THE DESIGN OF HF COMMUNICATIONS SYSTEMS
FOR SIGNAL QUALITY OPTIMIZATION

by

WALTER WILLIAM JACOBS, B.Sc.

PRESENTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
THE UNIVERSITY OF EDINBURGH
1977
ABSTRACT

The design of conventional HF communication systems with respect to signal quality is critically appraised and weaknesses are discussed. From a comparison of analogue and digital HF systems against a set of derived criteria, it is deduced that at present the optimum HF system must be based on some form of biphase coded signalling technique such as FSK and DPSK. This optimum is then examined for performance in a Rayleigh fading multipath channel. This study suggests that the matched filtering techniques used in conventional HF systems could be improved.

On this basis, the relatively new technologies of Surface Acoustic Wave (SAW) devices are investigated for application in HF system design. From this study an HF digital communication system based on radar pulse compression techniques is described using SAW matched filters and a method of time compression using CCDs and digital devices.

Finally it is proposed that the spread spectrum techniques, made more feasible by the advent of these technologies, are applied to the HF system design problem, and a tentative design in described and discussed.

It is hoped that the findings of this investigation will open the way towards the realization of an HF communication system which will offer the economic, technical and strategic advantages of ionospheric radio propagation without many of the disadvantages which make HF radio communication non-competitive with present-day satellites and undersea cables.
ACKNOWLEDGEMENTS

The author wishes to express his gratitude to his supervisors, Professor W.E.J. Farvis and Dr. L.M. Muggleton for their invaluable guidance, encouragement and advice throughout the course of this research. In particular Professor Farvis is thanked for his undying faith in the author and his constant introduction of new ideas and helpful discussion of the author's own ideas.

Thanks are also extended to the author's colleagues at Microwave, Electronic Systems Limited and in particular to Dr. M.B.N Butler who helped with his comprehensive knowledge of Surface Acoustic Wave devices and their use in radar pulse compression systems.

Grateful thanks are also given to Miss Fiona Hunter who typed the manuscript and Miss Gillian Sharp who drew the diagrams.

The author is also indebted to his father and mother - Sir Wilfred and Lady Jacobs - for their unending spiritual and tangible support during this work; and to his family-in-laws to be - the Sharps - for their constant faith, morale-boosting and encouragement.

Finally the author extends his loving gratitude to his wife-to-be, Rosemary, for emotional support and stability during this research.
## CONTENTS

### Chapter 1

General Introduction ......................................................... 1

### Chapter 2

**HF Signal Quality and System Design Fundamentals**

2.1 Introduction .......................................................... 5

2.2 Basic Concepts in HF Signal Processing .......................... 7
  2.2.1 A Measurement of Signal Quality ............................. 12
  2.2.2 Analogue Modulation Techniques ............................ 23
  2.2.3 HF Digital Modulation Techniques ......................... 32

2.3 Fundamental Ionospheric Effects .................................. 38

2.4 Considerations in HF Antenna Design ............................. 41

### Chapter 3

**HF System Design for Optimum Signal Quality**

3.1 Introduction .......................................................... 47

3.2 Criteria For General Optimality .................................. 48

3.3 Optimality of HF Systems ........................................... 51
  3.3.1 Signal Quality Comparison .................................. 51
  3.3.2 Reliability Comparison ...................................... 52
  3.3.3 Complexity Comparison ...................................... 52
  3.3.4 Cost Comparison ............................................... 53
  3.3.5 Bandwidth Comparison ........................................ 53
  3.3.6 Use of Technology Comparison ............................... 54

3.4 HF Digital System Performance ..................................... 56
  3.4.1 The Rayleigh Fading-Multipath Channel .................... 58
  3.4.2 Digital Signalling in Raleigh Fading ...................... 61
  3.4.3 Diversity in Digital Systems ............................... 67
  3.4.4 Error Control Coding ........................................ 74
3.5 The Optimum HF System
  3.5.1 Forward Transmission Systems 86
  3.5.2 Feedback Transmission Systems 87
  3.5.3 Overall System Appraisal 90

Chapter 4
Pulse Compression in HF Digital Data Transmission
  4.1 Introduction 92
  4.2 Matched Filtering Techniques 95
    4.2.1 Digital Matched Filtering 98
    4.2.2 CCD Matched Filtering 101
    4.2.3 SAW Matched Filtering 105
  4.3 Pulse Compression and HF Digital Data Transmission 111

Chapter 5
Proposal for an HF Spread-Spectrum System
  5.1 Introduction 116
  5.2 Spread-Spectrum Communications 118
    5.2.1 SAW Technology 121
    5.2.2 CCD Technology 122
    5.2.3 Digital Device Technology 122
  5.3 Spread-Spectrum Potential in HF Communications 124
    5.3.1 Introduction 124
    5.3.2 Comparison 125
  5.4 Proposal for HF Spread-Spectrum System Implementation 130
    5.4.1 System General 130
    5.4.2 The Spread-Spectrum Modem 133
  5.5 Discussion 136
## Chapter 6

### Summary and Conclusions

6.1 Summary  
6.2 Conclusions

### APPENDICES

#### Appendix 1

**Summary of Ionospheric Radio Propagation Fundamentals**

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1.1</td>
<td>Structure of the Ionosphere</td>
<td></td>
</tr>
<tr>
<td>A1.1.1</td>
<td>The D-Region</td>
<td>141</td>
</tr>
<tr>
<td>A1.1.2</td>
<td>The E-Region</td>
<td>142</td>
</tr>
<tr>
<td>A1.1.3</td>
<td>The F-Region</td>
<td>144</td>
</tr>
<tr>
<td>A1.2</td>
<td>Predicted Ionospheric Parameters</td>
<td>147</td>
</tr>
<tr>
<td>A1.3</td>
<td>Basic Ionospheric Radio Propagation Theory</td>
<td></td>
</tr>
<tr>
<td>A1.3.1</td>
<td>Flat Earth Ionosphere, No Magnetic Field</td>
<td>152</td>
</tr>
<tr>
<td>A1.3.2</td>
<td>Curved Earth Ionosphere - Magnetic Field</td>
<td>153</td>
</tr>
<tr>
<td>A1.4</td>
<td>Ionospheric Path Losses</td>
<td></td>
</tr>
<tr>
<td>A1.4.1</td>
<td>Spatial Attenuation Loss - Ld</td>
<td>158</td>
</tr>
<tr>
<td>A1.4.2</td>
<td>Ionospheric Absorption Loss - La</td>
<td>159</td>
</tr>
<tr>
<td>A1.4.3</td>
<td>Focussing Loss - Lf</td>
<td>162</td>
</tr>
<tr>
<td>A1.5</td>
<td>Fading and Multipath</td>
<td>164</td>
</tr>
<tr>
<td>A1.6</td>
<td>Ionospheric Radio Noise</td>
<td></td>
</tr>
<tr>
<td>A1.6.1</td>
<td>Atmospheric Noise</td>
<td>169</td>
</tr>
<tr>
<td>A1.6.2</td>
<td>Cosmic Noise</td>
<td>173</td>
</tr>
<tr>
<td>A1.7</td>
<td>Ionospheric Disturbances</td>
<td>176</td>
</tr>
</tbody>
</table>

#### Appendix 2

**Computer Aided Design of HF Antenna Systems**

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2.1</td>
<td>The Horizontal Dipole Array (HDA)</td>
<td>178</td>
</tr>
<tr>
<td>A2.2</td>
<td>The Rhombic Antenna</td>
<td>182</td>
</tr>
<tr>
<td>A2.3</td>
<td>The Log Periodic Dipole Array (LPDA)</td>
<td>191</td>
</tr>
</tbody>
</table>
REFERENCES

References to Chapter 1 197
References to Chapter 2 198
References to Chapter 3 200
References to Chapter 4 202
References to Chapter 5 206
References to Appendix 1 208
References to Appendix 2 210
Unless otherwise stated in the text, the following symbols will have the meanings listed below:

- \( A \): carrier amplitude in AM
- \( A_D \): carrier amplitude in DSB
- \( A(f_v) \): absorption as a function of frequency
- \( A_R \): carrier amplitude at receiver
- \( A_S \): carrier amplitude at antenna terminals
- \( A_T \): limit value for absorption
- \( a \): general subscript
- \( B \): general bandwidth
- \( B_D \): doppler spread in multipath channel
- \( B_Q \): bandwidth of random gaussian process
- \( B_N \): receiver noise bandwidth
- \( B_T \): transmitted message bandwidth
- \( E(t) \): generalised message signal
- \( F \): noise figure
- \( f \): generalised frequency
- \( G \): general gain parameter
- \( C_R \): receiver front-end gain
- \( G_T \): transmitter amplifier gain
- \( h \): general height above ground
- \( H_c(f) \): channel transfer function
- \( H_R(f) \): receiver transfer function
- \( H_T(f) \): transmitter transfer function
- \( I \): modulation index
- \( K \): general constant
- \( L \): loss in channel and antenna side length
- \( m(t) \): message signal in AM
n  general number
N  electron concentration
N_{ax}  additive noise power
N_D  destination noise power
N_T  noise temperature
N_x  noise power at general point x
P_P_e  error probability in digital communication
r  sampling rate
R  sunspot number (Zurich)
R_n(t)  envelope function
S  distance of wavefront from antenna
S_D  destination signal power
S_R  received signal power (front-end)
S_S  signal power at antenna
S_T  transmitted signal power
S_x  signal power at general point x
t  general time
T_M  multipath spread factor
v  electron collision frequency
W  bandwidth

\omega_c  carrier frequency

\alpha  recombination coefficient
\gamma  signal-to-noise parameter for analogue systems
\phi  general phase
\eta  noise power
X  sun's zenith angle
\rho  signal-to-noise parameter for digital systems
\lambda  wavelength
\theta  antenna elevation angle
\mu  refractive index
\sigma  standard deviation
CHAPTER 1

GENERAL INTRODUCTION
1.1 Introduction

Since the 1920's High-Frequency (HF) ionospheric radio has been one of the main forms of long-distance communication. Until World War II, the main interest was with radio-telephony, but with the enemy threat to the undersea cable system, emphasis shifted to radio-telegraphy. This shift continued after the war and since then new technological developments and increasing understanding of the ionosphere and its behaviour has led today to a radio communication system which has an important role in long-distance international communication.

Important as this role is, it is nevertheless secondary to that of satellite and transoceanic cable links which provide many more channels with greater reliability and better signal quality than ionospheric radio links.

Since the launching of the first communication satellites in the late 1950's and the extension of the undersea cable system, the bulk of long-distance traffic has been removed from the HF channel. Its major uses at present are for those purposes which do not justify the expense of the multichannel satellite and cable systems. Among these are:

1. Short-wave broadcasting to those parts of the world not yet serviced by satellite links.
2. International overseas ship-to-shore radio-telephony.
3. Commercial and military Telex.
5. Some aircraft navigational aids.
7. Back-up and overflow for the satellite and undersea cable systems.
The majority of these applications involve low data rates and channel capacity and are essentially non-critical with regard to absolute reliability and speed of response between communicating parties. These factors make them well-suited for the narrowband HF ionospheric channel. Even so the HF band is almost saturated and demands for availability and spectrum space are on the increase. It is thus reasonable to expect that there will be a continuing need for the HF channel in long-distance communication.

Besides its necessity, the HF system has many advantages over the other systems. Technologically it is simple in concept, economically it is relatively cheap and from the strategic point of view, the ionosphere, unlike satellites and undersea cables, is practically indestructible.

Thus despite the obvious attractions of satellites and cables for both military and commercial use, they are natural military targets. Should they be destroyed, the only fall-back position for long-distance communication in such emergencies would be the ionosphere. It is therefore judged to be of the uttermost importance that research be continuously carried out to improve and advance the technology in HF communications.

The advantages of the HF channel could be more fully exploited if solutions could be found to some of the major limitations involved in its use. Among these two of the most important are

1. The conventional, but presently necessary, use of narrow channel bandwidths.

2. The relatively low reliability of conventional HF circuits.

These constraints are due partly to the random nature of the ionosphere and partly to the present state-of-the-art in HF technology. The main effect of these constraints is to limit the quality of signal available using conventional HF communication systems.
The essence of this particular study is to understand and overcome by any technological means at our disposal the fundamental weaknesses of the ionosphere as they have been experienced over the past fifty years and to propose solutions which are essentially viable technologically at the present time.

To this end it is first of all necessary to examine all aspects of HF radio communication. This is done in Chapter 2 where the frames of reference for this study are discussed and defined. Appendices I and II supplement this discussion and in themselves represent a large part of the initial work done in this investigation.

In Chapter 3 analogue and digital HF systems are compared with reference to a set of criteria deduced from the implications of the initial work in Chapter 2. From this comparison the system which at present provides optimum signal quality is chosen and its performance in a realistic model - the Rayleigh Fading Multipath Channel - is discussed. This discussion highlights the main shortcomings of the presently available HF communication systems.

Accepting that the conventional system is fairly well optimized, the discussion in Chapter 4 investigates the use of techniques and technologies not normally encountered in HF system design. Of particular interest are the pulse compression techniques used in chirped radar systems, and especially those based on the recently introduced Surface Acoustic Wave (SAW) technology. The use of digital devices and Charged Coupled Devices (CCDs) are also considered in the context of HF pulse compression communication systems.

In this chapter it is also suggested how these new concepts may be applied to the design of an HF system which is much more resistant to the effect of multipath and other deleterious phenomena in the ionosphere, than is the conventional HF system.
In Chapter 5 these concepts are extended to consider the application of spread spectrum techniques in HF system design. This results in a proposal for a new kind of HF system in which the ionosphere is used as a common shared spectrum for all users of the HF band. Successful implementation of this system could go a long way to reducing the necessity for narrowband communication in the HF channel and could increase greatly the signal quality performance of the HF radio communication system.

The findings of this investigation are summarized and concluded in Chapter 6.
CHAPTER 2

HF SIGNAL QUALITY AND SYSTEM DESIGN FUNDAMENTALS
2.1 Introduction

The purpose of a communication system is to convey information from one point to another with as little distortion and interference as possible. The essential elements of any practical communication system are shown in figure (2.1.1).

It begins with a baseband message source which is usually up-converted in frequency before being fed to the transmitter. This upconversion is achieved by modulating a periodic carrier with the message signal. In the transmitter, the modulated carrier is amplified and matched to the transmitting antenna system from which it is launched into the channel. At this point the signal power is denoted by $S_T$.

The channel supports the message-carrying electromagnetic wave on its passage to the receiving antenna. During this passage, the signal suffers some loss, $L$, in power, due to attenuation and other lossy phenomena in the channel. The signal is also usually distorted by the channel. At the output of the channel, the signal power is $S_R$ and the message is contaminated with noise of spectral density $S_N(f)$.

This signal-plus-noise is then reprocessed at the receiver to extract the original baseband message from the carrier, the additive noise, and the distortion introduced in the journey from the transmitter. The output, or destination, signal and noise powers are $S_D$ and $N_D$ respectively in the figure.

This study is about the High Frequency (HF) communication process. Here the channel is the ionosphere, which means that the signal is at all times vulnerable to the diurnal, seasonal and other vagaries of the upper atmosphere. Reception in this channel is seldom free from distortion and other deleterious effects characteristic of ionospheric radio propagation. Besides these, there are those effects
Fig (2.1.1) COMMUNICATION SYSTEM PARAMETERS
introduced by the limitations in signal processing at the transmitters and receivers, and in the efficiency of the antennas, used in HF communication systems.

These effects all act to degrade the quality-of-signal performance in HF communications and are thus of basic importance in this investigation. In this chapter therefore it is intended to define the basic design parameters, and their performance limitations as regards HF signal quality, used in realizing practical HF circuits.
2.2 Basic Concepts in HF Signal Processing

In every communications system, signal processing consists of combinations of some or all of a number of basic operations; these are

1. modulation; demodulation
2. encoding; decoding,
3. amplification; attenuation,
4. additive and multiplicative mixing,
5. filtering.

Each of these operations will affect the quality of the signal being processed.

In the context of communication, this quality depends on the interests and needs of the communicating parties. A concept of such subjective nature is difficult to interpret and define in the practical engineering field of performance assessment. Nevertheless, for the purposes of this investigation, definite criteria must be established for enumerating signal quality.

As a first step in this signal quality may be equated to information intelligibility. It will thus depend on the amount of distortion and interference suffered by the signal during processing and transmission. Distortion will produce variations in amplitude, frequency and phase of the waves; interference will contaminate the signal with spurious information.

Initially therefore signal quality may be defined as 'the extent to which the amplitude, frequency and phase of the received signal equals the corresponding variables of the transmitted wave'. Ideally these parameters would be in exact agreement; in practice their values vary in a manner dependent upon the type and amount of signal processing used at the transmitter and receiver.

Conventional communication in the HF channel is of two distinct
types. Classically there is the voice or analogue channel in which the signal is processed by Amplitude Modulation (AM) in one form or other. More recently there is the rapidly growing and ever improving field of HF digital communication. Here telegraphy is the usual form of message transfer using derivatives of Narrow-Band Frequency Modulation (NBFM) and Narrow-Band Phase Modulation (NBPM); - Frequency-Shift-Keying (FSK) and Phase-Shift-Keying (PSK). These digital techniques are normally employed in various coding formats to utilize available channel bandwidth more efficiently, and for increased speed and accuracy.

The choice of a modulation system is one of the main decisions to be made in communication system design. It depends on several interrelated factors of which the more important are:

1. User Specification - the type of service to be implemented - analogue or digital - military or commercial.
2. Channel Specification - the characteristics of the ionosphere between the proposed sites for the transmitter and receiver.
3. Economic Specification - the budget available to meet development, production and running costs of the system.

Once the decision on modulation has been made, the other constituents of the system will generally follow as a matter of course. For example, if the system is to be a military digital communication system over a fairly predictable ionospheric path, then FSK or one of its derivatives will probably be chosen. This implies that the code used will be as efficient as possible regardless of cost, and that a coherent detection (demodulation) technique will be used at the receiver. In contrast, if the system were for commercial use where the budget was limited, then the less efficient FSK technique might be chosen with a less expensive code and non-coherent detection. This system would of course be cheaper than the FSK system and would also be more error-prone, but it would be normally adequate for the less demanding specification of the commercial user.
Besides modulation and coding which vary according to the system, the other signal processing operations such as amplification, filtering, and mixing all exist in some form in every communication system. The only criteria for these processes are that they be linear and as noise-free as possible. In this investigation they will be generally defined as 'black-boxes' with attendant parameters. For example an amplifier will be specified by its gain and noise figure. It will also be assumed that these operations are linear and therefore do not distort the signal.

It may now be asserted that in the context of signal processing, signal quality is dependent only upon the modulation and coding techniques used to enable message transfer in the HF band.

When designing the signal processing hardware for a communication system, the engineer is faced with two general types of constraints - those occurring naturally and those due to the state-of-the-art in relevant technologies.

In HF communication, the natural constraints are those due to the characteristics of the ionosphere. These define the available bandwidth and the noise environment in which the signal must travel. The available bandwidth may be defined as the interval of frequencies over which the magnitude of the signal remains within 3dB of its value at the midband frequency. In the HF band, the midband frequencies are limited naturally to those between 3 and 30kHz (Appendix 1). Since there is also a very high density of traffic in the band, the HF system designer is conventionally limited to a very small bandwidth defined by the world-wide assignment of high frequencies. This stipulates bandwidths that vary up to a maximum of 12kHz for every user of the HF channel. Conventionally assignments below the 3kHz mark are used for telegraphy and data communication, while assignments between 3 and 12kHz are used for analogue communication, or with further frequency division for
digital communication as well. This implies that the conventional HF system designer need only consider narrowband signal processing techniques.

