IFF systems are discussed in the Navigation text. Airborne sensors (radar, infrared, acoustic, etc.) are discussed in the Radar, Electro-Optical, and Undersea Warfare texts. Weapon control systems are discussed in the Integrated Weapon System text, and Electronic Warfare systems are discussed in the Electronic Warfare text.

3.2.2 Definitions of Communication System Characteristics -- The major communication system characteristics are listed below with brief definitions.

Transmitter Characteristics

Output Power -- The average power of the unmodulated carrier emitted by a transmitter.

Output Frequency Accuracy -- The accuracy with which the frequency of the unmodulated carrier can be set and maintained.

Output Frequency Spectrum -- The power density-versus-frequency characteristics of the transmitter output.

Output Impedance -- The internal impedance at the output terminals of a transmitter.

Modulation Level -- The degree of modulation on a carrier. Amplitude modulation is determined by the variation in amplitude of the modulation envelope. Frequency modulation is determined by the frequency deviation of the modulated carrier. Phase modulation is determined by the shift in the phase of the modulated carrier.

Modulation Bandwidth -- The allowable frequency range of the modulating signal, as defined by the half-power points in the frequency response of the modulation system.

Modulation Distortion -- The distortion present in the modulation impressed upon the carrier, due to nonlinearity, cross-modulation, etc.. Distortion is often determined by measuring the power contained in the harmonics generated by the distortion (harmonic distortion).

Carrier Noise Level -- The amount of non-informational signal present on the carrier. Noise level is usually expressed as signal-to-noise or signal plus noise-to-noise ratio.

Input Impedance -- The internal impedance at the input terminals (modulation input) of a transmitter.

Receiver Characteristics

Selectivity -- The sharpness of tuning (bandwidth) of the frequency-selective (tuned) circuits of a receiver.

Threshold Sensitivity -- The minimum signal level detectable by a receiver. Usually, the input signal for which the output signal-to-noise ratio is a specified value.

Sensitivity (Gain) -- The ratio of output level to input (modulation) level of a receiver.

Information Signal Bandwidth -- The range of frequencies allowable in the information signal, usually defined by the one-half power points of the allowable spectrum.

Output Distortion -- The distortion present in the output signal, (recovered modulation or information signal), due to non-linearities, cross-modulation, etc..

Output Noise Level -- The amount of non-informational signal in the output signal, usually expressed as signal-to-noise or signal plus noise-to-noise ratio.

Output Power -- The maximum average power supplied by the output of a receiver

Output Variation (AGC Function) -- The variation in the output level of a receiver, due to variations in the input level. (The purpose of the automatic gain control (AGC) is to prevent such variations.)

Squelch Level -- The input signal level above which the squelch circuit of a receiver opens (allows an output). (The squelch circuit measures the level of the input signal and blocks the output if the input level is below a set value).

Noise Limiting -- The input level above which the response of a receiver is deliberately reduced (limited). (Noise limiting circuits are intended to prevent high output levels due to high noise levels).

Input Impedance -- The internal impedance at the input terminals of a receiver.

Output Impedance -- The internal impedance at the output terminals of a receiver.

Transmission Line and Antenna Characteristics

Voltage Standing Wave Ratio (VSWR) -- The ratio of the maximum to the minimum amplitudes of a standing wave on a transmission line. VSWR is a measure of the ratio of reflected power to forward power. It is also a measure of impedance mismatch in the line. (See Section 2.1.20 of this text.)

Transmission Line Loss -- The loss of power (attenuation) of a signal due to transmission through a transmission line. Line loss is usually expressed, in decibels, as the ratio of output power to input power.

Maximum Power Rating -- The maximum allowable input power to a transmission line or antenna. Limitations are due to heating (temperature rise) and arcing or corona discharge. Some antennas have internal components that can be damaged by excessive power.

Antenna Efficiency -- The ratio of the power radiated by an antenna as useful carrier to the input power. The power loss is primarily due to internal heating.

Antenna Polarization -- The type, degree, and direction of polarization of the radiation transmitted (or received) by an antenna. (See Section 2.4.1 of this text.)

Radome Transmissivity -- The degree to which a radome transmits radiation without power loss.

Antenna Power Gain Pattern -- A three-dimensional representation of the signal level (power density) of the electromagnetic energy radiated by an antenna, as a function of spatial position. Also called antenna field pattern and antenna gain pattern. Antenna patterns are usually presented in two-dimensional plots, two or more being used to provide three-dimensional information.

Integrated System Characteristics

Signal Strength -- The power level or magnitude of a signal.

Signal Clarity -- The intelligibility or "readability" of a signal.

Signal Noise Level -- The amount of interfering noise in a signal.

Maximum Functional Range -- The maximum distance between two communication systems at which reliable communications can be performed.

Built-In-Test Operation -- The ability of the BIT provisions to detect system faults. Typical detected faults are: low output power, frequency error, high VSWR, no modulation, high noise level, and power supply failure.

3.2.3 Airborne Voice Communication Systems -- Airborne voice communications systems provide air-to-air and air-to-ground, two-way, voice communication. U.S. Naval airborne voice communication systems are generally UHF systems employing amplitude modulation. (U. S. Army and civilian aircraft generally use VHF frequency-modulated systems.) The reasons for the Navy's use of amplitude modulation are primarily historical. The inherent advantages of FM systems with respect to signal-to-noise ratio makes likely increasing use of FM in future systems. The advantages of digital processing makes likely an increasing use of digital systems. UHF systems are limited to line-of-sight operation. At higher altitudes, however, power (rather than radio horizon), sometimes limits range. (Typical UHF transmitter power is ten to twenty watts.) Many airborne UHF antennas are quarter-wave, skin-mounted monopoles (stubs or blades). The "image" reflection in the aircraft skin provides the other half of an effective dipole. Secure or covered communication is often provided by direct sequence spread spectrum processing. The above remarks apply to the majority of U. S. Naval installations. Special-purpose systems exist that have quite different characteristics; for example, the air-to-submarine communication systems that operate at much greater power, in the VLF frequency range, and utilize long, trailing-wire antennas.

The characteristics of typical U. S. Naval voice communication systems are presented in the following tables.