The natural noise environment is also a very important design consideration. Besides the noise occurring in the ionosphere, there is also the noise predicted by kinetic theory to be inherent in all electrical communication systems. Since noise limits the ability to correctly detect the intended message, it is critical to optimum communication design. Furthermore, since the rate of information transfer in a given system depends directly upon correct message identification, narrowband systems with inherently low information rates depend critically upon low noise environments for efficient operation.

A parameter used universally to quantify the effect of noise on communication system performance is the Signal-to-Noise Ratio (SNR), usually expressed in dB thus,

$$\text{SNR}_x = 10 \log_{10} (S_x / N_x)$$ (2.2.1)

where $x$ is the point in the system where the signal power ($S_x$) and noise power ($N_x$) are measured. In most communication systems, the SNR is of major importance at the input to the receiver where $S_x$ is lowest, and where therefore additive noise most seriously degrades the signal. The value of SNR at the receiver output is also very important since it is at this point that the overall system signal quality relevant to this investigation is assessed. This implies that receiver performance is all-important to efficient communication.

Receiver performance depends on state-of-the-art communication technology and is therefore representative of the second type of constraint mentioned previously. The important parameters in receiver performance are predetection gain and noise figure. The former must be automatically controlled to cater for reception of both the smallest
and the largest signals expected, such that the smallest expected signal may be suitably amplified before detection and the largest expected signal may be received without distortion due to amplitude limiting. The latter must normally be kept as small as possible so that the least amount of noise be added to the signal in its passage through the receiver.

The HF communication system designer is thus faced with a number of fixed constraints which must be overcome by proper choice of modulation or coding technique. Each of these techniques has associated with it a certain amount of necessary bandwidth and SNR performance. In designing for optimum signal quality, it is necessary to know the values of these parameters associated with a particular technique and so the intention here is to discuss both analogue and digital modulation techniques with the aim of establishing guidelines for choosing the optimum system performance. First of all, however, the yardstick by which these systems are to be compared must be defined.
2.2.1 Measurement of Signal Quality

This yardstick is the SNR. Equation (2.2.1) gives the SNR in dB at a given point of the system. Here the simple power ratio (i.e. \( S_x/N_x \)) will be used. Conversion to dB can be made by use of the equation.

The discussion here will be in terms of baseband signals represented by the general signal \( E(t) \) which defines an ensemble of probable messages from a given source. Though \( E(t) \) is not normally bandlimited, it will be safe here to assume that there is some upper frequency limit, \( W \) (e.g. 3kHz), above which the spectral content of \( E(t) \) is negligible. \( W \) may be defined as an analogue message bandwidth.

The general communication system, relevant to the discussion, is shown in figure (2.2.1). The transmitter is simply an amplifier with power gain \( G_T \) so that \( S_T \), the transmitted signal power is given by \( S_T = G_T E^2 \). The receiver filter is a nearly ideal low pass filter (LPF) with bandwidth \( W \), so that \( R_N \), the receiver noise bandwidth is given approximately be \( R_N \approx W \). The other parameters are as follows:

- \( L \) = Transmission power loss in channel
- \( S_R \) = Signal power at receiver input
- \( N/T_N \) = Noise temperature referred to receiver input
- \( n \) = Noise density at receiver input (assumed constant)
- \( G_R \) = Receiver front end gain
- \( (SNR)_D \) = Destination signal to noise ratio
- \( k \) = Boltzmann's constant
- \( Y_D(t) \) = Output signal

If the total transmission delay is \( t_d \) and there is no distortion over \(|t| < W \), then the output signal is

\[
Y_D(t) = G_T G_R E(t-t_d) + N_D(t)
\]

(2.2.2)
FIG. (2.2.1) ANALOGUE TRANSMISSION SYSTEM

Receiver Front End

\( LFE \) (Line)

\( R \) (Rm)

\( + \)

\( SE \)

\( GT \)

\( G \) (f)

Transmitter
Where \( N_D(t) \) is the distortion noise signal. Thus the destination signal and noise powers are

\[
S_D = \left\{ G_T \cdot G_R / L \right\} \cdot \frac{E^2}{2} \tag{2.2.3} \quad (a)
\]

\[
N_D = G_R \cdot W \cdot \eta \tag{2.2.3} \quad (b)
\]

\[
(SNR)_D = G_T \cdot \frac{E^2}{2} / L \cdot W \cdot \eta \tag{2.2.4}
\]

Now since \( S_T / L = S_R \), substitution gives

\[
(SNR)_D = \frac{S_R}{W \cdot \eta} \tag{2.2.5}
\]

Table (2.2.1) lists representative values of \((SNR)_D\) for selected analogue signals along with the frequency ranges of interest. These are baseband frequencies but the principles outlined here are applicable to the much wider range of modulated signals.

Equation (2.2.5) represents \((SNR)_D\) in terms of some very basic parameters. These are \( S_R \), \( \eta \) and \( W \).

These three parameters are very important performance variables and as such may be considered standard in communication systems. This they will be here and so the variable \( \gamma \) may be defined thus:

\[
\gamma = \frac{S}{W \cdot \eta} \tag{2.2.6}
\]

In the present context, \( \gamma \) is equal to the \((SNR)_D\) for baseband signals. More generally, since equation (2.2.5) assumes distortionless transmission conditions and a nearly ideal filter; it is more accurate to say that \((SNR)_D < \gamma \). \( \gamma \) therefore represents an upper band for analogue baseband performance that may or may not be achieved in an actual system. For example, the noise bandwidth of a typical LPF will be greater than the message bandwidth which gives \((SNR)_D = \frac{S}{W \cdot \eta} \), which is less than \( \gamma \). Similarly non-linearities that cause the output to include signal-times-
<table>
<thead>
<tr>
<th>Signal Type</th>
<th>Frequency Range (Hz)</th>
<th>$(SNR)_d$ dBs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barely intelligible voice</td>
<td>500-2000</td>
<td>5-10</td>
</tr>
<tr>
<td>Telephone quality voice</td>
<td>200-3200</td>
<td>25-35</td>
</tr>
<tr>
<td>AM broadcast quality audio</td>
<td>100-5000</td>
<td>40-50</td>
</tr>
<tr>
<td>HiFi Audio</td>
<td>20-20 000</td>
<td>55-65</td>
</tr>
<tr>
<td>Television Video</td>
<td>60-4.2 MHz</td>
<td>45-55</td>
</tr>
</tbody>
</table>

Table (2.2.1) Destination SNRs for typical baseband signals.
noise terms also reduce the effective SNR.

When the noise is non-white and/or channels require considerable equalization, the rather simple approach above must be replaced by a more sophisticated approach in which specially designed filters are incorporated at both the transmitter and receiver. This is the principle of optimal terminal filtering in analogue signal transmission. Figure (2.2.2) depicts an updated version of figure (2.2.1) which typifies the system more normally encountered in practice. The power gains $G_T$, $1/L$ and $G_R$ are absorbed in the frequency-response functions $H_T(f)$, $H_C(f)$ and $H_R(f)$ where $T$, $C$ and $R$ denote transmitter, channel and receiver respectively.

As far as distortionless transmission is concerned, any pair of terminal filters will suffice as long as

$$H_T(f).H_R(f) = K.e^{-\frac{i\omega f}{C}}/H_C(f)$$

(2.2.7)

where $K$ is an arbitrary constant and $|f| < W$. Hence if $H_R(f)$ is chosen to minimize the input noise (maximize the output SNR) and $H_T(f).H_R(f)$ satisfies equation (2.2.7), the terminal filters will have been optimized in the sense that $(SNR)_D$ is maximum and the output signal is undistorted. The optimization is however subtly constrained by the need for keeping $S_T$ within reasonable limits. Hence $S_T.N_D/S_D$ rather than $N_D$ alone is what is being optimized.

If this is done and equation (2.2.7) holds, expressions can be obtained for the optimum receiving and transmitting filters, as

$$|H_R(f)|^2_{opt} = \frac{G^\frac{1}{2}}{H_C(f)} |G_N(f)|$$

(2.2.8)
Fig (2.2.2) Generalised Perceptual Communication System
and

\[ |H_T(f)|^2_{\text{opt}} = k^2 \cdot G_N^2(f) / |H_c(f)| \cdot G_E^2(f) \]  

(2.2.9)

respectively, where \( G_B(f) \) is the message spectral density and \( k^2 \) can be determined from the desired value of \( S_n \).

Interpretation of these results shows that \( H_R(f) \) de-emphasizes those frequencies where the noise density is large and the signal density is small, while \( H_T(f) \) does the reverse. The phase shift of \( H_T(f) \) and \( H_R(f) \) does not appear here since spectral densities are being considered. Nevertheless, the overall phase must satisfy equation (2.2.7) for distortionless transmission.

Apart from the difficulties of synthesizing the filters mathematically, the major obstacle in practice is the assumption of the fact that \( G_B(f) \) is known in detail. Usually the communications engineer knows the general characteristics of the message or rather the class or ensemble of probable messages - but not the complete details. For example, if the messages are known to be bandlimited in \( W \), it may be assumed that the spectral density is flat over \(|f| < W\); there being no reason to believe that \( G_B(f) \) is larger or smaller at any particular frequency. Using this assumption will yield a good but non-optimum filter.

The above example is actually the underlying assumption in the approach at the beginning of this section when the noise was assumed to be white and the receiving filter served only to eliminate the out of band noise. If \( G_N(f) = n/2 \) and the channel is distortionless so that \(|H_c(f)|^2 = 1/l\) then the terminal filters become ideal BPFs.

Now for optimum filtering.
This may be written

\[
(SNR)_{D', \text{max}} = \frac{S_T}{(S_T - N_D - D)_{\text{min}}}
\]

(2.2.10)

In the bandlimited case, the denominator of this equation is integrated between the limits ± \( \omega \). This gives

\[
(S_T - N_D - D)_{\text{min}} = L\omega^2
\]

(2.2.12)

and so

\[
(SNR)_{D', \text{max}} = \frac{S_T}{L\omega^2 \eta}
\]

(2.2.13)

by substitution of (2.2.12) in (2.2.11).

Similar to the optimum terminal filters discussed above for analogue signals, there are optimum filters for digital signals. These filters process pulses of known shape and can thus be truly optimum since they can be exactly synthesized for a given signal. They are used extensively in data transmission systems using FSK, PSK and their derivatives.

In the case of digital transmission, the concept of SNR by itself is not normally used. Rather there is the idea of probability of error which defines the uncertainty that the symbol received is correct. This probability is however usually directly dependent upon the SNR performance of the keying or coding technique; due to the fact that the recognition of a particular pulse depends upon the amount of noise contaminating it when it arrives at the receiver filter.

The SNR performance of a keying technique may be defined in a way analogous to that used for the analogue baseband signal. However,
instead of bandwidth, the sampling rate, $r$, is used. Hence the parameter, $\rho$, can be defined as

$$\rho = \frac{S_p}{\pi r}$$  \hspace{1cm} (2.2.14)

The sampling rate, $r$, is given by $B/2$ where $B$ is the channel bandwidth. $\rho$ is thus defined for the ideal baseband digital system and will be used to evaluate signal quality for such systems in the same way that $Y$ will be used for analogue systems.

Taking the receiver section of figure (2.2.2) and assuming that the desire is to maximise the output pulse amplitude at some arbitrary time, and to minimize the output noise regardless of the shape of the pulse, which is unimportant in digital detection. In the absence of noise, the peak output signal at $t = t_0$ is

$$Y_D(t_0) = \int_{-\infty}^{\infty} H_R(f)X(f)\, e^{j\omega_0 f} \, df$$  \hspace{1cm} (2.2.15)

where $X(f) = F(E(t))$ - the Fourier transform of $E(t)$. The transform is used here rather than the spectral density function $G_E(f)$, since the pulse is a known energy signal. The output noise power is

$$N = \int_{-\infty}^{\infty} |H_R(f)|^2 G_N(f) \, df$$  \hspace{1cm} (2.2.16)

The matched filter frequency response $H_R(f)$ emphasizes the frequencies where $|X(f)| / G_N(f)$ is large and vice versa. Unfortunately in this case $H_R(f)$ is physically unrealizable because the corresponding impulse response is non-zero for $t < 0$. It may however be very closely approached in practice if the realizability constraint is taken into account. One important quality of the matched filter is that the impulse response has the same shape as the input pulse reversed in time and shifted by $t_0$. This is equivalent to synchronous detection.
which, as will become apparent, is a very desirable quality for any reception system.

A method for enumerating signal quality in both analogue and digital HF signal processing has now been established. However before it can be used to evaluate the performance of the various modulation and coding techniques, something must be said about the important bearing the noise figure of the receiver front-end has on this parameter.

In the discussion above, the receiver front-end was depicted as an amplifier with gain $G_R$ which was assumed noiseless. In practice this is never the case as was stated previously.

Figure (2.2.3) shows a typical configuration for a receiver front-end where it can be seen that the problem of receiver noise may be considered in three sections;

1. The source signal-to-noise ratio $(SNR)_S$ at the antenna terminals.
2. The predetection signal-to-noise ratio $(SNR)_R$
3. The destination signal-to-noise ratio $(SNR)_D$

The input noise is generally produced by random electromagnetic radiation intercepted by the antenna - this radiation being due to both natural and man-made phenomena. The former includes atmospheric noise (static) and emission from extra-terrestrial bodies (cosmic noise). The latter consists of such effects as commutator sparking, fluorescent lighting, x-ray machines and other processes involving electrical discharge. In addition to this, the antenna itself may be a noise generator, implying that the incoming noise power depends on the kind of antenna used, its position and orientation, the operating frequency and the characteristics of the environment in which it
Fig (2.2.3) Typical Receiver Configuration
operates. \((\text{SNR})_S\) is thus a difficult parameter to predict theoretically and is usually measured on site when needed.

At the other end of the predetection unit, the detector and any subsequent stages are normally relatively noiseless, so that \((\text{SNR})_R\) and \((\text{SNR})_D\) are related to each other only by the type of modulation used.

If the predetection unit is also noiseless, then \((\text{SNR})_S\) is equal to \((\text{SNR})_R\) and there is no problem. However this is seldom if ever the case since the RF and IF stages often add internally generated noise whose magnitude is comparable to, or even greater than, the input noise power.

Receiver noise is usually discussed in terms of two standard parameters—the Noise Figure (F) and the Noise Temperature (NT) which relate \((\text{SNR})_S\) to \((\text{SNR})_R\) for the predetection unit. They are thus invaluable as an aid to optimum system design.

The predetection unit of a receiver consists of several units connected in cascade, each of which produces its particular internally generated noise. Figure (2.2.4) shows one of these components as a noisy two port network with gain \(G_a\), noise equivalent bandwidth \(B_N\) and additive noise power \(N_{ax}\). The unit is also matched to its signal source and load. The source produces a signal power, \(S_{ai}\) and a noise power \(N_{ai}\). If this noise is white and represented by a noise temperature \(NT_i\), then \(N_{ai}\) is given by \((kNT_iB_N)\).

Under these conditions it is easy to show that

\[
(\text{SNR})_o = (\text{SNR})_i \left/ \left(1 + N_{ax} / (G_a kNT_i B_N)\right)\right.
\]

\[(2.2.18)\]

\((\text{SNR})_o\) is thus never greater than \((\text{SNR})_i\). Equation (2.2.18) also shows that despite additive noise the output and input SNRs can be
Fig (2.2.4) SINGLE UNIT OF PREDETECTION SECTION
nearly equal as long as \( N_{ax} < \langle G_a \cdot k \cdot N_{TF} \cdot B_n \rangle \). This implies that the absolute value of the excess noise power is not important as long as it is small relative to the source noise. It may be further noted that \( N_{ax}/G_a \cdot k \cdot B_n \) depends only upon the device parameters and has the dimension of temperature. Thus the effective input \( NT_i \), or the amplifier temperature, \( NT_E \), may be given by this expression. \( NT_E \) is a measure of noisiness referred to the input of the unit. If a device is noiseless then \( NT_E = 0 \). If \( NT_E \) is substituted in equation (2.2.18), then the output noise power can be derived as \( (G_a \cdot k \cdot (NT_i + NT_E) \cdot B_n) \) which gives an output noise temperature as \( G_a (NT_i + NT_E) \).

If the impedences are not matched, then all the powers will be less than those available. The above equations will still hold as \( G_a \) and \( NT_i \) are the predominant factors (i.e. \( NT_E < < NT_i \)). It may also be noted that as \( G_a \) is defined for matched conditions, some compensation will in general be necessary.

\( NT_E \) is a useful receiver performance parameter when \( NT_E < < NT_i \). If this is not the case, then it is generally preferable to use the Noise Figure. \( F \) is formally defined as the actual output noise power divided by the value of the output noise power if the two port network were noiseless, the source being at room temperature. \( F \) is usually greater than unity and \( F = 1 \) gives the noise figure for a noiseless amplifier.

Using the same notation as before, the output SNR may be defined by

\[
(SNR)_0 = (SNR)_i/(1 + (F-1) \cdot (NT_0/NT_i))
\]

(2.2.19)

If the source is at room temperature, then \( NT_i = NT_0 \) and \( (SNR)_0 = (SNR)_i/F \). In most modern systems, however, where noise is a critical parameter,
equation (2.2.19) must be used in almost every case.

The predetection unit consists of several of the units shown in figure (2.2.4) in cascade. This means that the individual noise characteristics must be combined in some way to give the total noise performance of the predetection stage. This process will relate $(\text{SNR})_S$ to $(\text{SNR})_R$ and thus provide the means of evaluating $(\text{SNR})_D$ - the parameter of most interest in this investigation.

With three or more cascaded stages, $NT_E$ and $F$ of the system so composed are given by

$$NT_E = NT_{E1} + NT_{E2}/G_{a1} + NT_{E3}/G_{a1}G_{a2} + \ldots$$

(2.2.20) (a)

$$F = F_{1} + \frac{F_{2} - 1}{G_{a1}} + \frac{F_{3} - 1}{G_{a1}G_{a2}} + \ldots$$

(2.2.20) (b)

Equation (2.2.20) (b) is the well-known FR11S formula and shows quite clearly that receiver noise is dominated by the first stage - the familiar front-end of the predetection unit. Thus if $G_{a2}>1$, then the overall $NT_E$ is essentially that of the first stage itself. If however the first stage is an attenuator of loss $L = 1/G_{a1}$ then the effect is compounded as $NT_{E1} = (L - 1)NT_{i}$, etc. This implies that front-end attenuators should be avoided wherever possible.

Assuming now that the predetection unit has a noise bandwidth of approximately the modulated signal bandwidth, $B_T$, and that at the antenna terminals, the signal power $S_S = S_T/L$ and the antenna noise temperature is $NT_S$. Then if the antenna is matched to the first stage, as is usual, then the predetection SNR will be given by

$$(\text{SNR})_R = G_a S_S / (G_a k (NT_S + NT_E) B_T)$$

(2.2.21)
Substitution and simplification then give

\[(\text{SNR})_R = \frac{S_T}{L \cdot k \cdot NT_N \cdot B_T} \]  \hspace{1cm} (2.2.22)

where \(NT_N = (NT_S + NT_E)\) is the system noise temperature referred to the antenna terminals. If \((\text{SNR})_R\) is taken as equal to \(s_R/\eta B_T\), where \(\eta\) is the uniform white noise density over the passband \(B_T\), then \(s_R = G_R \cdot s_T/L\) and \(\eta\) is given by \(G_R \cdot k \cdot NT_N\). Here \(G_R\) is the actual power gain of the predetection unit as before.
2.2.2 Analogue Modulation Techniques

The conventional AM signal used in standard radio broadcasting may be defined as,

\[ E_A(t) = A(1 + I.m(t)) \cos \omega_c t \]  
(2.2.23)

where \( \omega_c \) is the constant carrier frequency, \( I \) is the modulation index and the time varying amplitude, \( (1 + I.m(t)) \) is a linear function of \( m(t) \) - the modulating signal containing the information. It will be assumed here that \( m(t) \) is so scaled that \( |m(t)| < 1 \).

In Double Side-Band (DSB) AM, the carrier term, \( A \cos \omega_c t \) is suppressed and equation (2.2.23) becomes for DSB

\[ E_D(t) = A_D.m(t) \cos \omega_c t \]  
(2.2.24)

here subscript \( D \) denotes DSB.

If the DSB signal is filtered to remove one of the sidebands, Single Sideband (SSB) AM results as

\[ E_{SU}(t) = \frac{1}{2}A_S(m(t) \cdot \cos \omega_c t - \hat{m}(t) \sin \omega_c t) \]  
(2.2.25)

Here \( S \) represents SSB and \( U \) the upper sideband. The minus sign before the second term in brackets becomes a plus if the lower rather than the upper sideband is used. Finally \( \hat{m}(t) \) is the Hilbert transform of \( m(t) \).

Two variations of SSB, Independent Sideband (ISB) and Vestigial Sideband (VSB) are also used in particular applications but their characteristics will not be discussed here. The interested reader is referred to appropriate references.\(^1\),\(^3\),\(^4\)

The noise that affects HF analogue signals is narrowband as is the modulation used at HF. This type of noise may be represented in two ways which are of special use in evaluating the SNR performance of
The first of these is the quadrature carrier representation given by

\[ n(t) = n_1(t) \cos \omega_c t - n_2(t) \sin \omega_c t \]  

(2.2.26)

where \( n(t) \) is a sample function of an ergodic Gaussian random process in the interval 0 < t < T. The coefficients \( n_1(t) \), \( n_2(t) \) are themselves sample functions which are independent Gaussian variates with zero mean and equal variances in the limit \( T \to \infty \).

If the original random process has bandwidth \( B_q \) centred at \( \omega_c \), then \( n_1(t) \) and \( n_2(t) \) each have bandwidth \( B_q \) centred at the zero frequency, and may be regarded as low-pass, random noise waveforms having zero mean and positive frequency bandwidth \( B_q/2 \). They respectively DSB modulate the quadrature carriers \( \cos \omega_c t \) and \( -\sin \omega_c t \). This formulation is particularly useful for calculating the effects of small amounts of noise on system performance.

The second representation is the envelope-and-phase definition given by

\[ n(t) = R_n(t) \cos(\omega_c t + \phi_n(t)) \]  

(2.2.27)

where \( R_n(t) \) is the envelope function and \( \phi_n(t) \) the phase function of the noise signal. A physical realization of this representation is possible if noise is observed after it has passed through a narrow band-pass filter. It appears as a sinusoidal waveform, the amplitude and phase of which fluctuate slowly compared to the average period of the wave. \( R_n(t) \) and \( \phi_n(t) \) are similar to the coefficients \( n_1(t) \) and \( n_2(t) \) and in this form the expression is useful for SNR analysis when the signal is weak compared to the noise.

These two definitions of narrowband Gaussian noise are closely...
related in the context of the present discussion. This is due to the fact that modulated signals are generally bandpass signals rather than baseband signals which were used in the previous section. Besides having a modulation capability at the transmitter, systems using bandpass signals also have an extra filter at the receiver front-end. This filter ensures that only the small band containing the carrier frequency and the message sidebands are passed to the predetection unit of the receiver. Figure (2.2.5) shows a block diagram of a typical bandpass signal communication system.

From the figure, if \( G(f) = \frac{\eta}{2} \) at the input, then the spectral density of \( n(t) \) is \( G_N(f) = \frac{\eta}{2} \left| H_R(f) \right|^2 \). Assuming that \( H_N(f_c) = 1 \) and that the receiver bandwidth \( B_R \) is approximately equal to the transmission bandwidth \( B_T \), then if also the noise equivalent bandwidth of \( H_R(f) \) is approximately \( B_T \), the total filtered noise power is \( N_R = \frac{n^2}{2} = \eta B_T \), which is known as the received noise power.

Having now defined the signal and noise functions used in analogue HF communication, it is now possible to evaluate the SNR performance of HF analogue modulation systems.

Figure (2.2.6) depicts a typical CW system receiver. The modulated signal plus noise at the detection is given by

\[
u(t) = K_R E(t) + n(t) \quad (2.2.28)
\]

Since linear modulation has \( B_T = W \) or \( 2W \) depending on whether or not a sideband has been suppressed, there are three possible spectra for \( G_N(f) \) if \( H_N(f) \) is symmetrical. These are shown in figure (2.2.7). The average signal power at the detector input is \( K_R^2 \) multiplied by the time averaged value of \( E^2(t) \); the result is \( S_R \). The noise power, \( N_R \), is given by \( B_T \). Since signal and noise are additive in equation
Fig (2.2.5) CW COMMUNICATION SYSTEM

[Diagram of a communication system with blocks labeled LPF, Detector, BF, Channel, and Transmitter, with annotations like Ytt, V(t), and E(t).]
\[ G(f) = \frac{n}{2} \]

\[ V(f) = kR \cdot E(t) + N(t) \]

\[ kR = V \sqrt{L} \]

**Fig. (2.2.6) CW Receiver System**
FIG. 2.2.1) PREDETECTION NOISE SPECTRUM $G_{IN}(f)$ IN LINEAR MODULATION.
(2.2.28), it is meaningful to define the predetection SNR as

$$\text{(SNR)}_R = \frac{S_r}{N_R} = \frac{S_r}{N_B_T} \quad (2.2.29)$$

Taking the expression for $\gamma$ from equation (2.2.6), as $\gamma = \frac{S_r}{N_W}$, it can be deduced that

$$\text{(SNR)}_R = \frac{(W/B_T)}{\gamma} \quad (2.2.30)$$

Hence $\text{(SNR)}_R = \gamma$ for SSB ($B_T = W$) and $\gamma / 2$ for DSB ($B_T = 2W$).

Now $\gamma$ is equal to the maximum value of $\text{(SNR)}_D$ for analogue baseband transmission. By the same reasoning, equations (2.2.29) and (2.2.30) represent upper bounds for the bandpass signal.

Now $S_R$ is also defined for a modulated signal in terms of $S_T$, the transmitted signal power as

$$S_R = S_T/L = K_{R'}^2 S_T$$

$$= (A_R/A_c)^2 S_T \quad (2.2.32)$$

where $A_R = K_{R'}A_c$ is the carrier amplitude at the detector input, and $A_c$ is the value of this parameter before bandpass filtering.

With the above considerations, two questions came to mind.

1. Given $E(t)$ and the type of detector, what is the final output signal-plus-noise waveform, and

2. If the signal and noise are additive at the output, what is the value of $\text{(SNR)}_D$ in terms of $\text{(SNR)}_R$?

In other words; what is the measure of signal quality in HF analogue systems? The answer lies in an investigation of the two detection methods used in analogue communications.

The first of these is synchronous detection using a product demodulator. The incoming signal is multiplied with a locally
generated sinusoid which must be synchronized with the carrier in frequency and phase. The result is low pass filtered with a filter having bandwidth equal to or larger than the message bandwidth.

If the modulation is DSB, then equation (2.2.24) is substituted for \( E(t) \) in equation (2.2.28). If further \( n(t) \) is represented in quadrature carrier form, equation (2.2.26), then the total detector input is

\[
 u(t) = (A_R m(t) + n_i(t)).\cos \omega_c t - n_q \cdot \sin \omega_c t 
\]

(2.2.33)

This form of the bandpass signal, \( u(t) \), is itself the quadrature carrier representation of \( u(t) \). When this is applied to an ideal synchronous detector, the result \( y(t) \) in figure (2.2.6), is \((A_R m(t) + n_i(t))\) and after low pass filtering

\[
 y_D(t) = A_R m(t) + n_i(t) \quad (2.2.34)
\]

since both components of \( y(t) \) are band limited to \( B_T/2 \).

Three important conclusions may be drawn from equation (2.2.34).

1. The message and noise are additive at the output.
2. The quadrature noise component, \( n_q(t) \) is completely rejected.
3. The output noise power spectrum is \( G_{n(t)}(f) \) which has the shape of \( G_N(f) \) translated to zero frequency.

If \( H_R(f) \) is relatively flat over \( B_T \), the output noise is essentially white over the message bandwidth \( W \).

Taking the mean square values of the two terms in equation (2.2.34), \((SNR)_D \) is given by

\[
 (SNR)_D = A_R^2 m^2 / n_i^2
\]

\[
 = 2S_R / n_B T
\]

\[
 = 2(SNR)_R \quad (2.2.35)
\]
using equation (2.2.30) that gives for DSB

\[
(SNR)_D = \frac{E_{\text{SNR}}}{N_0} = \gamma \quad (2.2.36)
\]

Therefore in narrowband gaussian noise, DSB with ideal synchronous detection is equivalent to baseband transmission even though the transmission bandwidth is twice as great. This performance is obtained because the translated sidebands containing the signals overlap coherently, whereas the noise sidebands sum incoherently. This signal sideband coherence counterbalances the doubled noise power \(nB = 2nT\) compared to baseband.

The performance of conventional AM with synchronous detection may be easily inferred from the reasoning above and equation (2.2.31).

This process gives for AM

\[
(SNR)_D = \frac{I^{2-2}}{(1 + I^{2-2})} \gamma \quad (2.2.37)
\]

where \(I\) and \(m(t)\) are previously defined. Equation (2.2.37) shows that \((SNR)_D\) for AM is less than or equal to \(\gamma/2\) (since \(I \leq 1\)) which gives the well-known result that fifty per cent or more of the transmitted power is wasted in the AM carrier. Other factors being equal therefore, an AM system must transmit at least twice the power of a suppressed carrier system to achieve the same performance at the destination.

This means that on an average power basis, conventional AM is inferior to DSB by 3dB or more. Typically \(I^{2-2} = 0.1\), for which \((SNR)_D\) is about 7dB below the maximum value for \(I^{2-2} = 1\). Under these conditions, AM is typically 10dB worse than DSB.

Under peak power limiting conditions, this is a serious problem in AM radio broadcasting and special techniques such as volume.
compression and peak limiting are frequently employed at the
transmitter to ensure that the carrier is fully modulated most of
the time. These techniques actually distort the recovered message
and would be unacceptable for analogue data transmission. For
audio programme material the distortion is usually tolerable.

For SSB, the DSB analysis must be modified by a quadrature
component in \( E_S(t) \) and the offset carrier frequency compared to the
centre frequency \( f_0 \) of \( H_R(f) \). These changes to the analysis yield
for SSB

\[
(SNR)_D = A_R^2 m^2 / 4, \eta . B_T = \gamma
\]

(2.2.38)
since \( B_T = W \) and \( S_R = (A_R^2 m^2 / 4) \) for SSB.

This performance is similar to that for DSB except that it
is produced in half the bandwidth which means that SSB is generally
3dB better than DSB.

So far the modulation techniques have been considered for
equal values of \( S_R \), the average received power. It is generally
more realistic to consider equal peak powers, which reflect the peak
power constraint of the transmitters.

Under these conditions, it can be shown that \( (SNR)_D \) for DSB
is four times (6dB) and that for AM since the peak powers are
proportional to \( A_c^2 \) and \( 4A_c^2 \) respectively.

Peak power calculations for SSB with representative modulating
signals indicate that SSB is two to three dB better than DSB. If
however the message has pronounced discontinuities which cause
envelope 'horns' then SSB becomes inferior to DSB.

The suppressed-carrier systems need the complex and expensive
synchronous detection techniques for proper operation since they have no carrier for reference at the receiver. AM, on the other hand, has power in the carrier to spare and so is ideally suited to envelope detection, the second of the techniques used to demodulate analogue signals.

The main difference between the performance of the envelope detector and the synchronous detector, in the presence of noise, is the occurrence in the former technique of a threshold effect. In synchronous detection the signal is always present no matter the level of the noise; and though it might actually be lost in the noise, an unmutilated version of the transmitted signal always exists. In envelope detection, mutilation of the signal occurs at low predetection SNRs. This is the threshold effect. It implies that there is some value of \( (\text{SNR})_{R} \) above which mutilation is negligible and below which system performance rapidly deteriorates.

In fact, if \( (\text{SNR})_{D} \gg 1 \), then envelope detection is equal in performance to synchronous detection, but if \( (\text{SNR})_{D} \ll 1 \) the message is usually lost beyond retrieval. Note that \( (\text{SNR})_{D} \) for envelope detection is the previously derived expression for synchronously detected AM in equation (2.2.37). From this equation \( (\text{SNR})_{R} = \gamma / 2 \) and if \( (\text{SNR})_{R} \), \( (\text{th} = \text{threshold}) \), is given approximately by 10, \( (\text{actual S/N}) \) which gives a 99% probability that \( (\text{SNR})_{D} \gg 1 \), then \( \gamma_{th} \) is equal to 20. The threshold effect is therefore seldom serious in practice if enough power is available and since, for example, reasonably intelligible voice transmission demands a post detection SNR of 30dB or more, it is reasonable to expect that such systems will always be designed to provide enough power for this stipulation.
The results derived in this section for the SNR performance of the analogue modulation techniques used at HF are summarised for convenience in table (2.2.2). Results for baseband signals are also included for reference.

Of the several types of modulation, the suppressed carrier methods are superior to conventional AM on several counts. SNRs are better and there is no threshold effect which wastes transmitter power. When bandwidth conservation is important, SSB and VSB are particularly attractive. But one seldom gets something for nothing and the price of efficient linear modulation is the increased complexity and hence cost of instrumentation particularly at the receiver. Synchronous detection, no matter how it is implemented, requires highly sophisticated circuitry compared to that needed for envelope detection. For point-to-point communications (one transmitter, one receiver) the price may be worthwhile. But for broadcast systems (one transmitter, many receivers), economic considerations must tip the balance toward the simplest possible receiver and hence envelope detection.

From considerations of instrumentation AM is the least complex and costly of the systems, while suppressed-carrier VSB, (which was not discussed but it is included for completeness) with its special sideband filter and synchronization, is the most complex. Of DSB and SSB (in their proper applications) the latter is less difficult to implement because the synchronization is not so critical. In addition improved filter technology has made the required sideband filters more readily available.
<table>
<thead>
<tr>
<th>Type</th>
<th>$\beta$</th>
<th>$B_I/W$</th>
<th>$(SNR)_D/\gamma$</th>
<th>$\gamma^\prime$</th>
<th>DC</th>
<th>Complexity</th>
<th>Comments</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseband</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>No</td>
<td>Minor</td>
<td>No modulation</td>
<td>short-hand links</td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>2</td>
<td>$\frac{1^2 - m^2}{1 + l^2 - m^2}$</td>
<td>20</td>
<td>No</td>
<td>Minor</td>
<td>envelope detection $\leq 1$</td>
<td>broadcast radio</td>
<td></td>
</tr>
<tr>
<td>DSB</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td>Yes</td>
<td>Major</td>
<td>synchronous detection</td>
<td>Analogue data multiplexed systems</td>
<td></td>
</tr>
<tr>
<td>SSB</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>No</td>
<td>Moderate</td>
<td>synchronous detection</td>
<td>point to point voice multiplexing</td>
<td></td>
</tr>
<tr>
<td>VSB</td>
<td>$1+$</td>
<td>1</td>
<td>-</td>
<td>Yes</td>
<td>Major</td>
<td>synchronous detection</td>
<td>Digital data</td>
<td></td>
</tr>
</tbody>
</table>

$\beta$ = normalized transmission bandwidth

$W = \text{message bandwidth}$

$S_R = \text{received signal power}$

$\eta = \text{noise density referred to receiver input}$

N.B. Ideal systems are assumed so values of $(SNR)_D$ are upper bounds.

Table (2.2.2) Comparison of HF CW Modulation Systems
2.2.3 HF Digital Modulation Techniques

In the HF channel where narrowband (NB) systems are conventionally used, digital signalling is effected through the use of NBFM and NBPM. These two modulation techniques are produced when the modulation indices of wideband FM and PM, respectively, are constrained to be less than or equal to unity

In digital systems, the message is usually coded in binary which requires only two states of the carrier to produce completely coded words. Using NBFM, this would mean the use of two separate frequencies to represent the binary '1' or MARK, and the binary '0' or SPACE. Similarly with NBPM, two different phases would be used. Reception would then consist of being able to recognize a MARK rather than a SPACE and vice versa. In other words, the receiver must be able to distinguish between two distinct frequencies or phases. The digital signalling system based on NBFM is called Frequency-Shift-Keying (FSK), while that based on NBPM is known as Phase-Shift-Keying (PSK).

The ability of a given digital system to do a good job is expressed in terms of the probability of error in detection, (ie) the probability that a MARK will be interpreted as a SPACE and vice versa. The lower this probability, the better the signal quality performance of the system. This probability is usually directly dependent upon the parameter \( \rho \) given by

\[
\rho = \frac{S_R}{\nu r} \tag{2.2.39}
\]

where \( \rho \) is the parameter for digital systems corresponding to \( \gamma \) (equation (2.2.6)) for analogue systems. In equation (2.2.39), \( S_R \) and \( \nu \) are as defined previously; \( r \) is the sampling rate of the keying technique given by \( r = B_n/2 \) where \( B_n \) is the channel bandwidth. \( \rho \) is thus defined here for the ideal baseband digital system.
The simplest form of FSK radio system is one with rectangular NBFM and constant carrier amplitude. This system is delineated by ideal transmitted pulses of the form:

\[ E_p(t) = A_k \sin \omega_1 t \text{ for MARK} \]
\[ = A_k \sin \omega_2 t \text{ for SPACE} \]

(2.2.40)

where \( \omega_1 = 2nf_1 \) and \( \omega_2 = 2nf_2 \), and \( f_1, f_2 \) are assumed constant over a single information pulse.

In non-coherent FSK reception using a pair of tone filters, the output of each filter is envelope detected. The envelopes are then sampled once per information pulse, and the MARK/SPACE decision is made according to whichever is the larger of the samples. Assuming that the signals and filtering are ideal, and that there is no crosstalk etc, and also assuming that the signal is contaminated with bandpass noise of the type given in equations (2.2.26) and (2.2.27), then an error occurs when the filter containing noise alone has a larger envelope, \( R_2 \), than the filter with the signal plus noise, \( R_1 \), say. The probability of error is then given by:

\[ P_{eF} = \text{Prob}(R_2 > R_1) = \frac{1}{2} e^{-\beta/2} \]  

(2.2.41)

Coherent detection of FSK is seldom if ever used largely because if the money is available for a complex coherent detection system, then it is very much better to use a PSK modulation system. The discussion on PSK is best approached from the viewpoint of decision theory. This viewpoint points out that optimum (minimum error-rate) binary signalling/reception requires the encoding of each information element as the algebraic sign of a single pulsed waveform with detection by multiplication and integration (cross-correlation) against a perfect replica of this waveform. PSK of a constant amplitude carrier, with sharp
binary transitions between two phase states separated by \( \pi \) radians fulfills exactly the requirements for optimum signalling. The resulting PSK signal has the form of a sequence of plus/minus rectangular pulses of a continuously generated sinusoidal carrier.

Detection of the binary information in a PSK signal is implemented in either of two ways. The more widely used method is the coherent detector which receives a reference waveform accurate in frequency and phase. An exactly similar reference is required to realize the phase detector. The ideal phase detector is a zero memory device whose output is independent of the envelopes of the detector inputs and which specifically measures the cosine of the phase differences. A phase difference of zero results in a +1 (MARK) and of 180° in a -1 (SPACE), which is directly the kind of bipolar output required by decision theory in optimum binary communication systems. The coherent detector is also a zero-memory device, but it differs from the phase detector in that it has an output proportional to both the input envelopes and the phase difference cosine. Therefore both detection methods result in identical binary decisions from the same input signals and thus there will be no performance difference in binary PSK between the coherent detection as suggested by decision theory, and the phase detector implied by direct analogy to the demodulation of wideband PM signals.