System Characteristics

Item: AN/ARC-159 Radio Set
Type: Voice Communication and Automatic Direction Finding
Maker: Collins Radio
Specification: MIL-R-81877A(AS)
Vehicles: A-4, RF-4B, A-6E, EA-6B, A-7, F-14, F-18, H-1, LAMPS
Frequency: UHF (225 MHz to 400 MHz)
Propagation Mode: Line-of-Sight
Modulation: AM
No. of Channels: 7000 (Receives two channels simultaneously:tuned channel and 243 MHz guard channel).
Encryption: Yes (Bandwidth provided for use with KY-58)
Spread Spectrum: Yes -- direct sequence encoding for encryption.
Form Factor (Size): 5.75"(W)x4.875"(H)x9.25"(L).
Weight: 9.25 Lbs.
Input Power: 28 V.D.C. @ 4.45 Amps
Operating Life: 10,000 Hours
Mean Time Between Failures: 1000 Hours

Transmitter Characteristics

Frequency: 225.000 to 399.975 MHz
Frequency Accuracy: $\pm$ 2 KHz
Power: 10 Watts
Modulation: AM, 90-100% with standard input.
Distortion, Harmonic: <10%
Input Voltage: 0.5v, rms (Unsec.); 12v, p-p (sec).
Input Impedance: 150 Ohms (Unsec.); 2000 Ohms (Sec.)
Input (Audio) Bandwidth: 300 to 3500 Hz (Unsec.)
Input (Audio) Bandwidth: 300 to 25,000 Hz (Sec.)
Output Impedance: 50 Ohms

Receiver Characteristics

Frequency: 225.000 to 399.975 MHz
Frequency Accuracy: $\pm$ 2 KHz
Selectivity (6 dB Bandwidth): 38 KHz (Unsec.); 70 KHz (Sec.)
Sensitivity: 3 Microvolts for S+N/N = 10 dB.
Threshold Sensitivity: 1.5 Microvolts
Input Impedance: 50 Ohms
Noise Figure (Equiv. Out.): S+N/N = 30 dB (with Std. Input)
Noise Peak Limiting: Limit at 50% Modulation
Audio Bandwidth (+ 1, -3 dB): 300 to 3500 Hz (Unsec.)
Audio Bandwidth (+ 2, -4 dB): 300 to 25000 Hz (Sec.)
Output Impedance: 150 Ohms or 600 Ohms (Unsec.)
Output Impedance: 1000 Ohms (Sec.)
Output Power (Audio): 250 Milliwatts, Max. (Unsec.)
Output Power (Audio): 250 Millivolts (with Std. Input)
Non-linear Signal Processing: Input Limiting, Squelch, Automatic Gain Control

Antenna Characteristics (MIL-A-6224E)

Type: Quarter-Wave Dipole (Stub) or Broad-Band Dipole (Butterfly) (Typical)
Frequency: 225 to 400 MHz
Polarization: Linear, Vertical
Pattern: Vertical Dipole Pattern (Doughnut) 360° Az; ± 30° and ± 45° El.
Power Gain:
+ 45° to +20° El.: > -2 dB (wrt Isotropic)
+ 20° to -20° El.: > 0 dB (wrt Isotropic)
-20° to -45° El.: > -2 dB (wrt Isotropic)
Efficiency: 0.85 Min. (225 to 400 MHz)
Impedance: 50 Ohms
Max. Input Power: 100 Watts
Physical Size: 9" (Stub); 18" (Butterfly) (Typical)

Transmission Line Characteristics (MIL-A-6224E)

Type: Coaxial Cable
Bandwidth: 0 to 500 MHz
Characteristic Impedance: 50 Ohms
VSWR: < 2.0 (with antenna load)
Line Loss: < 10.5 dB/100 Ft. @ 400 MHz
Max. Power: 500 Watts

Item: AN/ARC-182 Radio Set
Type: Voice Communication, Automatic Direction Finding, and Automatic Homing
Maker: Collins Radio
Specification: NASC AS-4579(AV)
Vehicles: A-4, A-6, A-7, F-4, F-14, F/A-18, AV-8, E-2C, C-130, OV-10, H-1, H-3, H-46, H-53

System Characteristics

Frequency: VHF, 30 to 88, MHz; 108 to 156 MHz; 156 to 174 MHz
UHF, 225 to 400 MHz
Propagation Mode: VHF--Line-of-Sight (some ground wave)
UHF--Line-of-Sight
Modulation: FM--VHF (Lo), VHF (Hi), and UHF
AM--VHF (Mid) and UHF
Number of Channels: 7000 (UHF); 4640 (VHF); (Guard channels for each range)
Encryption: Provides wide-band channels for encrypted signals from KY-58
Spread Spectrum Processing: (Future) Frequency Hopping
Form Factor: 5.75"(W) x 4.875"(H) x 10.00"(L).
Weight: 10 Lbs.
Input Power: 28 VDC @ 100 Watts (Transmit), 20 Watts (Receive)
Operating Life: 20,000 Hours
Mean Time Between Failures: 1000 Hours

Transmitter Characteristics

Frequency: VHF, 30 to 87.975 MHz; 118 to 155.975 MHz; 156 to 173.975 MHz
UHF, 225 to 399.975 MHz
Frequency Accuracy: $\pm$1 part per million
Power: 10 Watts (AM); 15 Watts (FM)
Modulation: AM, 80 to 100% (with standard input)
FM, $\pm$5.6 KHz Deviation (with standard input)
Distortion: <5% (In demodulated audio)
Input Voltage: 0.5v, rms (Unsec.); 12 v, p-p (Sec.)
Input Impedance: 150 Ohms (Unsec.); 2000 Ohms (Sec.)
Input Bandwidth: 500 to 3500 Hz (Unsec.)
10 to 20,000 Hz (Sec.)
Output Impedance: 50 Ohms