An error occurs in ideal coherent PSK when the phase is negative when it should be vice versa. Assuming contamination of the pulse with narrowband gaussian noise, the probability of error for this keying method is given by

\[
P(e) = \frac{1}{2} \text{erfc} \ \rho \quad (2.2.42)
\]

when \( \rho \) is the (SNR) in the filter output at the sampling instant. When the \( Pe \) for PSK is compared to that of FSK (coherent) expressed as \( \frac{1}{2} \text{erfc} \ \rho/2 \), it is clear that a particular probability of error occurs at exactly 3dB lower SNR in ideal coherent PSK than in coherent FSK.
Thus where a system is needed to operate below some specified error-rate, there is a very real 3dB design advantage in using PSK rather than FSK. As is also evident, coherent PSK has no greater equipment requirements than does coherent FSK; in fact, less since only one frequency is used. This fact results in the very rare use of FSK in a coherent communication system.

Use of PSK does present two very real problems in practice, however; the first of which is the possibility of phase errors in the receiver reference with respect to the incoming signal. These errors are due to drifts of a relative nature in the transmitter and receiver master oscillators. These changes may be compensated for by correcting the receiver reference, but the information needed for this can usually only be obtained by operations on the received signal. In this situation relatively long-term smoothing is required to diminish additional apparent fluctuation caused by receiver noise.

The other important limitation in PSK detection is the additive noise associated with deriving the reference signal in the receiver. This may be briefly discussed in relation to an important derivative of ideal coherent PSK - differential PSK or DPSK. DPSK overcomes the long-term stability and high quality correction loops required for ideal coherent PSK. In DPSK systems it is assumed that there is enough stability present in the oscillators and the transmission medium to make negligible any change in phase from one information pulse to the next, aside from change caused by actual encoding. Information is then encoded differentially by encoding the information in terms of the phase change between successive pulses. For example, no (0°) phase shift from the previous pulse could be MARK and then a 180° shift would be SPACE. A coherent or phase detector is still used, one input being the 'current' phase and the other being the previous pulse appropriately delayed.
The major difference between PSK and DPSK is not however in the
differential encoding. Rather it is in the fact that the reference
signal in DPSK is derived from the receiver input over a single
previous pulse. This means that the reference is contaminated by
additive noise to the same extent as the information pulse; (ie) both
have the same SNR.
Therefore assuming that a MARK is transmitted and that in the absence
of noise, both the signal and reference portions of the filter output
have the same frequency and phase, it can be shown\(^5\) that the error-rate
for DPSK is given by
\[ P_{ed} = \frac{1}{2} e^{-P} \] (2.2.43)
Comparison of this with equation (2.2.41) shows that at all error
rates, DPSK requires exactly 3dB less SNR than non-coherent FSK, for
the same error-rate. It is also clear that at high SNRs DPSK performs
almost as well as ideal coherent PSK at the same keying-rate and power
level.

Therefore even though ideal coherent PSK is undeniably the best
system for best error-rate performance, DPSK is often recommended
because it avoids the problem of Transmitter/Receiver master oscillator
synchronization. However both PSK and DPSK require coherent detection
which produces considerable complexity and cost in the receiver. In
many cases therefore, it is found that non-coherent FSK is used, even
though it is prone to a greater error-rate than either PSK or DPSK.

From a consideration of the error-rates given above, it may be
deduced that the error probabilities of HF digital signalling techniques
are all directly proportional to \(P\). If for any of the systems described
above, the signal power is limited to some maximum value - a relevant
optimum design consideration - and errors are still unacceptably
frequent, then some other means of improving reliability must be found.
Often error control coding provides the best solution.

At this point it is not intended to discuss this vast and multifaceted subject. This will be done in Chapter (3). Suffice here to say that error control coding is the calculated use of redundancy to enable a detector to carry out a dynamic check on the bits of the code it is receiving. At the present time, through the use of interleaved block cyclic codes, convolutional codes, and many other types of code, error control coding is proving itself of major importance in improving HF digital communication.
2.3 Fundamental Ionospheric Effects

So far in this investigation, the HF channel joining the transmitter to the receiver has been described as a system block which attenuates the signal and contaminates it with narrowband Gaussian noise. Even for the basic nature of the discussion in this chapter, this picture is over-idealistic since HF radio waves propagate in the ionosphere.

Of course, long distance HF radio communication is possible only because the ionosphere exists. Every 3 to 30 MHz radio wave that is to be received beyond the line of sight joining the transmitter to the receiver must traverse one or more ionospheric layers, and must then be reflected from another layer before an HF communication link is established.

The ionospheric layers are ionized and as such have profound effects upon electromagnetic waves penetrating them. The most advantageous of these effects, and the one central to the whole concept of ionospheric radio propagation is the progressive refraction of the transmitted HF radio wave. This process eventually causes the signal to be reflected back to earth and thus to the receiver. The other effects all serve in one way or other to degrade signal quality.

Foremost among these effects are absorption and noise which are always present in ionospheric radio propagation. Absorption reduces the transmitted signal power, $S_T$, and as such makes up a part of the attenuation attributed to the HF channel. Ionospheric noise combines with the noise already present in the signal. It is additive, narrowband and Gaussian in nature. These two effects always produce signal quality degradation in that the $(SNR)_S$ is reduced in both cases.

Besides these permanent effects, there are also random phenomena which act against signal quality. The most important of these is fading.
which may occur at any time and on any path. It results in reduction of power and thus of SNR. Fading also occurs in a semi-permanent form due to multi path propagation effects which are seldom absent from the communication links.

Ionospheric characteristics vary continuously with the Sun's position and level of activity. This results in an ionosphere which changes continuously from day to day, season to season, and over a fairly consistent 11-year solar cycle. To some extent these changes are predictable and formal forecasts are made by several bodies involved in routine ionospheric observation. These predictions are indispensable guides to the HF design engineer who uses them initially to quantify the probable characteristics of the path over which he hopes to establish communication.

Designs based on ionospheric predictions can only be approximate. In an optimum design, it is more important to know the probable variation about the mean of the predicted parameters. This leads to a concept of design based on criteria for reliable communication.

These criteria are the service grades defined as the SNRs needed for links which are operational 90% to 95% of the time availability over a given path, using a given form of communication system. Typical SNRs for various types of services under non-fading, or stable conditions are given in table (2.5.1). These values are modified in practice by the type of modulation used in the system and by the effects of random fading in the channel.

When designing an HF system, the engineer needs to know the properties of the ionosphere and to understand how these properties vary with time. This information enables him to define a qualitative and quantitative model of the intended communication path through the ionosphere. Furthermore a basic understanding of the ionosphere is helpful in obtaining a proper knowledge of the correct use of these data.
SNRs required for different types of service (stable signal and noise)

**TABLE (2.31)**

<table>
<thead>
<tr>
<th>Type of Transmission</th>
<th>Carrier/noise ratio or PEP noise ratio for a band of 1kHz (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Telegraphy A1</strong></td>
<td></td>
</tr>
<tr>
<td>8 bauds, poor quality</td>
<td>1</td>
</tr>
<tr>
<td>24 bauds</td>
<td>16</td>
</tr>
<tr>
<td>120 bauds, recorder</td>
<td>8</td>
</tr>
<tr>
<td>50 bauds, teleprinter</td>
<td>10</td>
</tr>
<tr>
<td><strong>Telegraphy A2, modulated carrier</strong></td>
<td></td>
</tr>
<tr>
<td>8 bauds, poor quality</td>
<td>-1</td>
</tr>
<tr>
<td>24 bauds</td>
<td>10</td>
</tr>
<tr>
<td><strong>Telegraphy with FSK P1</strong></td>
<td></td>
</tr>
<tr>
<td>120 bauds, recorder</td>
<td>10</td>
</tr>
<tr>
<td>50 bauds, teleprinter</td>
<td>6</td>
</tr>
<tr>
<td><strong>Script telegraph system</strong></td>
<td>11</td>
</tr>
<tr>
<td><strong>Telephony</strong></td>
<td></td>
</tr>
<tr>
<td>Double band, skilled operators</td>
<td>20</td>
</tr>
<tr>
<td>Double band, barely commercial</td>
<td>29</td>
</tr>
<tr>
<td>Double band, good commercial</td>
<td>37</td>
</tr>
<tr>
<td><strong>SSB and ISB</strong></td>
<td></td>
</tr>
<tr>
<td>1 channel</td>
<td>31</td>
</tr>
<tr>
<td>2 channels</td>
<td>33</td>
</tr>
<tr>
<td>3 channels</td>
<td>34</td>
</tr>
<tr>
<td>4 channels</td>
<td>35</td>
</tr>
</tbody>
</table>
explain the basic properties mentioned above in great detail. For this reason it would be superfluous to include a detailed treatise in the main discussion of this investigation.

However in order to update and unify the present state of knowledge on the ionosphere, and also to aid the development of the ionospheric model used in this thesis (Chapter 3), Appendix I contains a review of the basic ionospheric properties and effects which have such an important bearing on the whole concept of signal quality and optimum HF system design.

The discussion in this appendix deals first of all with the basic structure of the ionosphere and the fundamental theory of ionospheric radio propagation. A section is then devoted to predicted ionospheric parameters and their importance in good HF design. The practical methods used to quantify ionospheric path loss are then described and then three important qualitative and quantitative ionospheric phenomena are discussed. These are

1. Fading and Multipath
2. Ionospheric Radio Noise
3. Ionospheric Disturbances

Appendix I will be denoted AI in references appearing hereafter in this investigation.
2.4 Considerations in HF Antenna Design

The antenna systems which match the transmitter and receiver to the ionosphere are normally the most expensive blocks in any modern HF radio circuit. By today's standards, an efficient and reliable arrangement requires very large amounts of land - hundreds of acres sometimes - which has to be very carefully selected and prepared for optimum antenna operation. The antennas themselves are large structures, (a typical rhombic\(^1\)) measures some 200 m along the major axis at the relevant frequencies) and use large amounts of metal and cable for both the radiating structures and their support arrangements. Besides this a modern antenna 'farm' typically contains between 30 and 50 different HF antennas to ensure adequate coverage world-wide and thus economic and efficient operation of the wideband transmitters and receivers now being built.\(^3\)

Besides the obvious cost, HF antennas are not renowned for their efficiency in radiating available transmitter power for ohmic and unwanted radiation losses tend to be relatively high. Also HF antenna bandwidths are not in general very wide and so more than one antenna may be needed to cover even one frequency band accurately. Again, because of the large transmission distances normally encountered HF antennas generally have to be directional which therefore limits their use in directions other than that for which they were designed. To overcome this requires several antennas pointed in different directions since mechanical rotation of such large structures is virtually impossible.

With these considerations in mind, it is not difficult to appreciate why the HF antenna is often referred to as the weak link in the HF communication circuit. Long distance ionospheric communications usually involve high overall transmission losses, especially under unfavourable propagation conditions. The use of expensive linear high power trans-
mitors is therefore practically mandatory to ensure reliable HF communication channels. To justify this expense, the antenna systems must be carefully designed with specific objectives in mind. The most important of these are now considered.

1. The antenna system must provide high gain and thus high directivity in the preferred direction of communication. This is needed to maximise the power transfer from transmitter to receiver and to minimize radiation in undesired directions - thus reducing potential interference to other services.

The preferred direction of communication involves both the vertical and horizontal angles at which maximum radiation is desired.

The vertical angles pertinent to ionospheric propagation depend on the distance between the transmitter and receiver, the effective ionospheric layer height and the mode of propagation (1-hop, 2-hop, etc). Figure (2.4.1) shows the elevation angles used for 1-hop propagation for various virtual heights. This is however only approximate and in general it is desirable to carry out a detailed propagation study over the proposed path before fixing the vertical angles to be allowed for in the antenna design. For communication over distances greater than 4000 Km which are not covered by figure (2.4.1) the maximum signal will result from low-angle transmission in the range from 2° to 15°, with the lower angles generally providing better results.

The horizontal range of angles for beamwidth needed for point-to-point circuits depends on irregularities in the ionospheres and the effects of magnetic storms which cause deviations from great circle paths. It has been shown that direction of
FIG. (2.1.1)\textsuperscript{16} VERTICAL RADIATION ANGLES FOR I-HOP PROPAGATION
arrival of short-wave signals may vary as much as ± 5° in the horizontal plane because of these effects. This makes use of antennas with extremely narrow beamwidths undesirable, especially on paths which skirt or traverse the auroral zone. International broadcasting requires that the horizontal angle should subtend the target area with due allowances for path deviation effects.

2. The antenna system must have good input impedance characteristics over wide frequency ranges. This is necessary to ensure the optimum power transfer from the modern wideband, self-tuning HF transmitters to their antenna systems and to guarantee the converse at the receiver. In general, a compromise must be made between the acceptable impedance match, the antenna bandwidth and the power derating of the feeder due to a standing wave being set up. In practice, this is usually satisfied if the VSWR on the feeder does not exceed 2:1.

3. The antenna system must be properly sited for optimum operation. Where long distance circuits are involved, it is essential that a reasonably flat area of good ground conductivity is selected. The flat area should extend from the antennas to beyond the point at which the ground reflection of the signal will occur in the transmission direction. Similar requirements hold for receiving antenna siting, moreover in this case it is also important that the site is removed from sources which might produce interference to reception. Among these are nearby transmitting stations, power lines and other types of man-made interference.

Good ground conductivity is needed to provide effective radiation at the transmitting antennas. For horizontal polar-
ization, soil of moderate to good conductivity provides effective reflection for the range of angles usually used in HF radio transmission. For vertically polarized antennas, high ground conductivity is needed to prevent losses at the lower radiation angles. One way of aiding this in practice is to run 'counterpoise' wires under the ground surface in the required direction of transmission. Graphs for evaluating reflection coefficients are given by Terman.

Mechanically the antenna system must be designed to withstand expected wind and ice loading and to properly provide for the expected antenna sizes with regard to support structures.

The types of antennas satisfying the requirements in (1) and (2) above, and which also make relatively economic use of the rather special type of land needed for good operation fall into two classes. Their application depends upon the type of circuit and transmission distance involved. They are travelling wave antennas; in particular the rhombic antenna and the horizontal dipole antenna array, suitable for long distance fixed services operating over distances of 4000 Km or more; and the log-periodic dipole array for short distance fixed services below 4000 Km.

Horizontal dipole or curtain arrays of good design are characterized by relatively high efficiency and low power loss due to heat dissipation or radiation in undesired sidelobes. They are capable of providing high directional gain with good performance over a fairly wide frequency range.

Rhombic antennas provide effective operation and excellent input impedance characteristics over a fairly wide range. This antenna is normally terminated in a resistor load equal to its characteristic impedance. This results in a power loss from 2 to
3dB. In addition, the power loss due to sidelobe radiation is generally higher than in well designed curtain arrays. Despite this, the rhombic antenna is the most widely used in HF communication because of its reliability and low cost compared to the horizontal dipole arrays. It is also remarkably simple to construct.

The Log Periodic Dipole Array is a comparative newcomer to the HF field being first reported in 1957. However because of its wide-band characteristics (in a single structure) it has found immediate extensive use in short range fixed services.

In the initial stages of this investigation, the author spent about a year studying HF antenna systems with the aid of a computer. The basic aim was to provide efficient computer subroutines to model the antenna systems discussed above. In the first instance these subroutines would manipulate basic dimensional data of the antenna and generate values of gain, directivity and other relevant antenna parameters to be used in an ionospheric radio propagation programme being developed at the time by Dr. L. M. Muggleton at Edinburgh University.

In the context of this investigation, this study also provided the author with a necessary understanding of the antenna parameters which affect HF signal quality and thus added to his knowledge of the overall communication system.

During this study, the author was also to attempt to find ways of improving HF antenna performance with respect to signal quality; either by use of different construction techniques based on types of antenna designs. The results of this were not very encouraging in the short term. The study was therefore discontinued after the year and other relevant aspects of HF communication were considered instead.
However the basic aim was realized and programmes were produced for the three types of antenna systems. The sub-routines for the curtain array and the rhombic antenna were incorporated in Dr. Muggleton's programme and worked well. The programme for the log periodic dipole array was very time-consuming - and therefore expensive - and was thus not included as part of the main programme. Instead it was decided to use published data when necessary for the log periodic dipole array parameters.

Since very little emerged from this study programme which directly improved HF signal quality, these programmes will not be included in the main discussion. However since they represent a significant part of the work done in this investigation they are included in Appendix (2) where the programmes are described and discussed for the interested reader.
CHAPTER 3

HF SYSTEM DESIGN FOR OPTIMUM SIGNAL QUALITY
3.1 Introduction

The basic elements of an HF communication system have been discussed with particular attention being paid to the way in which they individually affect HF signal quality. It is now possible to design an HF system for signal quality optimization.

Such a design is based upon certain criteria. These are simply and intuitively deduced in section (3.2). They are then used in section (3.3) to compare the two main signalling techniques at HF—analogue and digital, and to make a choice of the optimum technique. The practical problems associated with this method are then discussed (section (3.4)); the treatment being based upon observed ionospheric behaviour rather than on the nearly ideal conditions used in chapter 2 where the emphasis was on basic concepts. Finally in section (3.5), the practical implementations of this optimum system are discussed. This will realize the concepts necessary to enable criticism of the optimum which in turn will suggest areas in which improvements may be possible, in conventional HF systems, to further improve signal quality.
3.2 Criteria for General Optimality

A successful engineering design is so adjudged on its ability to meet or surpass a given set of specifications. In general there are several designs which will satisfy these requirements. Each of these designs will be 'best' in one or more senses, but there will be only one which is 'better' in an overall sense than every other design. This design will normally be taken as optimum for the given set of specifications.

The choice of an optimum HF communication system must thus be based upon criteria which will accurately define the most important practical limitations to ideal communication in the HF channel. Among the most significant of these constraints are those due, directly and indirectly, to the characteristics of the ionosphere.

The ionosphere is the only HF circuit element which is not controlled by the designer and thus affects every aspect of system design. Its direct effect is the inevitable degradation in quality of signals passing through it. Certain signal types are affected less than others, and vice versa, which implies that one criterion may be defined with respect to ionospheric signal quality.

Directly related to this is HF system reliability. This is defined as the ability of a given system to maintain effective communication even under adverse conditions which might totally defeat other less reliable systems. The reliability specified for a given system is one of its most important parameters - having direct influence on the choice of transmitter and receiver hardware as well as modulation.

The designer's ability to achieve high reliability will be limited by specifications on system cost - and thus complexity. The budget allowed for a given system will thus have immense bearing on the ability
of that system to perform reliably and efficiently.

Another constraint on design is the traditional requirement for HF systems to be narrowband. This is due to congestion in the HF band and to the fact that wideband signals are normally severely affected by selective fading in the ionosphere. The optimum HF system will thus be required to use available channel bandwidth efficiently.

Finally, there are criteria based on the efficient use of modern HF technology. Foremost among these is the ability of the system to interface with the very efficient modern communication aids such as computers and video displays. Also greatly important is design flexibility, which provides for easy modification to suit foreseeable future advances in technology.

The criteria which will form the basis for the choice of the generally optimum HF communication system in this investigation may now be stated.

An HF system will be taken as optimum in a general sense if it is capable of satisfying a larger number of the following criteria than any other HF system.

1. Signal quality - defined as the intelligibility of the received signal. The system whose performance is least affected by ionospheric conditions will be regarded as optimum.

2. Reliability - defined as the proportion of time compared to available time that a system is able to maintain continuous communication. The system that is able to perform effectively for the longest time in steadily worsening ionospheric conditions will be taken as optimum.
3. **Complexity** - Defined as the amount of sophisticated hardware needed to realize a particular system. The optimum system will provide the best trade-off between complexity and good performance.