Receiver Characteristics

Frequency: Same as Transmitter plus 108 to 117.975 MHz (AM)
Selectivity (6 dB Bandwidth): 37 KHz (Unsec); 69 KHz (Sec); 28 KHz (Guard)
Demodulation: See transmitter characteristics.
Sensitivity (Threshold): VHF (Lo) = -117 dBm; VHF (Mid) =-103 dBm;
VHF (Hi) = -113 dBm; UHF (AM) =-103 dBm; UHF (FM) =-113 dBm
Sensitivity: S+N/N $\geq$ 10 dB for threshold inputs
Noise Figure: AM, -53 dBm Std. Input produces S+N/N $\geq$ 30 dB.
FM, -70 dBm Std. Input produces S+N/N $\geq$ 40 dB.
Input Impedance: 50 Ohms
Audio Bandwidth: (+1, -3dB), 500 to 3500 Hz (Unsec.)
(+2, -4dB), 10 to 25,000 Hz (Sec.)
Output Impedance: 150 Ohms or 600 Ohms (Unsec.)
20,000 Ohms (Sec.)
Output Power: 0 to 250 mw (600 Ω); 200 mw (150 Ω)
Output (Audio) Distortion: < 5% (with standard input)

3.2.4 Airborne Tactical Data Links -- Tactical data links provide for the transmission of digitized information. Both analog and digital quantities are conveyed, the analog quantities being digitized before transmission. (Analog data telemetry systems are sometimes used in airborne test and evaluation. These systems utilize FM, PWM, PPM, and other analog forms of modulation.) The advantages of digital systems are discussed in Sections 2.5.5 and 2.6 of this text. The Joint Tactical Information Distribution System (JTIDS) also provides digitized voice communication.

Current applications of tactical data links are various command, control, and communications (C^3) functions: intercept vectoring, automatic targeting and bombing, exchange of sensor data, automatic carrier landing (ACLS), and CAINS inertial navigation system alignment. Current U. S. Naval airborne tactical data links employ UHF (Link 4 and Link 11) and HF (Link 11). The UHF systems are restricted in range to line-of-sight operation but have lower probability of intercept and relative freedom from interference. The HF systems have ranges well beyond line-of-sight, but are more subject to interference and interception. To gain the greater range, the HF systems utilize much greater transmitted power (1000 watts) than do the UHF systems (100 watts). The Link 11 system provides for encrypted transmission, thus somewhat alleviating the security problem. There is no provision for secure communications with the Link 4 system. Neither system has special provision for anti-jamming protection.

U. S. Naval aircraft utilizing the Link 4 system include the EA-6, A-7, F-4, F-14, F-18, S-3, and E-2C. Aircraft utilizing the Link 11 system include the P-3, S-3, and E-2C. Special purpose data links are employed in a few other aircraft. The F-14 has a Link-4A two-way data link system designated the AN/ASW-27. No

rotary-wing U. S. Naval aircraft employs either the Link 4 or the Link 11 system. The LAMPS, Mark III utilizes a special purpose tactical data link from the aircraft to the ship.

The information (data) rates for the Link 4 and Link 11 systems are quite low (5 KHz). This data rate limits, for example, the maximum number of aircraft with which an E-2C can simultaneously communicate. Commonality requirements are an obstacle to upgrading the data rate.

The antennas and transmission lines employed with data links are those commonly used for the appropriate frequency bands.

The characteristics of typical U. S. Naval tactical data link systems are presented in the following tables.

Item: AN/ASW-25B Receiver
Type: Link 4A Tactical Data Link Receiver
Maker: Harris Corp.
Specification No.: MIL-D-81124B(AS)
Vehicles: A-6, A-7, F-4, F-18, S-3

System Characteristics

Frequency: UHF (300 to 325 MHz)
Modulation: PCM (70 - Bit word); FSK (± 20 KHz)
Data Rate: 5 KHz
No. of Channels: 249
Encryption: No
Spread Spectrum Provision: No. (Planned for future).
Multiplexing: No
Error Correcting Code: No. (Parity checks only).
Anti-Jamming Code: No

Receiver Characteristics

Frequency: 300.000 to 324.900 MHz
Bandwidth (3 dB): 70 KHz
Demodulation: PCM
Sensitivity: None specified.
Threshold Sensitivity: 5.6 Microvolts
Input Impedance: 50 Ohms
Noise Figure (Eq. Input): None specified.
Digital Outputs: Open Collector (10 ma sink).
Analog Outputs: 0 to 26 V, rms; 400 cycles; 250 Ohms
0 to ±2.2 V, dc; 1000 Ohms

Item: Joint Tactical Information Distribution System (JTIDS)

Currently under development is a multi-purpose system for all U. S. military platforms called the Joint Tactical Information Distribution System (JTIDS). JTIDS is an integrated, communications, navigation, and identification (CNI) network utilizing a single, wide-band (960 to 1215 MHZ) channel with time division multiple access (TDMA) multiplexing.

The use of a pseudo-random encoding signal to control the time multiplexing process provides secure communication and spread-spectrum signal processing. In addition, provision is made for direct sequence encoding for encryption and frequency hopping to further spread the spectrum of the transmitted signal. Error detecting codes are employed to provide for noise rejection and error correction. The result is a secure, noise-rejecting, jam-protected, digital, error correcting, high-data-rate communications system that provides air-to-air, air-to-surface, and surface-to-surface data link and voice communication. Transmission beyond the line-of-sight is achieved by providing for each member of the net to act as a relay repeater.

JTIDS provides a relative navigation (ranging) capability by virtue of the precise time synchronization required for the time multiplexing process. In addition, the system also provides standard TACAN navigation. An identification-Friend or Foe (IFF) capability is provided by the address codes assigned to members of the net. In addition to manned aircraft operations, JTIDS will be used as a data link for ground and amphibious operations and for guidance and control of RPV's and tactical missiles.

4.0 Communication System Performance Test and Evaluation

4.1 Levels of Testing

4.1.1 Stages of Testing -- Testing can be categorized as developmental, functional, or operational, depending upon the stage of development of the test item. Developmental testing is concerned with the evaluation of design features for the purpose of design development. The end result of developmental testing is the proposed final design. Functional testing is concerned with the performance evaluation of the final design as a whole. The principal method of evaluation is the quantitative measurement of the ability of the test item to perform its intended functions. The end result of functional testing is final design acceptance or rejection. Operational testing is concerned with the evaluation of the final design and production implementation of the test item. Of primary interest is the ability of the test item to accomplish its intended operational mission. The end result of operational testing is acceptance or rejection of the test item for service use and the recommendation of operational procedures.

4.1.2 Testing Criteria -- The basic purpose of any stage of testing determines the criteria used to evaluate the test results. The testing criteria, in turn, are reflected in the tests to be performed and the test methods employed. Testing criteria derive from one of three objectives: data acquisition, determination of specification compliance, and evaluation of mission performance. In developmental testing the intent is to acquire comprehensive information on the

characteristics of the item under test. Usually, no a-priori criteria are imposed for performance acceptance or rejection. Functional testing, however, is primarily intended to evaluate the performance of the test item against specific criteria -- that is, for specification compliance. As previously indicated, operational testing is primarily concerned with mission performance. While some specific, quantitative requirements are imposed, test criteria for operational testing often are of a qualitative nature.

It should be recognized that the three stages of testing; developmental, functional, and operational; are not mutually exclusive. That is, the differences are primarily ones of emphasis. For example, functional testing often produces data that result in a design change. Thus, functional testing often takes on some aspects of developmental testing. For that reason, it is necessary, in functional testing, to test to a depth sufficient to allow engineering analysis of the problem. A "go" or "no-go" answer often is not sufficient. On the other hand, functional testing cannot ignore mission suitability in evaluating a new design. Compliance with published specifications is not sufficient if functional testing reveals an operational problem. Thus, while the following sections of this text will be concerned primarily with quantitative tests for specification compliance, it should be noted that functional testing should reflect mission requirements, including non-quantitative considerations when appropriate.

4.1.3 Test Regimes -- Functional airborne system tests are performed in the laboratory, in the aircraft on the ground, and in flight. For various reasons, testing is usually performed in that order. Tests performed on the bench in the laboratory are most convenient, quickest, cheapest, and safest. Flight

tests are least convenient, take the longest time, are most costly, and present the greatest danger to personnel and equipment. They also are most susceptible to uncertainties in the weather and availability of equipment. For the above reasons, tests should be performed in the laboratory, before installation in the aircraft, when feasible. Tests that can only be performed installed in the aircraft should be performed on the ground when feasible. Flight tests should be performed only when necessary and only when laboratory and ground tests have reduced the uncertainties to the greatest extent feasible. Of course, some tests can be performed only in flight; and, in any event, flight performance eventually must be demonstrated.

In the following sections, brief descriptions will be given of the methods employed to determine system compliance with the major communication system functional specifications. General testing, such as environmental, electromagnetic compatibility, reliability, and maintainability testing, is discussed in a separate text devoted to tests common to all airborne systems.

4.2 Communications Transmitter Testing

4.2.1 Transmitter Test Setup -- In Figure 4.2.1.1 is shown the block diagram for a typical communications transmitter test setup. The Information Signal Generator supplies the input (modulating) signal to the transmitter. A power meter with integral load is used to measure the output signal power level. A signal coupler is inserted in the transmission line to extract a small amount of the modulated carrier power for use as inputs to the various measurement devices shown in the figure. A frequency meter, (generally a digital counter),