4. **Cost** - There are two components to be considered here:
   1. Design, development and installation
   2. Operation and maintenance

The system which makes most use of existing plant and of future investment to provide efficient reliable communication will be regarded as optimum.

5. **Bandwidth** - Defined as the amount of available channel bandwidth needed by a given system to perform efficiently. The system which makes 'best' practical use of the available channel will be considered optimum.

6. **Technology use** - The system which makes 'best' use of available HF technology and also produces the best interface with other forms of communication systems and aids will be taken as optimum. This system must also be flexibly designed with regard to the ease of modification to accommodate advances in the technology.
3.3. Optimality of HF Systems

The HF channel supports both analogue and digital communication links. A comparison of these on the basis of the criteria defined in section (3.2) will lead to definite conclusions regarding the general optimality of one or other of these systems with respect to HF communications.

The basic difference between these two types of information transfer is that in an analogue channel signals are transmitted in the form they occur whereas in the digital channel messages are encoded before being transmitted. This difference has far reaching consequences when the systems are compared as described above.

3.3.1 Signal Quality Comparison

Under the same ionospheric conditions, the coded digital signal is capable of much more accurate reception than the uncoded analogue signal. This stems from the fact that the binary coded signal has only two important states - MARK and SPACE, while the analogue signal has many states of amplitude, frequency and phase. The detector in the former case has only to accurately decide which signal state is present at any one time. In the latter case the detector has to accurately reproduce the message in amplitude, frequency and phase. This is difficult under even the normal ionospheric conditions of flat and selective fading which can severely affect any one of these parameters. The effect of these conditions on the digital signal is much less pronounced as the states have to be very severely distorted before it becomes impossible to recognize them - especially if matched filtering techniques are incorporated into the design of the system.

This means that a digital system operating under the same conditions as an analogue system invariably has a better signal quality capability.
This leads to the further important conclusion that in general, digital systems require less power than analogue systems to produce the same signal quality.

3.3.2 Reliability Comparison

Communication system reliability can be defined as the proportion of time that the system can be used for intelligible transmission from one point to another compared to the amount of time that its use is actually required. Due to unexpected ionospheric variations, the HF channel is never very reliable, but obviously the system which is least affected by these fluctuations will be optimum in this sense. For reasons similar to those expressed in section (3.3.1), the HF digital system will be inherently more reliable than the analogue system.

3.3.3 Complexity Comparison

The basic AM analogue channel is very simple in conception as is the basic non-coherent FSK digital channel. However the most complex analogue system is not nearly as complicated as its digital counterpart, as the latter would almost certainly use advanced error control coding techniques and regenerative repeaters to improve accuracy and reliability. This is one of the main factors in support of HF analogue systems - their basic simplicity and low cost compared to even moderately complex digital systems.

Where however the desire is for optimality, the disadvantages of analogue systems more often than not outweigh their advantages. Thus in general a relatively simple digital system will outperform a much more complex analogue system. This factor usually more than balances the extra cost of plant involved in choosing a digital system rather
3.3.4 Cost Comparison

HF system cost consists of two main items - system implementation costs and system operating costs. The former is directly related to the design complexity necessary to attain a required operational reliability; the latter includes such costs as those for maintenance, wages and power on a day-to-day basis.

Basically an HF system provides a communication service which is paid for by its users. Like any other service, its economic success will be judged on its ability to profitably offset investment against returns. This success will depend on its reliability and availability.

As demonstrated in section (3.3.2) the HF digital system is inherently more reliable than the HF analogue system, but from section (3.3.3) it is seen to be usually more complex and thus more costly to install and maintain. However the digital system has greater availability. First of all, because it is more reliable, the digital system will be accessible to users over longer periods of time. In the second place, as will be shown in section (3.3.5), a greater number of digital channels are available in a given system bandwidth. These two factors mean that the digital system will provide financial returns for longer periods and from a greater number of users. Therefore, even though the digital system would cost less to the individual user, it would almost certainly earn more over a given period.

3.3.5 Bandwidth Comparison

HF analogue systems are used for telephony and each channel requires a minimum bandwidth of 3kHz to adequately accommodate the range of voice frequencies. HF digital systems provide communication
using telegraphy typically implemented through teleprinters and data modems. A typical teleprinter circuit requires about 300Hz of bandwidth for adequate quality (100 baud FSK with 80Hz total shift between MARK and SPACE). On a channel to channel capacity, the digital system is capable of providing many more channels for a given system bandwidth.

3.3.6 Use of Technology Comparison

Commonly HF systems use multiplexing techniques to conserve bandwidth and to provide a real-time facility to the communicating parties. Use is also usually made of diversity techniques to promote accurate reception through the reduction of multipath effects. The fully comprehensive HF communication facility will thus include both frequency and time multiplex capabilities as well as most forms of diversity techniques. If these facilities are generally available, then the digital system will make most use of the system hardware.

Digital system performance is enhanced by both multiplexing systems and by most forms of diversity reception. Analogue systems find improvement only in the use of frequency multiplexing and in some forms of spaced diversity.

Existing HF technology is thus more widely applicable to digital systems. This gives them greater flexibility in design and thus more range of application from the standpoint of the user requirement.

Moreover, in terms of future prospects, digital systems have definite advantages. The modern trend is increasingly towards the storage and manipulation of data in digital form using high speed digital computers. The interface between a digital communication
system and the computer is obviously ready-made.

Again with the advances in integrated circuit technology and
application and with their ever decreasing costs, it should not be
long before the digital system begins to rival the analogue system
in cost and easy implementation.
3.4 HF Digital System Performance

It is not enough to assert that optimum HF radio communication is achieved by using digital transmission/reception techniques. The actual performance of these methods in the ionosphere must also be investigated. Such a study will determine the real performance limitations of the HF digital systems and will thus help to identify their weaknesses. This will in turn suggest areas in which technological effort will prove most fruitful. In order to investigate performance successfully, the communication channel must be properly characterised with respect to the transmitted signal format.

In chapter (2) the major ionospheric effects on HF signal quality were described. In the main, these are due to the variation of propagation conditions such as absorption and noise which are functions of geomagnetic activity, solar activity (11-year cycle) and solar storms. These changes take place on a time scale that is long, say, compared to one second. Alternatively signal format design is primarily affected by fluctuations which occur in fractions of a second, usually in the order of a millisecond or even microseconds. Neither signal design, filter design, nor detection design can alter the effects of the slower fluctuations and the only solution is to design enough power, antenna size, etc (system margin) into the system to achieve satisfactory operation under whatever conditions of long-term signal degradation are anticipated in the operational environment. Consequently, HF systems often appear to be greatly over-designed for their routine performance; but this excess ensures adequate performance under the relatively rare conditions upon which the design was based.

A degraded environment may also be made tolerable by changing the signalling. This method is used in adaptive systems where the information transmission rate can be altered to cope with changes in SNR.
common forms of adaptation for dealing with long-term variations on these fading channels include such simple expedients as changing the frequency in use on a long-distance HF circuit.

The digital signal format is thus assumed to be affected only by short-term variations, typically evidenced in fractions of a second. These arise because of multipath conditions associated with ionospheric transmission (Appendix 1) rather than from longer-term variations in the gross nature of the medium.

The basic characteristics of multipath propagation have been discussed (Appendix 1). The task here is to describe the effects of a channel thus affected on digital system performance. To this end the initial discussion must be enlarged to produce a more complete definition of the HF channel with particular reference to its fading multipath characteristics.
3.4.1 The Rayleigh Fading Multipath Channel

The characteristics of the fading multipath channel are best discussed by quantifying its effects on a general modulated signal. The main analysis in this investigation follows that given by Stein and Jones.¹

When a general modulated signal is transmitted through a multipath structured channel, the received signal varies from that transmitted in that the modulation of the former contains additional multiplicative components due to the effects of the fading radio channel.¹

This multiplicative modulation may be represented for a single frequency, f, transmitted by the complex-valued, equivalent low-pass, time-varying transfer function for the medium. This function¹ is a complex gaussian time process with respect to its dependence on time. This implies that its two quadrature components have the characteristics of random, independent gaussian variables with equal variances and zero mean. The resultant of these components is thus Rayleigh distributed in amplitude - hence its name. If, as is sometimes the case, a specular component is present, then the components have a non-zero mean and are better defined in amplitude by the Rice distribution.¹

In the case of digital data transmission, more than one frequency is sent and the relationship between several transfer functions over wide frequency ranges must be defined to deal with this. The relationship is now given by the normalized complex frequency cross-covariance. This correlation coefficient is defined in terms of the transfer functions of the channel for different f-values.

The definition assumes that the transfer functions are all locally stationary processes and that the covariance depends only upon the frequency difference, and not on the frequency itself. These assumptions are especially valid for digital signal pulses.
If two tones are close in frequency, as in most digital codes, the fading of the tones as they pass through a multipath channel will be clearly correlated, unlike the case when wider frequency differences are used and correlation weakens, as in most analogue transmissions. In this latter situation, severe fading (frequency-selective) is usually encountered. This point upholds the observation made in section (3.3) regarding the suitability of signals for HF communication.

This condition of fading becomes relevant if the signal bandwidth exceeds or is the same order as the frequency cross-covariance bandwidth. Therefore signalling systems for fading channels are usually limited in bandwidth to values over which the normalized frequency cross-covariance is near unity. This arrangement ensures that flat fading is the only significant fade-type associated with HF signalling systems.

For the ideal flat fading channel, aside from overall transmission delay, the received signal complex modulation component is equal to that transmitted multiplied by a complex Gaussian process.

The frequency selectivity of the channel is related to its multipath structure in the time domain. This relationship may be defined in terms of the complex frequency cross-covariance; this latter defines the rate at which the multipath structure of the channel causes independent fluctuations of the signal components. The form of the relationship is such that it demonstrates a Fourier transform relationship between the two parameters, and its significance is relevant to rapidity of fading considerations.

Fading rapidity is extremely important when system performance depends, as it often does, on obtaining a relatively noiseless monitoring of the instantaneous state of the channel. Systems designed for this need relatively long smoothing times which will be limited in useful duration by fading rapidity. This variable will also be
important in limiting digital system performance when the fading becomes rapid enough to affect coherence within a single pulse.

Fading rapidity is related to motions in the medium evidenced by Doppler shifts in the received pulses, and the multipath time covariance is clearly characterised in its spectrum by the Doppler broadening expected of the channel.

In conventional long distance HF radio propagation, at frequencies below the MUF, the fade rate at median level is usually found to lie between 0.1 and 1 fade per second, with 0.1 fade/sec (6 fades/min) being typical. The multipath is made up of two sections. Each layer reflection, owing to irregularities and layer depth, typically corresponds to a multipath continuum no more than a few hundred microseconds long with 50 - 200 μs being typical. However, the dominant multipath that restricts signalling bandwidths at HF arise due to the multihop nature of long distance propagation. Here some of the energy arrives at a distant receiving point via one, two or perhaps as many as ten hops. The differences in path delay between such modes are anywhere from a few hundred microseconds, for relatively short paths (hundreds of miles), to many milliseconds on extremely long paths. Typical multipath delays on circuits 1000-3000 miles long are in the range of 1 ms to 3 ms. Because of this signalling bandwidths at HF are conventionally below 500Hz for any individual pulse, although many signals may be sent on parallel paths in a multiplex arrangement.
3.4.2 Digital Signals in Rayleigh Fading

The model of slow, non-selective, purely Rayleigh fading describes the majority of HF (and other multipath) channels with enough accuracy for realistic performance appraisals, particularly with respect to digital signalling techniques. Specifically the assumption here is that the multiplicative modulation, or noise, varies so slowly with respect to the signal pulse lengths that it may be taken as constant over the period of a single pulse. Of course, over a long succession of pulses, the variation will become relevant. If this assumption is combined with the others made in section (3.4.1), then the mathematical model is such that each pulse transmitted is modified by

1. a simple amplitude multiplier, selected from the Rayleigh ensemble of the fading envelope, and
2. an additive carrier phase shift chosen from the associated uniform phase distribution.

These two components will then also be present in the output of the receiver filters applied to the detector at the sampling instant for making the binary decision for that pulse.

The error probability may thus be calculated on the above basis by adopting the expressions given previously in chapter 2. These results gave the conditional probability of error, for each kind of binary communication channel, as an appropriate function of the SNR for each pulse where the signal part of the SNR involved the multiplicative envelope factor for that pulse. In fading conditions, the average performance over fading must be ascertained. This involves the averaging of the conditional error probability over the ensemble of probable values of the multiplicative factor - over the Rayleigh distribution of this factor and, if present, the uniform distribution of the phase.

In general however practical digital communication systems use
sequences of pulses rather than single pulses (e.g., teletype and block codes). In this case, all pulses are assumed to have the same multiplicative noise factor and the same probability of bit error. The probability of sequence error may now be calculated conditional upon the value of the multiplicative process. The net probability of sequence error is then obtained by averaging this conditional probability of sequence error over the statistics of the multiplicative noise. Since a number of binary decisions are involved, which are non-independent due to the multiplicative noise being completely correlated over all pulses in the sequence, this assumption becomes less and less realistic as the sequence length increases. Further more the relation between the conditional probability of sequence error and individual bit error probability will in general involve a combinatorial expression. It is then practically impossible to carry out the final averages analytically, although numerical quadratures are possible. Moreover, the combinatorial expression involved is very particularly related to the nature of the character or sequence coding.

Fortunately, however, it appears that very little improvement is offered by coding in a communication system in which the data must be transferred at a fixed rate with fixed transmitter power over a slowly fading channel. Lower bounds have been derived for the probability of symbol error for a given transmitter power, order of diversity, and size of symbol alphabet. These bounds lie only a few dB (4-7dB) in terms of transmitter power, below the performance attainable without coding. This implies that the requirement for fixed information transfer or for fixed transmitter power must be removed if improved communication is to be obtained with or without coding.
Moreover, in dealing with sequence errors, the really important problem, particularly with respect to the use of 'long codes' or 'burst-error codes' for increasing transmission rate or a fading channel, is the removal of the non-selective, slow-fading assumption and the analysis with respect to more rapid or selective fading or both. In general, this problem has not been largely explored. The assumption that every pulse error is unrealistic when coupled with an assumption of slow-fading over each individual pulse except in one very important case. This is when interleaving is used - a process which does produce this effect of error independence within a sequence without regard to fading.

For the moment, therefore, the effects of slow, non-selective Rayleigh fading on individual bit error rates will be examined particularly with respect to systems based on FSK and PSK.

With the slow, non-selective, Rayleigh fading assumption, FSK and PSK can be considered simultaneously. As described in chapter 2, there are only two forms of conditional probability of error on individual pulses,

\[
P_{e}^{(1)} = \frac{1}{2} e^{-\alpha \rho} \quad \rho < 0.5 \quad \text{noncoherent FSK} \\
\rho = 1 \quad \text{ideal coherent DPSK} \\
\]

\[
P_{e}^{(2)} = \frac{1}{2} \text{erfc} \left( \sqrt{\alpha \rho} \right) \quad \rho = 0.5 \quad \text{coherent FSK} \\
\rho = 1 \quad \text{ideal coherent PSK} \\
\]

The average probability of single pulse error for the various signalling techniques may now be defined by averaging \((3.4.1)\) and \((3.4.2)\) over the fading statistics. However a problem is immediately apparent for either ideal PSK of coherent FSK. In view of the random fluctuations of carrier phase implied by fading how is the requirement for a phase reference for detection satisfied. Superficially it is possible to assume that the fading is slow enough to enable the derivation of a
suitably noiseless phase reference for either of these systems. For the moment, this assumption will be accepted as true, making it possible to derive the following average probability of error in slow, non-selective Rayleigh fading.

Noncoherent FSK: \[ P_e = \frac{1}{2} + \rho_0 \] \hspace{1cm} (3.4.3)

DPSK: \[ P_e = \frac{1}{2} + 2\rho_0 \] \hspace{1cm} (3.4.4)

Ideal PSK: \[ P_e = \frac{1}{2} \left[ 1 - \frac{1}{(1 + \frac{1}{\rho_0})^2} \right] \] \hspace{1cm} (3.4.5)

Coherent FSK: \[ P_e = \frac{1}{2} \left[ 1 - \frac{1}{(1 + \frac{1}{2\rho_0})^2} \right] \] \hspace{1cm} (3.4.6)

where \( \rho_0 = \rho \) is the mean SNR averaged over fading at the sampling instants at the filter output. These results are plotted in figure (3.4.1). Again there is an exact 3dB difference in performance between coherent FSK and ideal PSK, as well as between noncoherent FSK and DPSK. These are simply a reflection of the exactly 3dB difference in the relative performance of these systems at all levels of SNR.

For comparison the noncoherent FSK and ideal PSK curves for error rate with steady signal are reproduced as the dotted curves in figure (3.4.1) using the same abcissa to now denote the steady SNR. The severe degradation owing to fading fluctuations about the mean received signal is obvious. For a single binary channel subject to slow, non-selective Rayleigh fading, additional system margins of the order of 10dB (mean level received in fading as compared to required mean SNR in the absence of fading) must be provided for error rates around 0.01, with approximately an additional 10dB required for every order of magnitude further decrease in allowed system error-rate. Indeed from the equations, the asymptotic performance for large average SNR shows that the error-rate for all these systems is exactly inversely proportional to the mean SNR. By way of comparison, changes in SNR produce exponential changes in error-rate with steady signals, whereas with Rayleigh fading,
FIG. (3.4.1) Pe for several binary systems in Rayleigh fading.
changes are only on a 1:1 basis. This is perhaps the dominant qualitative difference. It also indicates the nature of the additional design burden imposed upon systems required to operate in fading and the resulting importance of techniques for overcoming the effects of fading.

Qualitatively, the severe degradation introduced by fading is of course due to the finite percentage of very low signal levels occurring in Rayleigh law fading, during which the conditional probability of error is correspondingly very close to 0.5. In this sense the importance of diversity techniques may be considered to lie in modifying the probability diversity function for $p$ so that it tends to be very small for small $p$ rather than the exponential behaviour normally predicted for which the p.d.f. actually peaks at $p = 0$. This modifying capability is probably the most important single characteristic of any method employed to improve digital system performance in the fading HF channel.

It must be remembered that the results presented above are based upon the assumption of slow and non-selective fading. This model is applicable for a wide range of practical applications but is in essence only a first order model. For instance, for every Rayleigh fading channel, there are instants during which the signal level fades through or near zero value. At such instants the signal level may be changing rapidly, even though on average it is slowly fading. A pulse at this instant may be badly distorted, so badly that the wrong value might be chosen even if there were no additive noise. If, for example, this occurs on the average to one pulse out of every $10^4$ pulses, there would be an irreducible error-rate of $10^{-4}$.

Considerations such as these limit the use of the above model to the exploration of expected performance rather than actual performance except in the rare circumstances when the assumptions are valid.
However the use of these approximate descriptions is extremely helpful in predicting the areas in which technological innovation can be applied to improve performance.
3.4.3 Diversity in Digital Systems

The performance of single channel digital data transmission using non-fading signals is severely degraded by Rayleigh fading in the ionosphere. The extent of the degradation is often measured in tens of dB and major additional expenditure can be incurred if increased system margin is to be used for compensation. Indeed increased costs are often great enough to prohibit the use of an HF circuit altogether. Fortunately there is an alternative solution which uses more sophisticated modulation and reception techniques than already discussed.

These methods are less sensitive to the fading effects, and are in general cheaper than when system margin increases. Of these, the best understood and most widely used are the multiple receiver combining techniques - the diversity reception techniques.

The diversity principle may be applied to both analogue and digital systems in the sense that it is based simply on reducing the fading range, especially the probability of near-zero fades for the SNR.

There are many different methods of diversity reception depending upon the techniques used for signal transmission. Among these are

1. Spaced antenna diversity $^3,4$
2. Frequency diversity $^5$
3. Angle (of arrival) diversity $^6$
4. Polarization diversity $^3,4$
5. Time (signal-repetition) diversity $^1 (p.362)$
6. Multipath diversity (Rake) $^1 (p.362)$

Of these, the last two are at present practicable only when used with digital data transmission.

Except in the case of polarization diversity, there is no theoretical limit to the number of diversity branches provided for combining. Moreover, selection of the instantaneously strongest signal
is not mandatory; other combining techniques are widely used and include

1. Selector (or selection) combining
2. Maximal-Ratio combining
3. Equal-gain combining

These all involve linear combining networks and are thus the only type applicable in principle for distortionless reception of analogue signals. Several digital data streams may however be usefully received by other combining methods. The most familiar of these is square-law combining which is based on an extension of the matched filter concept to determine optimum diversity combining techniques.

Since long-term fading effects of the type already discussed affect all diversity branches similarly, the above techniques cannot compensate for degradations due to these variations. Diversity can therefore only be used to reduce fading effects over the statistically stationary short-term conditions described in section (3.4.1).

The diversity principle assumes the availability of M, say, distinguishable, differently fading, signal reception channels. The diversity receiver then chooses at each 'instant' the best of the M signals, or some desirable additive combination of all the signals. Since the system does not have infinite bandwidth, it cannot truly function on an instant-to-instant basis and must have some reasonable response time. This combiner time constant has in practice to be substantially shorter than the reciprocal of the fading rate to ensure proper diversity operation.

Again linear combining techniques depend upon the derivation of an estimated signal level at each diversity branch. Ideally this estimate should be noiseless except for additive receiver noise. This assumption is valid for most terrestrial radio channels but must be used with care.
When it is not valid, some other form of combining must be used.

Within the above limitations, diversity performance may be discussed with reference to the so-called diversity gain defined in terms of an 'outage' rate. This definition recognizes that while the median value of the SNR at the combiner output represents a greater value than that available on any single branch, the most significant aspect of diversity lies in reducing significantly the fraction of time in which the signal drops to unusable levels. This fraction of time is the 'outage' rate specified relative to a particular reference level based on the mean output noise level of the combiner. Since 'outage' rate may also be reduced by increased system margin, a criterion for diversity improvement may be taken as the saving in the median carrier-to-noise ratio required per branch in order to remain within some specified 'outage' rate, as compared to the median carrier-to-noise ratio required on a single channel for the same 'outage' rate.

The calculation of very low(1%) 'outage' rates involves the near-zero 'tails' of the Rayleigh fading pdf, and since the relation between 'outage' rate and median SNR is not linear, diversity improvement as defined here is only distantly related to average SNR improvement. This implies that a direct engineering meaning is obtainable only if signal quality is directly related to the 'outage' rate at some threshold value specified with respect to the mean receiver noise. With the emphasis on digital data transmission, a more directly meaningful definition would be given in terms of the requirements for meeting a certain error-rate specification. With diversity a significantly lower median signal level will be required on each branch than for the non-diversity channel, to achieve the required error-rate performance. Any such decrease is immediately equatable to a saving in non-diversity system margin and hence represents a more specific definition of diversity.
gain.

From the above it may be deduced that combining will produce better results than just selection of the instantaneously strongest signal. This follows since the combiner at the very least is capable of using the instantaneously best signal.

Anything, usefully added then must, by definition, represent an improvement. The advantage of the general combiner is that when no acceptable signal exists on any of the M channels, but when at least some of the signals are large enough, an effective combination can still produce an acceptable output SNR. However, when M is small, as is generally the case, and when error-rates are reasonably low, the use of a selector results in greater diversity improvement than the use of a general combiner. This effect is strongly evident in quantitative studies of diversity performance.

The effective SNR distribution achieved by various linear combiners is shown in the curves of figure (3.4.2) for several values of M, in the case where all channels are of equal SNR. These are the probability distributions on arithmetic probability paper, with the non-diversity curve for the same mean SNR plotted for reference (0dB is mean). It is clear that increasing orders of diversity increasingly narrow the fading range. Also, as mentioned above, the differences between the results for any particular combiner types are much less significant than the difference between any of them and the non-diversity curve. The optimum (best) is obviously maximal-ratio combining, as is generally accepted, but equal-gain combining for all orders of diversity is within 1dB of optimum. If further pdfs of the combiner outputs for low SNR are examined, it will be found that the parallelism in the curves holds for all levels of diversity even at low SNRs when most errors occur in digital data transmission. This is an important
FIG. C3.4.24  COMBINER O/P PROBABILITY DISTRIBUTIONS
factor in the theoretical analysis of complicated situations, where, for example, mathematical analysis is convenient for one form of combining only when accurate inferences may be made regarding the actual performance of other, perhaps more easily implemented, techniques.

So far independently fading channels have been assumed. It may be noted here that fairly high correlation can appear in the fading before any significant degradation in diversity performance becomes apparent.\(^9\)

The performance of digital data streams with diversity reception may now be discussed. A simple example of this kind of system is noncoherent signalling with maximal-ratio combining and independently fading channels\(^1\). In slow fading, the error-rate depends only upon the probability distribution of the output SNR at low values of \(P\). From this it may be deduced that the performance difference between the several combiners is exactly the same as noted earlier in comparing outage rate performance. Furthermore, it may be expected that the comparative results are not importantly changed when coherent detection systems are considered.\(^1\)

So far the assumption has been made that the signal level can be monitored to provide a reference for the combiner. This cannot always be done in adverse ionospheric conditions. In this case coherent signalling and detection may not be usable due to the unpredictability of phase variations in the signal. With noncoherent signalling, such as two-tone FSK, it has been shown\(^10\) that under the assumptions of slow multiplicative fading it is possible to use a best possible combining technique known as square-law combining. In each branch, noncoherent matched-filter detection is carried out separately for each binary state. Assuming that all branches fade independently and have
the same SNR, the squares of all the MARK filter output envelopes are then added; as are the squares of the SPACE envelopes, and the decision made on the basis of the larger sum.

Figure (3.4.3) shows a comparison of diversity combining techniques, for several H values, for binary FSK with Rayleigh fading. It may be noted that at high SNRs there is only 6dB performance difference between square-law combining with noncoherent signalling and maximal-ratio combining with ideal coherent FSK. Moreover, of this 6dB, 3dB may be recovered if the channel is slowly fading, by replacing noncoherent FSK with DPSK in an noncoherent mode with square-law combining.

There is another interesting possibility with square-law combining. If a channel is known to be fading so rapidly that there is a frequent loss of coherence, even over the period T of a single information pulse, then it will no longer be suitable to use a linear receiver which smooths over intervals of duration T. Its averaging action over each of the quadrature components will tend to produce a zero overall average. On the other hand, there will be subintervals of duration $T/k$ (k integer) over which the coherence may be considered to hold, and over which linear-filter smoothing will be effective. Over T, k smoothed samples of the received waveform may be obtainable, for which the signal contents do not further add coherently but which may be combined for decision purposes.

The optimum combining now becomes obvious. The envelope of each of the samples is squared; the MARK filter outputs are then added and compared to the sum of the SPACE filter outputs. In the limit the squared envelope of the received signal-plus-noise is continuously formed and integrated over the signal pulse duration. Even when k is finite, this square-law detection-integration is equivalent to choosing the k envelope samples for the processing described above.
Figure (3.1.3) Comparison of diversity combining techniques for binary FSK with Rayleigh fading.
This type of detection is called energy detection or radiometric detection and is a common technique in radio- and radar astronomy. Its performance is akin to time diversity operation. Care must be applied in the use of this form of detection for binary waveforms since the signals are required to be orthogonal over the subintervals if distinguishability is to be preserved since the result from each of the subintervals is noncoherently detected. This is readily achieved in FSK with sufficiently wide tone spacing. The usual DPSK signal will however, no longer satisfy this criterion over subintervals.
3.4.4 Error Control Coding

The design of the optimum HF communications system involves the use of techniques such as diversity, to compensate for signal degrading effects in the ionosphere. Another such technique of particular relevance to HF digital signalling is Error Control Coding (ECC).

Error probability in digital communication is a direct function of the SNR. If the system margin has been maximised and errors are still unacceptably frequent, then ECC generally provides the required performance improvement.

Basically ECC is the calculated use of redundancy indicated by information theory as being necessary for attaining errorless communication. Extra bits are systematically added to the transmitted message. These bits convey no information but enable the receiver to detect and even correct errors in the received message. In practice, cost considerations prohibit the achievement of totally errorless communication, but acceptable error rates for most needs are generally achievable by using ECC as a standard in optimum system design.

In the context of HF data communications, only binary codes require attention. Due to the complex nature of ionospheric variations, the use of more complex multilevel codes is prohibitively expensive. Besides this, binary codes have one unique feature — if the digits in error are known, then the correct bits are immediately derivable. The basic theory concerning ECC is well documented and requires little discussion here. However its performance with respect to HF digital signalling is important.

ECC performance in the HF channel is directly related to the error characteristics of the ionosphere, and an accurate description of these properties is essential for effective ECC design. As seen earlier for digital signalling without ECC (section (3.4.2)), the HF
channel is difficult to analyse for its error characteristics and several
simplifying assumptions had to be made to enable any reasonable conclusions
about performance to be made. Efficient ECC design depends upon more
realistic definitions based upon statistics derived from years of
observation of typical errors and of error connected characteristics
that exist in the ionosphere. The most important of these are:-

1. The Average Bit Error Rate (BER) - defined as the ratio of
total number of errors to the total number of transmitted bits. This
is an important first order statistic, but it does not contain any
information about how errors occur.

2. The Gap Distribution - A gap is defined as a region of error-
free bits between two errors; its length being the number of error-free
bits. The Gap Distribution is the plot of the cumulative relative
frequency of the gap length against the length of the gap. This parame-
ter gives some indication of the randomness of the channel.

3. The Burst Distribution - This is one of the most important
of the error characteristics. If a channel is not random, there are
more errors in certain portions of the data stream than there are in
others. These portions are normally referred to as Bursts. A Burst
is normally defined as being that region of the data stream within
which a minimum of two errors exist, and always begins with an error
bit immediately preceded by a correct bit. A specified minimum error
density must exist in the region, defined as the ration of error bits
to total bits in the Burst region for maximum length Burst. The Burst
region always ends with an error bit that is immediately followed by
a correct bit. With the Burst well-defined, the Burst distribution is
then the plot of the cumulative relative frequency of the Burst against
the Burst length.

The Burst information obtained above enables better use of a
communication channel characterised by this property. If the Burst distribution is known for a particular channel, then an ECC technique may be designed to correct the errors in the Burst. This is however not practical in the HF channel where the bursts are usually quite long and thus would require a very long code. A more practical method is to use interleaving together with a short random ECC code. The interleaving disperses the errors in the Burst into code words within the capability of the ECC code. Unlimited increase in the degree of interleaving tends to randomize the errors and eventually converts the Burst channel to a random channel. However interleaving also reduces channel capacity and so there is an optimum degree of interleaving.

4. The Burst Interval Distribution - Burst interval regions contain those sequences of bits which lie between Bursts. Every Interval always begins with a correct bit immediately preceded by an error bit, and ends with a correct bit. The error density in the interval is always less than the minimum error density defined above for the Burst region. The error-free gap may be considered as the limiting case for the Burst Interval.

This fourth parameter is a high order statistic which gives some indication of the dependence between the Bursts. It provides further aid in designing the interleaving, and together with the Burst distribution provides the basis for determining the optimum degree of interleaving with respect to the specific code.

5. Cluster Distribution - A cluster is defined as a region of consecutive errors. This distribution is a plot of the probability of the number of consecutive errors against the number of such errors. Usually in a reasonable transmission interval in an HF channel, say four months, there are not enough clusters to yield good statistics. However they do account for some spurious phenomena and must therefore be noted.
The HF channel may thus be characterized by random, burst and periodic errors occurring either singly or in various combinations at any given time. Various codes and ECC techniques have been widely investigated for digital channels prone exclusively to each of these error types on its own. In general the normal HF channel would defeat the efficiency of these codes in every instance because of the possibility of more than one error type occurring at the same time. However, various techniques may be used which will make use of these codes practicable for HF digital signalling.

There are two types of code relevant to HF digital data transmission - block and convolutional. To enable efficient usage of these codes requires a variety of time dispersion techniques such as diffuse convolutional coding, time-spread coding and interleaving which has already been introduced.

The most commonly used codes are those which use as many parity bits as informative bits (ie) rate $\frac{1}{2}$ codes. The use of these codes is dictated in military systems by the requirement that data rates be greater than $75 \times 2^h$ where $h \geq 0$ (cf MIL - STD - 183c). Since most commercial manufacturers of communication equipment also build for the Government, commercial users are frequently placed under a similar constraint.

Block, or cyclic, codes used at HF are usually derived from Bose-Chaudhuri-Hocquengheim (BCH) codes. These are a class of cyclic random error-correcting codes. They are described by the notation $(n,k,e)$ where $k$ information bits are encoded to form $n$ total bits, of which $e$ error bits may be corrected. The code rate is defined as $k/n$. The HF Burst error problem may be solved by interleaving these codes and for this reason interleaving is always applied to block codes used at HF. In interleaved codes, the words of length $n$ are corrected by
encoding information bits \((l)\) to generate parity bits \((P)\) according to

\[
I(k-l)_{m+1} + \ldots + I_{2m+1} + I_{m+1} + I_{1} + s \quad \text{where} \quad s = 0, \ldots, \ldots, (m-1) \\
\]

\[
P(n-k-l)_{m+1} + \ldots + P_{2m+1} + P_{m+1} + P_{1} + s 
\]

The interleaving parameter is \(m\). If \(m = 1\), there is no interleaving.

To demonstrate the use of interleaving the following example is given.

The modified Goly 35 is chosen. This code may be constructed as a BCH code with definition \((24,12,5)\), from which it may be seen that the code rate \(= \frac{3}{4}\) as expected. If \(m = 1\), then twelve consecutive information bits are encoded to create a 24 bit word. In this code word, if three errors are introduced they can be corrected. If six errors occur, no correction is possible and further errors are generally created. If however \(m = 2\), then twelve bits chosen as either the odd or even bits from the first twenty four bits are encoded together to obtain an interleaved block of two code words. If six errors now occur, they will fall three each in the two distinct code words and if the new total block contains no more errors, all errors will be corrected.

From the viewpoint of consecutive errors, interleaving redefines the code as \((mn, mk, me)\). It can therefore be seen that the object of interleaving is to spread a burst of errors into the guard space which immediately follows the burst. Since the guard space has few errors, a code block is created where the actual number of errors in the code block is less than \(me\) and, since the bursts have been spread out, error correction can be achieved. The penalty for improved performance is time delay.

In an \((n,k,s)\) code, all \(n\) bits must be received before any corrections can be made. Thus there is a delay of \(n\) bit times at the data rate. The delay in the interleaved system is \(2mk\) information bit times, with the
general delay equation given by (16)

\[ \text{Delay} = \frac{2m_k}{(\text{Information Data Rate})} \]  

(3.4.8)

The performance of these codes is described in terms of the percentage of input errors corrected. However for ease of interpretation, the Improvement Factor is defined as

\[ \text{IF} = \frac{1}{(1 - (\% \text{ of errors corrected/100}))} \]  

(3.4.9)

Brayer (16) has investigated the performance of a large number of BCH codes using a computer programme which took the actual error patterns measured by other investigators and found the percentage of errors corrected for various codes. The results were presented graphically showing the IF as a function of code rate. The channel data rate was fixed at 2400 bits/sec since all the measured error patterns were at this rate. The BCH codes evaluated all fell within the definition \( n = 2^q - 1 \) for \( 2 \leq q < 8 \). He found that for the same delay all codes were approximately equal in performance, and concluded that performance was a function of the product \( mn \) and not of \( m \), nor \( n \) individually, if delay is taken as the basis for comparison. The values of \( m \) used were 1, 5, 19, 41 and 89. In all cases it was possible to obtain 100% correction if the code rate was sufficiently reduced; (ie) a greater portion of the channel allocated to parity checks. It was therefore found that the error rate in HF digital communications, at the cost of a reasonable penalty delay, could be substantially reduced using interleaved BCH codes. He also compared these codes to symbol codes which have a greater dense burst correcting capability and found that at HF they were inferior to the interleaved BCH codes. The reason for this was that HF burst are generally low density diffuse types.
Brayer also considered a modification to the single level of interleaving described above. This was Tandem Interleaved Cyclic Coding. In this format, after interleaving by an 'outer' coder the encoded information is passed to an 'inner' coder which performs interleaved encoding on the total data stream. When this was done, the result was that the BER at the output was rarely poorer than $10^{-6}$. The coded blocks were almost all error-free provided that the overall code rate is in the neighbourhood of $\frac{1}{2}$ and the delay was between 2 and 4 seconds.

Convolutional codes may also be used to reduce Burst errors in the HF channel. These codes, also known as sequential or recurrent codes, differ from block codes in that the check digits are continuously interlaced in the coded bit stream rather than being grouped into words. The encoding/decoding process is therefore a continuous process, eliminating the buffering and storage hardware required by the delay characteristic of the block codes. Convolutional coding theory is very intricate and has not to date been unified. However a good account of its important characteristics may be found in a paper by Viterbi which also discusses the difficulties involved in unification. Nevertheless several investigations have compared the performance of convolutional codes to that of block codes. Among them are McManamon et al and Brayer.

McManamon discussed two convolutional coding formats. These were Massey's Try and Error (TE) and Robinson's Self Orthogonal (SO), denoted respectively (12,6) and (14,7). These are both orthogonal codes and enable the use of threshold decoding with its added advantage over other methods of using majority logic. These codes were compared with two block codes of the type already discussed - the modified Golay (24, 12, 3) and ordinary BCH (15, 7, 2). All codes were
compared using both diversity and non-diversity reception. With non-
diversity reception it was found that the two block codes reduced the
error-rate by about half while the two convolutional codes had negligible
effect. Interleaving was then introduced and though the performance
improved in every case, the block codes maintained better error-rate
reduction capability. Diversity operation produced better code performance.
Modulation diversity, itself a forward 1/2 rate error control technique,
is as effective as the four short forward 1/2 rate codes used without
interleaving. However when interleaved the four codes achieved up to
an additional order of magnitude improvement with diversity when compared
with non-diversity. In all cases it was found that time dispersion
techniques such as diffusion and interleaving were necessary for the
short codes to perform measurably better than dual inband diversity
without coding.

The more recent paper by Brayer also considered two convolutional
codes - a non-adaptive Massey diffuse convolutional code with threshold
decoding and an adaptive Gallagher code. These codes were chosen because
they corrected approximately the same proportion of errors in a random
channel and this proportion was the highest of all the 1/2 rate codes.
The block code used for comparison was the modified Golay (24, 12, 3)
used in two decoding modes to achieve adaptive and non-adaptive coding.
The error correction technique used was the random correction method due
to Kasami which is widely used for forward error control with block
codes.

It was found that the order of merit with respect to performance was

1. adaptive convolutional coding,
2. non-adaptive convolutional coding,
3. cyclic coding (without interleaving).
It was also found that the concatenated cyclic coding described previously outperforms all these and thus rate highest of all ECCs. It may be deduced therefore that for HF digital data transmission, interleaving used with the short block code is the most effective method of improving error-rate performance at the present state of the art.

Besides the use of ECC, performance may also be further improved by the use of the automatic request for repeat facility (ARQ). When the two are used together the symbol, character or word-error-rate is more significant than the BER. Word or block retransmission requests are significantly reduced using forward error correction, thus preventing the ARQ from temporary breakdown, caused by the continuous request for retransmission of a block under high-error conditions.