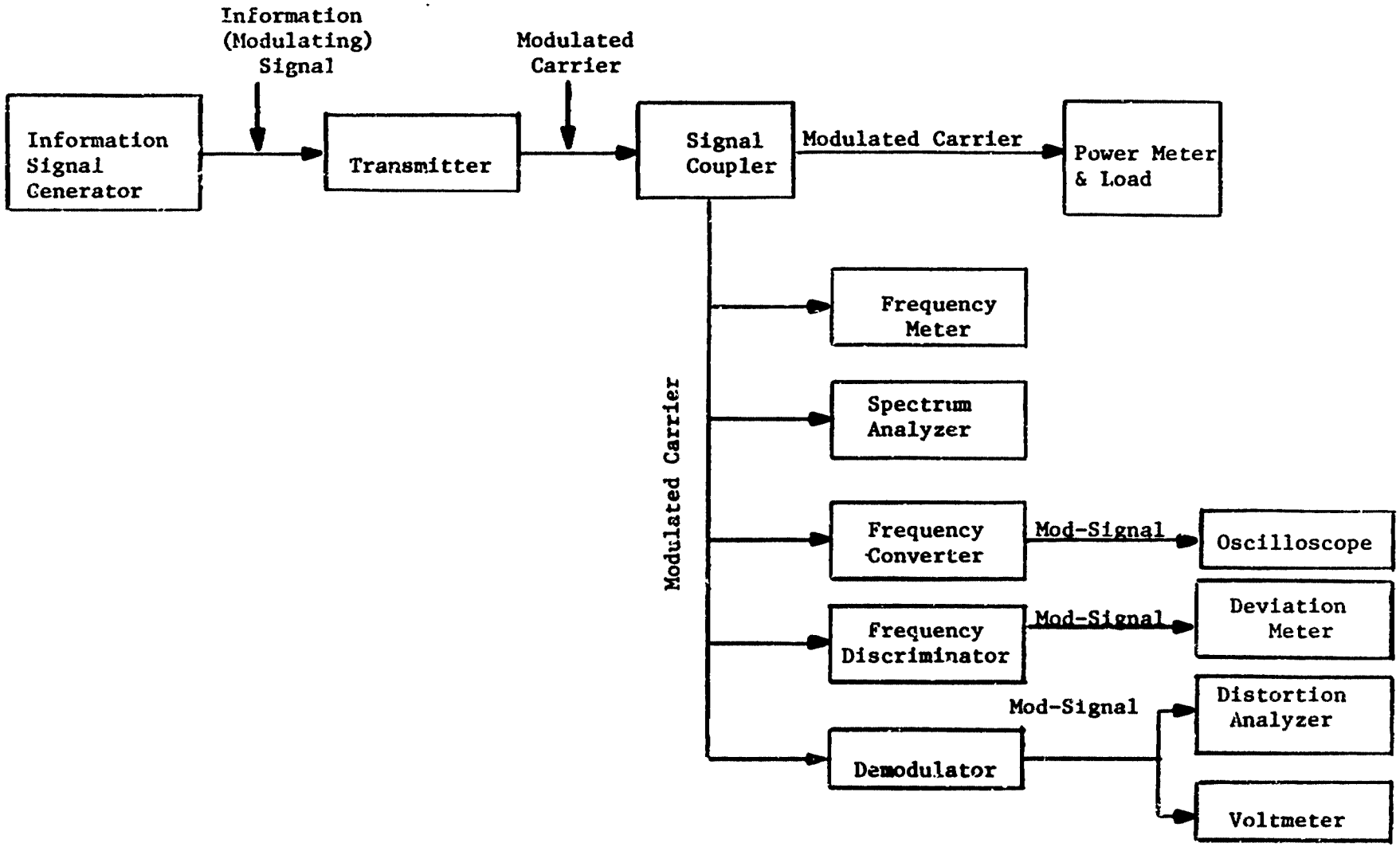


Figure 4.2.1.1 -- Communications Transmitter Test Setup

is used to determine the dominant frequency of the carrier. A spectrum analyzer is used to determine its spectral content (power density as a function of frequency). The modulation level of an amplitude modulated carrier is determined by viewing the modulation envelope on an oscilloscope. A frequency converter is used to shift the carrier frequency to a range suitable for oscilloscope input. The modulation level of a frequency modulated carrier is determined by demodulating the carrier, using a frequency discriminator, and measuring the frequency deviation. A distortion analyzer is used to determine the distortion present in the modulation signal recovered from the modulated carrier by a demodulator or detector.

4.2.2 Transmitter Test Methods -- The major transmitter characteristics are listed below, with brief descriptions of relevant test procedures.

1. Output Power

   a. Set modulation to zero.
   b. Set load to specified value.
   c. Set carrier frequency to specified value.
   d. Key transmitter.
   e. Record power.

2. Frequency Accuracy

   a. Set modulation to zero.
   b. Set load to specified value.
   c. Set carrier frequency to specified value.
   d. Key transmitter.
   e. Record frequency.
   f. Repeat steps c through e as required.

3. Output Frequency Spectrum

   a. Set modulation as specified.
   b. Set load to specified value.
   c. Set carrier frequency to specified value.
   d. Key transmitter.
   e. Record spectral data.
   f. Repeat steps c through e as required.

4. Output Impedance

a. Set modulation to zero.
b. Set carrier frequency to specified value.
c. Set load to specified value.
d. Key transmitter.
e. Adjust load to maximize output power.
f. Record load impedance (Max. power).

5. Modulation Level

a. Set load to specified value.
b. Set carrier frequency as specified.
c. Set modulation as specified.
d. Key transmitter.
e. Measure modulation envelope values and calculate percentage of modulation (AM); or, measure frequency deviation (FM).
f. Record modulation level.
g. Repeat steps c through f as required.

6. Modulation Bandwidth

a. Set load to specified value.
b. Set carrier frequency to specified value.
c. Set modulation voltage to specified value.
d. Set modulation frequency as specified.
e. Key transmitter.
f. Measure and record modulation signal level.
g. Repeat steps d through f as required.

7. Modulation Distortion

a. Set load to specified value.
b. Set carrier frequency to specified value.
c. Set modulation as specified.
d. Key transmitter.
e. Measure and record distortion in demodulated output.

8. Carrier Noise Level

a. Set load to specified value.
b. Set carrier frequency to specified value.
c. Set modulation to zero.
d. Key transmitter.
e. Measure and record level of demodulated output (noise).

9. Input Impedance

a. Set load to specified value.
b. Set carrier frequency to specified value.
c. Set modulation to specified value.
d. Key transmitter.
e. Measure and record input impedance or voltage and current.

4.3 Communications Receiver Testing

4.3.1 Receiver Test Setup -- In Figure 4.3.1.1 is shown the block diagram for a typical communications receiver test setup. The signal generator supplies the input signal (modulated carrier) to the receiver under test. A passive padding network is included in the input circuit to ensure impedance matching and provide standard attenuation. A voltmeter and a power meter are used to measure the output voltage and power, respectively. An oscilloscope is used to view output voltage waveforms and a distortion analyzer is used to measure output (modulation or information signal) distortion.

4.3.2 Receiver Test Methods -- The major receiver characteristics are listed below, with brief descriptions of relevant test procedures.

1. Selectivity

   a. Set modulation as specified.
   b. Set carrier level to specified value.
   c. Set carrier frequency as specified.
   d. Measure output power level.
   e. Vary carrier frequency to determine frequencies corresponding to output half-power points.
   f. Record bandwidth.
   g. Repeat steps d through f as required.

2. Threshold Sensitivity

   a. Set modulation as specified.
   b. Set carrier frequency as specified.
   c. Set carrier level to zero.
   d. Measure output (noise) power.
   e. Increase carrier level until output power increases by specified amount (typically 10 dB).
   f. Record carrier level.
   g. Repeat steps b through f as required.