Finally with reference to optimum digital signalling using ECCs and ARQ is the question of adaptive HF communications. The adaptability of a communications system may be loosely defined as its ability to modify itself to produce continuous optimum performance within a randomly varying environment. This implies that adaptive systems generally operate on the basis of real time monitoring of conditions in the communication link. Betts discussed the whole concept of adaptability in his recent paper. The analysis was theoretical and demonstrated how the phase perturbations due to ionospheric propagation could be employed for real time channel monitoring through their interpretation as a 'phase-error' rate. This phenomenon could also be used to assess the likely error rate associated with the HF digital channel. From analysis of these data it would be possible to control

1. the time at which to change carrier frequency,
2. the best standby frequency,
3. the degree of error control required to protect the information.
These tasks have traditionally been the responsibility of human operators and so one of the major advantages of an adaptive system as described above would be the elimination of the often intuitive system-update decision of the experienced human operator. Besides this, the time span between the system becoming inoperable due to channel conditions and a change of the system operating parameters will be much reduced by continuous monitoring followed by updating on an automatic basis.
3.5 The Optimum HF System

It has been demonstrated that HF digital communication systems satisfy more fully than the HF analogue systems the criteria for optimality defined at the beginning of this chapter. For this reason, the generally optimum system for ionospheric communication will be based upon digital signalling techniques. The performance attainable with these systems depends on several factors among which the most important are:

1. accurate characterization of the channel parameters between the two points of interest,
2. correct choice of the basic signalling technique such as FSK, PSK, and DPSK,
3. proper use of techniques such as diversity, ECC and ARQ in the design stage.

These decisions can be made based upon the discussions in this and other chapters of this thesis where the relevant factors have been discussed in detail.

The basic components of a digital communication system are shown in figure (3.5.1). The source symbols come from a digital data source such as a teletypewriter and may undergo source encoding through the application of block or convolutional coding techniques. This converts them to channel symbols which are then modulated into a carrier, using some form of shift keying, before being transmitted into the forward channel. Here fading and interference may degrade the signal and thus produce errors in the received symbols. If only forward transmission is being used the forward channel signal design (e.g. FSK), or ECC introduced at the source encoding stage must meet the required error probability ($P_e$) specified for the channel. If an error detection capability is added to the signal and a feedback channel is available the ARQ can be
FIG. (3.5.1)  BASICS OF DIGITAL COMMUNICATION
used to eliminate the errors, but a time delay must be accepted.

In the forward channel, received power and bandwidth are the two basic communication resources. The per-symbol transmission quality, measured by $P_e$ is entirely a function of the SNR, the distance between alternative signals and the symbol integration time. Power-limited cases (SNR constrained) occur when it is difficult to increase transmitted power, reduce physical distance between transmitter and receiver, or improve the antenna system.

If signal quality must be improved in power-limited cases, one alternative is to allow the channel symbols to become longer and hence to increase the final decision integrating time. This requires reducing the symbol rate which is often not possible. A second alternative is to encode groups of source symbols and thus to form channel signals which have a greater distance between the alternate signals.

Bandwidth-limited cases occur when adequate received power is more easily achieved than increased bandwidth. Ionospheric radio channels are bandwidth-limited due to competition among contending users. In these cases the power adequacy is exploited by efficient use of the available bandwidth. This may be achieved by grouping source symbols and forming channel signals which are closer and more dense in signal space through the use of some form of multiplexing.

The receiver decides which of the alternative forward signals was transmitted. If forward error control (FEC) coding is used, then the receiver design, along with the signal design, transmitter power and channel characteristics determines the error-rate and hence the performance. If time delay is acceptable and a feedback channel is available so that ARQ can be used, the final performance measure is the average effective rate of the final correct symbols and the undetected error-rate.
From the above it is clear that the configuration of the optimum digital communication system in a given situation is dependent upon whether delay is acceptable or not in effecting the required communication facility. Where speed is important, as for instance in real-time military operations, then a forward transmission capability is almost mandatory as long as an adequate error-rate can be achieved. Where however accuracy is important and data might be stored for later use or for hard-copy acquisition and delay is thus unimportant, then ARQ is desirable.

The decisions for optimum system realization in the two cases are different and bear further discussion.

3.5.1 Forward Transmission Systems

Ideally the forward transmission system is required to transmit the source symbols at the specified channel rate with an acceptable BER through a channel characterised by multipath fading and random interference. Correct operation is constrained to take place within a symbol decision interval based on the time interval over which the receiver makes a final decision about the received symbol. This interval contains one or more source symbols of lesser interval. In any decision interval the transmitter can send one of a number of alternative signals and given that reception takes place, the receiver must decide which symbol was sent.

Figure (3.5.2) depicts the general configuration of a forward transmission system. The source produces one of m symbols at a rate of R (bits per second). As already discussed (section (3.4.1)) m = 2 in general for HF systems. Source symbols may be entered directly into the channel modulator, or they may first be encoded. Based on previous discussion (section (3.4)) this latter step ensures the best performance
Fig (3.5.2) Forward Transmission System Components.
HP digital systems and is implemented using some form of ECC. This also ensures best use of channel bandwidth and transmitter power. The modulator then produces the channel signals as discussed previously (section (2.5)) and optimum modulation is chosen based on the discussion in section (3.4). The demodulator removes the carrier and the received contaminated and distorted signals are fed to the correlator. This process would almost certainly be based on some form of diversity (section (3.4.3)) to achieve good performance and the correlator would then consist of a matched filter type as discussed in section (2.2). The decoder then operates to decode the encoded message for display.
3.5.2 Feedback Transmission Systems

The major alternative to the forward transmission techniques described in the last section is the additional use of a feedback channel to enable ARQ. This operation makes use of forward error detection, as opposed to the error correction described above.

In this system, the stream of channel symbols is segmented into blocks and an error detection code is applied. The algebraic error detection codes are relatives of the block codes used in ECC. Since unpredictable retransmission will be required, the data at the transmission must be temporarily stored in a buffer. Figure (3.5.3) depicts the general configuration of a practical ARQ channel. If an ECC code is used on the pass through the forward channel, it would appear in the position shown by the dashed boxes.

ARQ systems are implemented in either of two ways:

1. Stop and wait.
2. Continuous.

In the first of these, which is the most widely used in present day digital data systems, the transmitter sends one block of encoded data at a time and waits for an acknowledge (ACK) signal from the receiver before proceeding. If a negative acknowledge (NACK) is received, the transmitter repeats the previous block. The advantages of stop and wait ARQ systems are:

1. They are easy to implement.
2. They require buffering at only one block.
3. A single channel can sometimes be used in a turnaround mode. Here a single channel serves alternately as both a forward and feedback channel. Usually a settling time is involved in each turnaround, during which no symbols can be transmitted.
Fig (3.5.3) ELEMENTS OF AN ARQ SYSTEM
In continuous ARQ systems, the transmitter sends the error-encoded blocked symbols continuously and a simultaneous feedback channel is required. When an error is detected, the NACK signal received at the transmitter via the feedback channel causes the transmitter to repeat the block found to be in error and the subsequent blocks.
The number of blocks retransmitted now depends on the channel round trip delay, and is usually more than one. Although more efficient, this method is more complex to organize and requires more storage. In addition, a separate feedback channel is always required. The cost and amount of bandwidth needed to implement such a system at HF is usually prohibitive, but its quasi-real-time operation could justify the expense in some circumstances.
3.5.3 Overall System Appraisal

The optimum HF system has so far been defined and implemented using current HF theory and practice. Choice of the optimum, based on the criteria developed in section (3.2) also requires that the optimum system should most easily accommodate advances in communication theory and technology. The implication here is that the system design should employ techniques which encompass concepts compatible with current research trends, both in HF and other communication fields.

Success in some areas of current research would indeed benefit both analogue and digital HF communication systems equally. These areas include:

1. Ionospheric simulation.
2. Antenna Design.
3. HF linear amplifier design.
4. Stable oscillator design.
5. Multiplexing design.
6. Real time measurement of ionospheric characteristics.
7. Diversity reception techniques.

However there are some areas in which success would mean improved operation only in the case of HF digital communications. The most important areas are:

1. Error control and detection methods.
2. Adaptive HF systems.

Of these the last offers the most exciting possibilities involving as it does the use of very new technologies and the realization of greatly improved signal quality in the HF channel. Adaptive communication in the HF channel also offers improved reliability and cost effectiveness in ionospheric digital communication and it is obvious that improved
coding techniques would always be welcome.

It is not intended to delve deeply into the usefulness of adaptive communication or new coding methods since these are well covered in the literature already referenced. New implementations of the matched filtering concept are however most attractive and bear importantly upon the optimization of signal quality in HF communications, which is the basic aim of this investigation.
CHAPTER 4

PULSE COMPRESSION IN HF DIGITAL DATA TRANSMISSION
4.1 Introduction

Optimum detection requires the use of some form of matched filtering. In conventional narrowband HF digital systems using FSK and PSK, simple frequency filters and phase discriminators are used to determine whether or not a MARK or a SPACE was transmitted. The decision is based upon whether or not the filter outputs are larger or smaller than a particular threshold (Chapter 3). In Chapter 3 it was described how this process was enhanced by various forms of diversity reception and error-control coding. However, the system is very prone to errors caused by signal distortion and other effects due to various ionospheric phenomena such as multipath propagation. This is due to the fact that the receiving elements merely detect the presence, or absence, of a particular frequency or phase - in which sense they are matched - and do nothing to enhance the particular frequency or phase.

This is not the case in systems using pulse compression techniques. In these the matched receiving filter adds gain to the particular signal, if it is present, no matter how corrupted it is by white gaussian noise or other additive distortion. This is due to the fact that, unlike the signal pulse of FSK or PSK, the transmitted pulse in the pulse compression system is itself coded.

Pulse compression is most widely used at present in radar systems where the most common form of code is linear FM - known familiarly as chirp. The use of chirped pulse compression techniques in digital data transmission was first suggested by Winkler who recognized that there could be definite advantages to be gained over the conventional methods.

The chirp technique involves frequency modulating each transmitted signal element to give a linear change of frequency with time along the pulse. Each received signal element is operated on by a matched
filter, which combines the signal spectral components coherently into a narrower signal of increased amplitude. This pulse compression operation may typically give 20dB peak gain.

In digital systems the MARK and SPACE signals are defined by chirps of opposite slopes (frequency sweeps) and are transmitted in the same frequency band. At the receiver two matched filters are used, one for the MARK signal, the other for the SPACE. Conventional interfering signals such as white gaussian noise and the carrier do not contain the frequency code needed for time compression and pass through the matched filter essentially unchanged in amplitude. Therefore it is expected that, at the output of either matched filter, the signal peak will generally be well defined under interference conditions.

By superposition, compressed multipath chirp-signal components will appear without significant mutual interference at the matched filter output, provided that the differential time delay between paths is greater than the time duration of a single compressed output pulse. This time duration is given approximately by the reciprocal of the transmitted signal frequency sweep, or bandwidth. The chirp technique thus offers inherent protection against signal fading due to multipath propagation. This property is exactly equivalent to the ability of chirp radars to detect multiple target returns discretely.

The chirp-signal repetition frequency must, of course, be significantly greater than the longest multipath time differential delay. This is a fundamental restriction on non-adaptive systems using serial data formats in dispersive channels, and restricts the keying-rate in the HF band to a maximum of about 100 baud as described previously.

This chapter will thus begin with a discussion of matched filters and the various methods by which they are implemented. Particular attention will be given to recent techniques such as the use of digital
devices (DDS), charge-coupled devices (CCDs) and surface-acoustic-wave (SAW) devices to implement the matched filters, since it is predicted that it is with these technologies that the future improvements in HF digital communication will come. The use of pulse compression in HF digital data systems is then introduced and a possible way in which the new technologies can be used to implement the pulse compression method in the HF field is described.
4.2 Matched Filtering Techniques

Matched filtering is a form of signal processing which is the optimum for arbitrary signals corrupted by white gaussian noise. It operates to transform the raw input signal into a form suitable for optimum detection decisions. As discussed in Chapter 2, the matched filtering concept refers specifically to digital signal processing.

Matched filters (MFs) are thus always specified with respect to given waveforms which have both finite length (T) and bandwidth (B). The product of these quantities - the time bandwidth (TB) product - is one of the most important parameters in MF performance.

In the frequency domain, the MF transfer function is equal to the complex conjugate of the signal spectrum multiplied by an appropriate linear phase factor which depends on the time delay through the filter. In the time domain, the MF impulse response is simply the time reverse of the desired signal, delayed to satisfy causality requirements.

When driven by its specified input signal, the MF produces the autocorrelation function of this particular waveform. This autocorrelation is characterised by a single large compressed pulse whose half-width is approximately equal to the reciprocal of B. In addition there are smaller subsidiary peaks - the sidelobes - which are placed symmetrically about the main lobe for a time duration ± T.

MFs therefore operate to produce pulse compression\(^1,2\) such that the input-pulse TB product is equal to that of the output pulse. The result of this is coherent processing gain, the amount of which is dependent upon the magnitude of the TB product. This, combined with the fact that MFs operate only upon signals specified in their design, makes them useful for a variety of fundamental signal processing functions. Among these are frequency filtering, spectrum analysis, code division
Basic MF operation is shown in figure (4.2.1) and its theory is well documented\textsuperscript{1,2}. The main emphasis in this investigation will thus be with the practical implementation of the MF concept.

The principles of matched filtering have been known for many years\textsuperscript{3}, but it is only recently that suitable implementation techniques have become available to enable practical MFs to enjoy widespread use in signal processing systems.

There are three techniques currently available for MF implementation which have particular significance in this discussion. Each of these techniques (DDs, CCDs and SAW) has advantages and disadvantages in terms of processing time, available bandwidth and other characteristics peculiar to the particular technology. In general however DDs are most efficient for very low bandwidths and long time delays, CCDs cover the important intermediate regions including audio through low MHz frequencies and SAW devices are useful for wideband short time duration operations. Figure (4.2.2) shows the approximate operating ranges for MFs based on these three technologies. Here it can be seen that SAW and CCD techniques are complementary and do not compete (overlap) except at the high-speed end of the CCD performance range. The same conclusions may be deduced for SAW and DD. CCD and DD implementations do however compete over a wide range representative of the many important MF applications mentioned previously. The choice between CCD and DD will thus generally be based upon other system considerations besides bandwidth and processing time. For instance, at high data rates the power requirements of DDs far outweigh those for CCDs which are also generally smaller in size. Again, time delays in excess of one second will normally require the use of digital techniques as shown in the diagram.
**Fig. (4.2.1) Matched Filter Operation**

**Input**
- Mark signal compressed in mark matched filter.

**Space**
- Linear FM downchirp
- Space signal dispersed in Mark matched filter

**Output**
- Matched Filter 1: Upchirp (Mark) downchirp (Space)
- Matched Filter 2: Mark matched filter (downchirp)

**Amplitude (Frequency)**

**Block Diagram**
FIG. (4.2.2) APPROXIMATE OPERATING SIZES FOR CCD's, SAW DEVICES AND DIGITAL DEVICES IN TERMS OF BANDWIDTH AND TIME DELAY.
Despite this competition, CCD and DD techniques are also capable of complementary operation. It may be noted here that the relative costs of these three techniques also play an important part in deciding which to use.

Because the three technologies can complement each other it is reasonable to expect that they may be used in combination. This would enable their advantages to accrue with the result that many complex signal processing operations may become possible.\(^8,9\)

In this section therefore, MF implementation using the three technologies will be discussed and the suitability of each with respect to HF digital communication will be considered.
4.2.1 Digital Matched Filtering

Digital signal processing devices are fabricated using basic components which perform simple functions. Among these are the shift register, the memory, the arithmetic and the logic circuits, commonly used in digital circuit design. Digital Matched Filters (DMFs) made from these fundamental blocks offer accuracy and versatility in system design. They can asynchronously process the strings of digital data which represent the signals and do not suffer from the usual problems of analogue filters - ageing and instability. They are however a type of transversal filter, and as such are constrained by the sampling structure implementation common to such filters. The specific limitations consist of quantization error of samples and coefficients, and estimation errors on intermediate results. DMF performance is usually measured in terms of dynamic range, determined by the number of basic components used in design, and of accuracy, defined by the number of samples required to produce a specific response. These two characteristics decide the complexity of the DMF and in conjunction with the type of semiconductor technology used (CMOS, ECL, I^2L, LSI), determine the processing speed, size, power consumption, and cost of device.

The basic operation of the DMF as conventionally implemented may be described with reference to figure (4.2.3). This particular filter uses baseband input data and produces baseband output data. The information is converted to digital format in the analogue-to-digital (A/D) convertor. The required dynamic range (SNR) determines the number of binary digits in this conversion. This information is then stored as analogue data bits in a serial input/parallel output shift register which has each of its N analogue bits composed of several binary digits. The parallel outputs of this register feed N separate multipliers whose
**FIG. (4.2.3) CONVENTIONAL DIGITAL TRANSVERSAL FILTER ARCHITECTURE**
other ports are driven from a reference shift register holding the desired filter impulse response. The outputs from the multipliers then feed a single N-input adder whose output is reconverted to an analogue signal by a digital-to-analogue (D/A) converter.

This conventional approach is normally complex and costly. For example a 100-bit PN sequence correlator (DMF) with $2^8$ levels can currently be implemented as above, in TTL for a cost of about £4,000. It would include 400 TTL networks and an A/D convertor. However more recent techniques have improved this situation somewhat. For example, implementation of the above correlator using large scale integration (LSI) techniques with a different architecture reduces significantly the size and fabrication difficulty of the device even though the cost is very much the same. This implementation is shown in figures (4.2.4) (a) and (b). In this case, the A/D convertor drives M LSI correlator chips (figure (4.2.4a)), one chip for each significant binary digit in each analogue data bit. As shown in part (b) of this figure, each correlator chip consists of 2 64-bit binary shift registers with independent clock lines, 64 exclusive OR circuits for providing the necessary binary correlation, 64 D/A current sources and a common sum line to provide an analogue correlation output for a pair of binary input signals. The design and implementation of DMFs using LSI technology is fully discussed by Turin. In his paper, Turin sees LSI techniques as making digital devices competitive with the alternative technology of CCD and also sees further improvement in performance as the technology advances.

In addition, with the advent of LSI techniques, the microprocessor is now a powerful tool which is seen as the future if not present, building block of a new generation of digital signal processing devices which will have the powerful capability of programmability in addition to the
FIG (1.2.4)a. IMPLEMENTATION OF DIGITAL CORRELATOR USING LSI CORRELATOR CHIPS.

FIG (1.2.4)b. BLOCK DIAGRAM OF ONE CHIP.
already proven advantages of using digital devices in communication systems.
4.2.2 CCD Matched Filtering

The CCD, though more complex and monolithic in construction, is similar in implementation and operation to the more familiar bucket brigade device (BBD). It can be fabricated using any standard MOS or bipolar process, and like the BBD, may be treated operationally in the generic class of Charge Transfer Devices (CTDs). The CCD moreover has performance advantages compared to the BBD and is thus expected to dominate future CTD signal processing applications.

Because of their inherently analogue qualities, CCDs are ideally suited to a large number of sampled data signal processing functions. These have been described as digital so far in this investigation, but as seen in section (4.2.2) the functions are often analogue, and CCDs thus offer the advantageous possibility of performing many sampled-data filtering functions directly in the analogue domain. CCDs may thus be regarded as combining the best features of digital and analogue techniques. Like digital filters, CCDs are controlled by a master clock and the CCD filter characteristic is as stable as the master oscillator. However, the requirement for A/D conversion is eliminated and all functions are performed in the analogue domain.