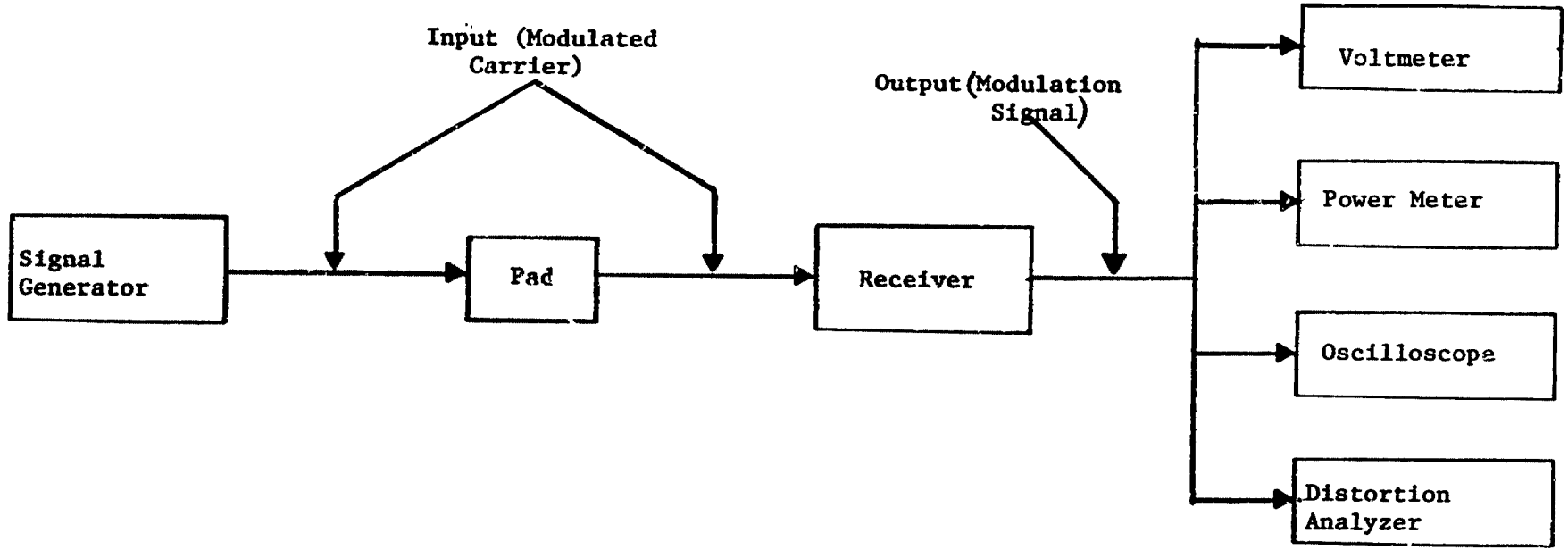


Figure 4.3.1.1 -- Communications Receiver Test Setup

3. Sensitivity (Gain)

a. Set carrier frequency as specified.
b. Set carrier level as specified.
c. Set modulation to zero.
d. Measure output power.
e. Set modulation as specified (below AGC level).
f. Measure output power.
g. Determine and record sensitivity as ratio of increase in output power to input level.

4. Information Signal Bandwidth

a. Set carrier frequency as specified.
b. Set carrier level as specified.
c. Set modulation level to specified value.
d. Set modulation frequency as specified.
e. Measure and record output level at half-power points.
f. Repeat steps d and e as required.

5. Output Distortion

a. Set carrier frequency as specified.
b. Set carrier level as specified.
c. Set modulation as specified.
d. Measure and record output distortion.

6. Output Noise Level

a. Set carrier frequency as specified.
b. Set carrier level as specified.
c. Set modulation to zero.
d. Measure and record output (noise) power.

7. Output Power

a. Set carrier frequency as specified.
b. Set carrier level as specified.
c. Set modulation as specified.
d. Measure and record output power.

8. Output Variation (AGC Function)

a. Set modulation as specified.
b. Set carrier frequency as specified.
c. Set carrier level as specified.
d. Measure and record output power.
e. Repeat steps c and d as required.

9. Squelch Level Adjustment

a. Set modulation as specified.
b. Set carrier frequency as specified.
c. Set carrier level as specified.
d. Verify that squelch control can be set to just open at specified carrier level.

10. Noise Limiter Operation

a. Set carrier frequency as specified.
b. Set carrier level as specified.
c. Increase modulation level until output saturation occurs.
d. Verify specified saturation level.

11. Input Impedance

a. Set carrier frequency as specified.
b. Set carrier level well below point where AGC activates.
c. Set modulation as specified.
d. Measure and record input impedance or input voltage and current.

12. Output Impedance

a. Set carrier frequency as specified.
b. Set carrier level as specified.
c. Set modulation as specified.
d. Measure output power.
e. Vary output load resistance to maximize output power.
f. Record load impedance (Max. power).

4.4 Antenna and Transmission Line Testing

4.4.1 Transmission Line and Antenna Electrical Test Setup -- The block diagram for a typical transmission line and antenna electrical test setup is shown in Figure 4.4.1.1. The power oscillator supplies an unmodulated carrier, through the transmission line, to the test antenna. Directional couplers are inserted into the oscillator-transmission line-antenna path to tap off a small, known fraction of the forward and reflected powers. In this manner, the incident and reflected powers are determined at the inputs to the transmission line and the antenna. It should be noted that VSWR and line loss testing is often performed during integrated system testing, utilizing the test transmitter, rather than a signal generator, to supply the unmodulated carrier.

4.4.2 Transmission Line and Antenna Electrical Test Methods -- The major transmission line and antenna electrical characteristics are listed below, with brief descriptions of relevant test procedures.

1. <u>Voltage Standing Wave Ratio (VSWR) for Transmission Line/Antenna Combination</u>

   a. Set carrier frequency as specified.
   b. Measure and record forward and reflected powers at coupler A.
   c. Repeat steps a and b as required.
   d. Calculate reflection coefficient and VSWR, and record versus frequency.

2. <u>Voltage Standing Wave Ratio (VSWR) for Antenna Only</u>

   a. Set carrier frequency as specified.
   b. Measure and record forward and reflected powers at coupler B.
   c. Repeat steps a and b as required.
   d. Calculate reflection coefficient and VSWR, and record versus frequency.

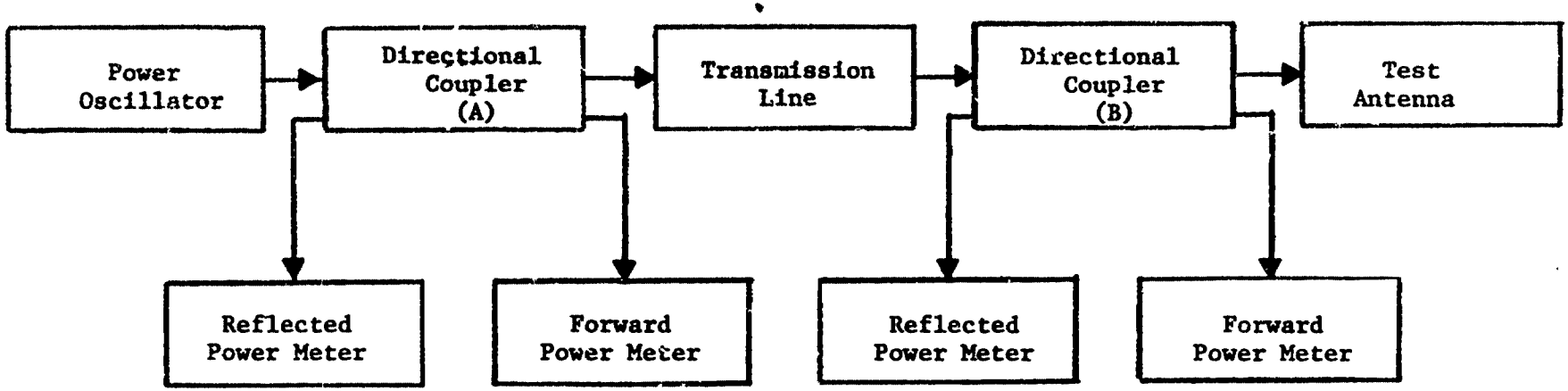


Figure 4.4.1.1 -- Transmission Line and Antenna Electrical Test Setup

3. Transmission Line Loss

a. Set carrier frequency as specified.
b. Measure and record forward powers at couplers A and B.
c. Repeat steps a and b as required.
d. Calculate power loss in line, and record versus frequency.

4. Maximum Power Rating

a. Apply specified maximum power to transmission line and antenna.
b. Monitor line and antenna for excessive temperature rise.
c. Monitor line and antenna for arc-over.

5. Antenna Efficiency

a. Set carrier frequency.
b. Measure forward and reflected power at coupler B.
c. Calculate input power to antenna and record.
d. Measure and record total radiated power using special antenna "hat".
e. Repeat steps a through d as required.

4.4.3 Antenna Radiation Test Setup -- The block diagram for a typical antenna radiation test setup is shown in Figure 4.4.3.1. The power oscillator supplies an unmodulated carrier to the test antenna. A directional coupler is inserted at the input to the antenna in order to extract a small, known fraction of the forward and reflected powers. These power samples are measured to determine the power delivered to the antenna. The transmitted carrier is received by a calibrated antenna and receiver, thus allowing measurements to be made of the radiated field power density at the point in space at which the receiving antenna is located. Utilizing the communications (range) equation, calculations can be made of the field power density that an isotropic radiating antenna would produce at that point in space, thus allowing determination of the absolute power gain of the test antenna as a function of space position. An alternate method of obtaining absolute calibration of the system is to substitute another calibrated antenna for the test antenna, thus providing for comparison of the test antenna characteristics to those of the (known) calibrated antenna, and hence to those of an isotropic antenna.

Measurements of airborne antenna gain patterns are often performed on the ground. The test antenna can be installed on the aircraft or mounted on a text fixture with a ground plane, if required. The effects of radome transmissivity and other structural effects can be evaluated by measuring the field both with and without the accompanying structures.

In order to reduce the size of the required antenna test range, ground or laboratory measurements of antenna field patterns can be made utilizing reduced-scale models of the test antenna, surrounding structure, and test range. In order to preserve the required ratio of antenna dimensions to wavelength, the

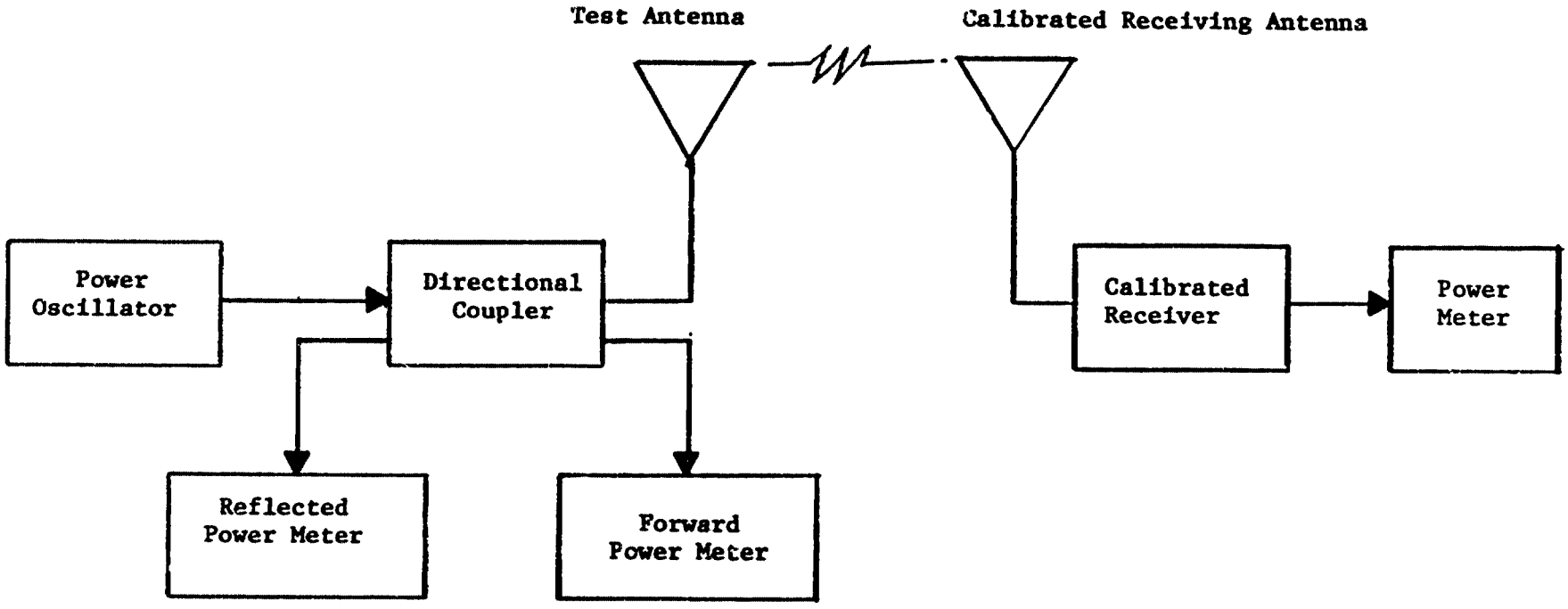


Figure 4.4.3.1 -- Antenna Radiation Test Setup

frequency of the radiation must be scaled up by the same ratio by which the model is scaled down.

Another method of reducing the dimensions of the required test range is that of employing near-field test techniques. This method makes use of the fact that, if the characteristics of the field in the immediate vicinity of an antenna (near field) are known, the characteristics of the far field can be calculated using mathematical transform techniques.

The most reliable method of determining the radiant field pattern of an airborne antenna is to measure it with the antenna installed as intended, on the proper aircraft, and in flight. The spatial position of the test antenna is then best determined by ground tracking equipment and test antenna orientation is best determined by instrumentation on the test aircraft. By virtue of the reciprocity principle, an antenna can be tested in either the transmitting or the receiving mode.

4.4.4 Antenna Radiation Test Methods -- The major antenna radiation characteristics are listed below, with brief descriptions of relevant test procedures.

1. Polarization

   a. Set carrier frequency as specified.
   b. Properly orient a plane-polarized receiving antenna.
   c. Measure and record the received signal strength (power).
   d. Repeat steps b and c, varying the orientation of the receiving antenna to determine the type, degree, and direction of polarization of the test antenna.

2. Radome Transmissivity

   a. Set carrier frequency as specified.
   b. Position test antenna as specified.
   c. Position receiving antenna as specified.
   d. Measure and record received power.
   e. Install radome as specified.
   f. Measure and record received power.
   g. Remove radome.
   h. Repeat steps a through g as required.

3. Antenna Power Gain Pattern

   a. Set carrier frequency and power level as specified.
   b. Position and orient the test antenna, with respect to the receiving antenna, as specified.
   c. Measure received signal strength and record.
   d. Measure space position of test antenna.
   e. Repeat steps a through d as required.

4.5 Integrated Communications System (Transmitter/Receiver) Testing

4.5.1 System (Transmitter/Receiver) Test Setup -- In Figure 4.5.1.1 is shown the block diagram for a typical communications transmitter/receiver test setup. The system under test is installed in the test aircraft and evaluated as an integral unit. Two-way communication is conducted with the calibrated ground station in order to evaluate functional performance and the quality of communications, including such factors as signal strength, clarity, and noise level. The calibrated ground station and power meter provide quantitative measurement of signal strength. Two-way communication tests are conducted both on the ground and in flight. Maximum functional range is determined by continuously conducting communications while increasing the range between the ground station and the test aircraft. Maximum range is that range, for a given altitude, at which communication becomes unreliable. Tests are also performed during various aircraft operational maneuvers and flight configurations, (landing gear, flaps, etc.), to determine the effects of these factors.

The operation of the built-in-test provisions of the system are tested by "injecting" faults into the system and noting the success or failure of the BIT to detect the fault. When such tests require access to the internal components of the system, they are normally performed in the laboratory. When "faults" can be injected externally, they can be performed installed in the aircraft. The specific "faults" to be injected depend upon the BIT design of the system under test. Typical "faults" are antenna or transmission line impedance mismatch, low output power, no modulation, and power supply failure.

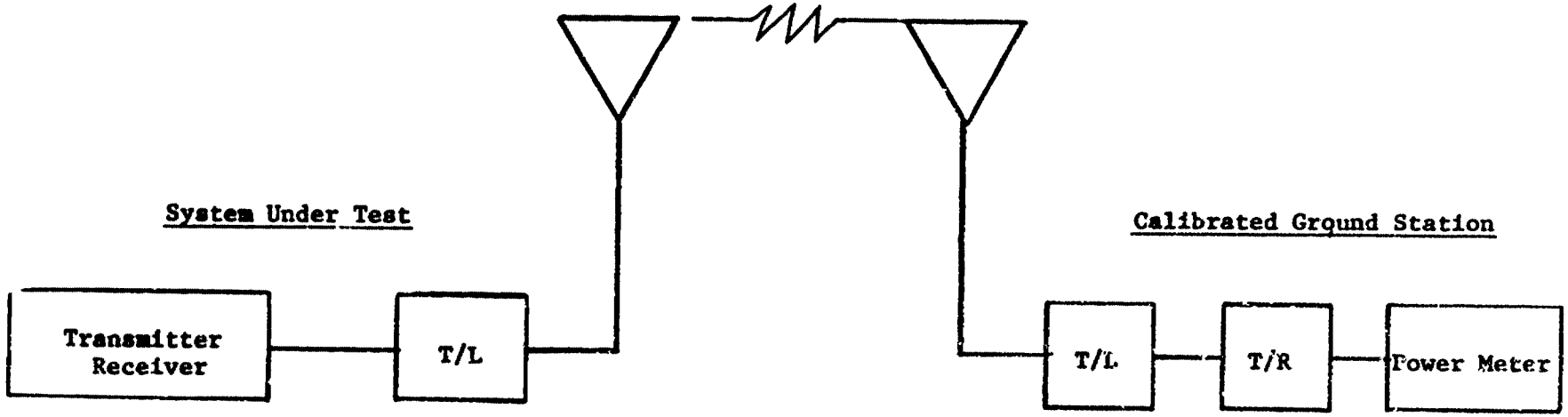


Figure 4.5.1.1 — System (Transmitter/Receiver) Test Setup

4.5.2 System (Transmitter/Receiver) Test Methods -- The major characteristics of an integrated communication system (transmitter, receiver, transmission line, and antenna) are listed below, with brief descriptions of the relevant test procedures.

1. Two-Way Communication (Ground)

   a. Set T/R frequency and mode.
   b. Establish and maintain two-way communications.
   c. Operate appropriate system controls.
   d. Observe and record system function.
   e. Repeat steps a through d as required.

2. Two-Way Communication (Flight)

   a. Set T/R frequency and mode.
   b. Establish and maintain two-way communications.
   c. Fly prescribed flight path (direction and altitude).
   d. Perform prescribed maneuvers at specified points in flight path.
   e. Note and record quality of communications (signal strength, clarity, noise level, etc.).
   f. Repeat steps a through e as required.

3. Maximum Functional Range

   a. Set T/R frequency and mode.
   b. Establish and maintain two-way communications.
   c. Fly prescribed flight path (specified direction at constant altitude) to systematically increase range between test system and ground station.
   d. Perform banking and turning maneuvers at specified ranges.
   e. Note and record quality of communications.
   f. Repeat steps a through e as required.

4. Built-In-Test Operation

   a. Establish proper system operation.
   b. Inject "fault" as specified.
   c. Note and record success or failure of the BIT to detect "fault".
   d. Repeat steps a through c as required.