Despite these apparent advantages, CCDs have performance limitations when compared to digital devices precisely because they are analogue devices. Foremost among these limitations are representation accuracy of both the signal and reference function and a limitation on the lowest operating frequency - this latter being due to thermal dark currents in the semiconductor which degrade the signal with length of delay time. Accordingly the main advantage of the CCD filter over the DMF is generally taken to be reduced cost of implementation. Additional advantages include lighter weight, smaller size, lower power consumption and
improved reliability. CCDs are thus not expected to make DEs obsolete, (especially with the advent of LSI and the microprocessor); but in applications where the requirement is for modest performance and sufficient volume so that low cost is the principal design goal, they are expected to have a large impact. 17

The CCD MF is a special case of the CCD transversal filter 5 just as the DMF was a special case of the digital transversal filter. A block diagram of a typical transversal filter is shown in figure (4.2.5). It consists of N delay stages D. The filter output is obtained by multiplying voltages $V_n$, $n = 0 \rightarrow N-1$ at each intermediate node by weighting coefficients $h_n$, $n = 0 \rightarrow N-1$ and summing. The CCD transversal filter is quite easy to implement 17 and its computational power is impressive, particularly when compared to an equivalent digital implementation. For example, a CCD filter having $N = 500$ stages clocked at $f_c = 1$ MHz performs the equivalent of 500 8-bit multiply-and-add operations each microsecond. The transfer function of a transversal filter may be written as 17

$$
U(z) = \sum_{n=0}^{N-1} h_n z^{-n}
$$

which describes a finite impulse response (FIR) 18 filter. This is a special case of a more general discrete time filter in which there is no feedback. The transfer function has only zeros (no poles) and its impulse response is therefore finite.

The multiply-and-add function described in figure (4.2.5) may be simply implemented by way of the split electrode technique illustrated in figure (4.2.6). Here, as charge transfers under an electrode, deposited on a silicon substrate, an opposite charge must flow from the clock line on to the electrode. On one phase, all electrodes are split into two sections of varying area. The phase selected is not
FIG (4.2.5) BLOCK DIAGRAM OF A TYPICAL TRANSVERSAL FILTER
Fig. (4.2.6) Split Electrode Multiply-And-Add Implementation.
important; and in the figure the $\beta_3$ electrode is split. One side of each of the split electrodes is connected to the $\beta_3$ clock line and the other to the $\beta_3$ clock line. These two lines are clocked simultaneously and in phase (compare figure (4.2.4b)). Each portion of the split electrode receives some signal dependent charge which is proportional to the area of that electrode section and to the signal flowing on to that electrode. The charge differential in the two sections of a split electrode is then measured, and this performs the non-destructive sampling and weighting operations. Since the $\beta_3$ electrodes, as well as the $\beta_3$, are tied together the summation occurs automatically. At each clock period, the output signal is proportional to the difference in charge required by the two lines of split phase $-\beta_3$ and $\beta_3$. The filter output is now obtained by integrating the difference currents flowing in the $\beta_3$ and $\beta_3$ clock lines using a differential current integrator (DCI).

The CCD transversal filters discussed above are inherently performance limited owing to:

1. Imperfect charge transfer efficiency (CTE)\(^{17}\) which modifies the transfer function of the ideal CCD. Its effect is to shift the frequency response of the ideal theoretical filter used as the basis for practical CCD implementation. Imperfect CTE thus introduces dispersion which limits the bandwidth when $N$, the number of serial delay stages (electrodes), is large. The amount of dispersion which can be tolerated depends on the application, but if it is specified that the dispersion must be less than $N_E = 1$ ($E = 0$ for an ideal CCD), and if $E = 0.5 \times 10^{-4}$, then up to $N = 2000$ serial delay stages can be implemented. Since $TB = N/2$ for a CCD filter\(^{17}\), this means that $TB$ products of up to $10^3$ can be achieved as shown in figure (4.2.2).
2. Weighting coefficient errors due to inaccurate mask production. This limits the precision with which a given filter response can be realized.

3. Noise, which limits the ultimate dynamic range attainable with CCDs. The CCD noise level is determined predominantly by the DCI noise figure and at present 75–80dB of dynamic range \(^1\) (P-P signal with acceptable distortion/rms noise level) is obtainable.

4. Linearity which is limited by imperfect sampling of the input signal voltage and by the nonlinear depletion layer capacitance of the device.

5. Peripheral MOS circuitry used for clock drivers and output amplifiers limits the maximum clock frequency, \(f_2\), to be < 20KHz. This in turn limits the practical CCD bandwidth to values below 10 MHz. (\(\frac{1}{2}\) max clock frequency).

6. Thermal leakage which limits the maximum delay achievable with a CCD. If it is specified that thermal leakage must be less than 1% of the total CCD charge capacity, then CCD delay time is limited to \(T < 1\) sec. \(^1\)

Because of the above performance limitations, CCDs will probably never totally replace DDs, but in some areas, because of those advantages they do have, their use will become more widespread especially in sampled data processing of analogue signals.
4.2.3 SAW Matched Filtering

A Surface-Acoustic-Wave (SAW) device consists of an interdigital metallization deposited on the polished surface of an axially oriented cut of piezoelectric crystal substrate. The SAW device fabrication process uses the standard photolithographic and etching techniques of the semiconductor industry. A simple interdigital structure is shown in figure (4.2.7). When an alternating voltage is applied to the metallization, at the bus-bars, a strain, oscillating at the frequency of the excitation voltage, develops between the interdigital fingers. This alternating strain is produced by virtue of the piezoelectric effect and launches a Rayleigh surface-wave front that travels in both directions from an origin at the centre of the interdigital transducer (IDT). The wave thus exists on the crystal surface as an electro-acoustic vibration in its primary mode. Other modes are also excited at levels below the surface, but these are severely attenuated with increasing depth into the substrate.

The SAW thus retains the frequency of the electromagnetic signal applied to the IDT. However since it propagates acoustically along the surface of a substrate, its velocity is some five orders of magnitude less than that of the excitation voltage. This means that a large number of wavelengths can be accommodated on a short length of crystal and hence a relatively large delay time is possible for a relatively small piece of hardware.

In a typical layout, two IDTs are deposited on a crystal substrate as in figure (4.2.8). The excitation of one IDT with a source of centre frequency, \( f_0 \), generates an SAW front at the centre of this IDT— the launching IDT.

The SAW propagates in both directions. An absorbing material placed on the crystal dampens the energy travelling to the left. The
FIG. (4.17) SIMPLE INTERDIGITAL STRUCTURE.

FIG. (4.28) TYPICAL SAW DEVICE LAYOUT.
SAW propagating to the right traverses a distance, \( L \), corresponding to a specified delay, across the crystal surface before passing under the recovery IDT. As a result of this passage, the electric potential difference - a part of the SAW - excites this transducer (TDI) and an energy transfer takes place from the SAW to the IDT which then drives the load, \( Z_L \).

The energy transfers occur because the centre-to-centre spacing of the TDR fingers is designed to be one-half wavelength (\( \lambda / 2 \)) at the device centre frequency, as shown in figure (4.2.9). Adjacent fingers are driven by opposite polarities. The associated changes of the resulting strain wave are generated by distortion of the piezoelectric dipoles on the crystal surface.

The delay properties of the SAW device enables the realization of practical transversal filters in much the same way as described in sections (4.2.2) and (4.2.3). The performance and response characteristics of these filters are critically dependent upon IDT design, and the range of SAW filter applications in signal processing is determined by the degrees of freedom available in the design of the IDT. Fortunately the IDT is a very versatile structure as may be shown by investigating its impulse response characteristics.\(^{21}\)

When an IDT is impulsed, the resulting field in each inter-electrode gap produces an elemental acoustic source whose polarity is determined by the sense of the field. The duration of the resulting SAW is then directly proportional to the electrode (finger) periodicity (determined by the placement of the fingers) and its amplitude may be controlled by varying the amount of finger overlap (aperture) - a process known as apodization.\(^{21}\) There is thus a one-to-one correspondence between the impulse response and the geometry of the IDT. The result is that for an \( N \)-electrode TDR, there are \( 2N \) degrees of design freedom - the \( N \)-electrode positions and the \( N \)-electrode lengths.
Fig. (4.2.9) Surface wave energy transfer process
Because of the flexibility of the IDT, an SAW device can be designed for almost any signal type. This is particularly useful for the implementation of matched filters which are usually specified with reference to a given waveform.

SAW matched filters are presently implemented in two different ways. The first method finds application in classic pulse compression systems and uses two Dispersive Delay Lines (DDLs)\(^1\),\(^2\),\(^3\). In these devices one TDR is designed to vary linearly with frequency along its length and thus has an impulse response characterized by linear FM or chirp\(^1\),\(^2\). In a practical system, one filter is impulsed to produce a pulse of chirp. The other filter which is designed to have a response of opposite slope to its matched counterpart acts on this pulse to produce its autocorrelation function in the form of a compressed pulse. This process thus gives the SNR enhancement characteristic of MF reception techniques. SAW MFs are now widely accepted and used in modern radar systems\(^4\). They are also used in implementing the chirp z transform technique\(^5\) which enables a highly efficient method of implementing real-time fast fourier transformation; a particularly useful artifice in real time spectrum analysis.

The second method used Tapped Delay Lines (TDLs) to perform the implementation of various codes used in communication systems. The principal application for TDLs has been for Analogue Matched Filters (AMFs) to produce bi-phase or PSK waveforms\(^6\),\(^7\). The waveform is generated by an input TDR having an impulse response equal to the chip (code bit) waveform, and an array of broadband taps spaced on delay intervals equal to the chip length. Coding is determined by the polarity of the tap connections. An AMF which produces the autocorrelation of the PSK waveform can be implemented by simply placing the input TDR at the opposite end of the tap array to time reverse its
response. AMFs with up to 255 taps, chip rates (bandwidths) of up to 20MHz and a length of 60µs are available at present. The full potential of the AMF is realized when the TDL is integrated with tap-switching circuits which then allows the coding to be varied electronically. These devices, called programmable Analogue Matched Filters (PAMFs) are still in the development stage but show great promise for the future.

SAW device performance is limited by several factors among which the most important are

1. practical substrate size;
2. the resolution capabilities of device fabrication techniques;
3. insertion loss (IL), and
4. various second order effects as a result of SAW device design and fabrication.

In practice, the maximum obtainable lengths of piezoelectric crystal substrates are limited to a few tens of centimetres; (ie) ST-X cut quartz may be obtained up to about 20 cm long. This limits the maximum TDR length possible in a practical device and is also the reason why ideal infinite impulse response must be truncated to realize practical SAW devices. The main consequences of this limitation with respect to achievable designs using two (or three) TDRs on a single surface of a given substrate are

1. to limit the minimum achievable centre frequencies to about 10MHz,
2. to limit the minimum bandwidth to about 50kHz, and
3. to limit the maximum delay to about 50µs on ST-X cut quartz.

The resolution capabilities of SAW device fabrication techniques limit primarily the maximum achievable centre frequencies to about 1GHz. This corresponds to a spacing between electrodes of just under
one micron (1μm), which is just possible; (eg) (using an Electromask machine). This also limits the absolute accuracy with which TDRs can be fabricated for any centre frequency since each space between fingers must be made up of an integral number of approximately 1μm steps using this machine.

Insertion loss IL is not really a limitation to SAW device usefulness since it may be recovered in a system by the use of appropriate amplifiers. However large insertion losses (>40dB) are not happily tolerated since they increase system size, complexity and cost. More importantly they may degrade the dynamic range and thus the SNR of a system and must thus be avoided where possible. In the SAW devices relevant to this investigation (DDLs and TDLs) IL is usually quite large (about 30 to 40dB) and is a key trade-off parameter in optimum design,\textsuperscript{30,31} affecting as it does the bandwidth, VSWR and Triple Transit (TT)\textsuperscript{31} level of the SAW device. Insertion loss comes from many sources\textsuperscript{21} the most important of which are due to electrical mismatch (in the networks used to couple the device to other system components) and to the bidirectional propagation of the SAW. Bidirectionality loss for typical 2-TDR SAW devices is 6dB\textsuperscript{21}; electrical mismatch loss\textsuperscript{21,31} is dependent upon the specified levels of TT, VSWR and the amount of bandwidth as a fraction of the centre frequency - the SAW device fractional bandwidth. Other losses are due to

1. parasitic loss in the TDR pattern,
2. losses in matching network components,
3. propagation losses in the substrate,
4. losses due to beam spreadings, and
5. apodization losses.

These seldom exceed 2 to 3dB (collectively) when care is taken in device design and fabrication.
Careful design also minimizes the limitations caused by second-order effects due to SAW propagation under metal TDRs deposited on a crystal substrate. Most of these effects distort, or tend to distort, the passband with superimposed ripple and other spurious and must be catered for at the design stage. The procedure for dealing with them are more or less standard at present and are well cataloged in reference 32.

The SAW matched filter is thus a very useful processing device at IF frequencies and has great advantages in flexibility, cost, reproducibility, ruggedness and size over more conventional IF filters in use at present (eg) (lumped element, bulk wave). If carefully designed and used properly it has obvious application in many future communication systems - a fact reflected by the numerous review articles and conferences concerning SAW devices in the past few years.
4.3 Pulse Compression and HF Digital Data Transmission

At the beginning of this chapter it was demonstrated that Pulse Compression (PC) systems would have advantages over the narrowband systems conventionally used in HF digital communications. It now remains to discuss how best to implement PC techniques in HF systems.

PC depends on the ability to generate some form of coded transmitted signal in pulsed form for each bit of digital data. In radar systems, where PC has been used for years, the most popular code is linear FM or chirp. This may also be used to advantage in HF digital systems.

Chirp signals may be generated using a saw-tooth pattern (corresponding to the data input) to modulate a voltage controlled oscillator (VCO). Alternatively a dispersive network may be impulsed, remembering that the HF impulse response is the time reverse of the matched signal. Both of these methods will produce a signal pulse of finite duration and amplitude containing a waveform whose frequency varies linearly along the length of the pulse. If the frequency decreases from beginning to end of the pulse then a downchirp is produced; if it increases an upchirp is generated. The duration of the chirp in the former case will be equal to the duration of the saw-tooth pattern, and in the latter case to the impulse response duration of the dispersive network. A typical chirp waveform is shown in figure (4.3.1). Its defining parameters are $T$, its duration as defined above; $B$, its bandwidth determined by the frequency sweep of the saw-tooth waveform of the frequency characteristics of the dispersive network; and $CF$, its centre frequency determined by the VCO range of the centre frequency of the dispersive network. The product $TB$, encountered previously as an MF performance parameter, is known as the signal dispersion factor of the system.

Signal parameters depend on the system application. For instance,
Fig (4.3.1) TYPICAL CHIRP WAVEFORM WITH DEFining PERIMETERS
in a long-range HF data transmission system, the 2.5kHz BW of the standard voice channel might be the most practical value for the frequency sweep or the dispersive network bandwidth, B. For a dispersion factor of 50, the signal duration T would be 20ms. This would give protection against intersymbol interference due to multipath propagation, the largest differential time delay on a North Atlantic HF band route being typically less than 5ms.

In an HF chirp system described by Gott et al., a VCO was used, modulated by a saw-tooth pattern. Chirp signals were generated at a centre frequency of 100kHz and down-converted to 5kHz at the receiver for compression. An alternative mark/space signal format was used with a signal dispersion factor \( T_B \) of 50.

At the receiver identical dispersive networks were used for both MARK and SPACE pulse compression, the received signals being frequency reversed in one branch. This was achieved by correct sideband selection after mixing by way of MARK and SPACE branch mixer oscillators separated in frequency by twice the CF of the dispersive network.

The dispersive network's output envelopes were then sampled over the duration of the compressed pulse using receiver synchronization.

A Gaussian filter with 20dB band edge attenuation relative to the CF was used to weight the received chirp signals before compression. Weighting is an important aspect of chirped radar systems since it minimizes the compressed pulse time sidelobes which might otherwise obscure secondary targets. In communication systems, its advantages are less obvious since it operates to widen the main compressed lobe which will then be smaller in amplitude for a given signal energy. Weighting also affects interfering white noise but not in the same way as it affects the signal. In this system this had the effect of reducing the SNR by about 1dB. However in the case where two adjacent
signals were the same, the advantages of weighting come to the fore. For example, two MARK signals occurring sequentially will be compressed as expected in the MARK dispersive network. At the SPACE dispersive network output, the two dispersed signals will overlap for a time equal to the input signal duration $T$. Without weighting these signals will add to give a composite dispersed signal whose maximum will be $6\,\text{dB}$ greater than the output due to a single MARK signal passed through the SPACE dispersive network. With Gaussian weighting the maximum was reduced by $4.4\,\text{dB}$ to $1.6\,\text{dB}$.

Another advantage of Gaussian weighting is the inherent phase linearity of such filters. If sharp cut-off filters are used without phase equalization, the band-edge components of the chirp signal will not effectively contribute to the coherent signal build-up in the dispersive network, thereby reducing the peak output.

The dispersive networks used in this system consisted of lumped element (capacitors and inductors) all pass bridged T-sections which have a time delay vs frequency response. The series combination of many appropriate bridged T-sections can give a composite time delay vs frequency response which is linear across the finite bandwidth. Each network in the system used 46 of these bridged T-sections. This type of lumped element configuration was used in order to effectively produce low-frequency chirp signals. However, low frequency pulse compression has one great drawback. This is that the number of cycles contained within the compressed signal may be so small that an arbitrary phase angle may cause the envelope peak to be poorly defined. This problem may be solved by carrying out the pulse compression process at a higher frequency, which will also enable the use of more efficient chirp signal compressors.

The system down converts to about $5\,\text{kHz}$ at the receiver before
pulse compression. At this point the signal may be compressed along
the time axis using a CCD serial register with different input and
output data rates. Samples of the received chirp signal are clocked
into the register at the Nyquist rate or greater. The input clock is
then stopped when the whole signal has been read and the output clock
is started. If the output clock rate is 1000 times as fast as the
input clock rate, then the signal will be compressed by 1000:1 along
the time axis. Assuming the parameters already discussed, a chirp
signal of duration 20ms with a B of 2.5kHz and a CF of 5kHz would
be compressed into a signal 20μs long with bandwidth 2.5MHz and a CF of
5MHz. If this signal is now upconverted to a CF of about 25MHz, it
will become compatible with the range of frequencies where SAW devices
can be used. This means that a SAW DDL with the following definition
can be used as a compressor in place of the low frequency lumped element
dispersive network described above.

\[ \text{CF} = 25\text{MHz} \]
\[ \text{BW} = 2.5\text{kHz} \]
\[ T_{\text{delay}} = 20\text{μs} \]

This device will have a TB product of 50 which is easily realizable
using conventional techniques as described in section (4.2).

The advantages of using this higher frequency compressor are well
documented in the reference literature \(^6\),\(^24\),\(^33\),\(^34\) and need not be
detailed here. Of course a DD time compressor could be used instead
of the CCD serial register. The device will depend on the particular
design being considered.

Another example which serves to reinforce the technique of combining
CCD/DD and SAW technologies and which may also find application as a
high performance HF scanning receiver is the pulse-doppler processor
described by Roberts et al.\(^39\)
This system is used to process 'slow' Doppler radar returns between 0 and 2MHz which occur in 25ms bursts. A CCD serial-to-parallel arrangement is used to provide a 1000:1 time compression rate between input and output data rates. As before the 25ms burst of Doppler is compressed into a 25μs time slot which occupies a BW of 2MHz, making the data SAW compatible.

The basic SAW analyzer described by these workers has been developed into a sophisticated SAW spectrum analyzer by the Author et al. This analyzer has a real-time processing delay of 50μs, 40dB dynamic range and the ability to discriminate between approaching and receding targets.

As a high performance scanning receiver for the HF band, the SAW module would be used on its own to accept HF carrier frequency inputs and to display their frequency and amplitude content for identification and monitoring purposes. At the moment BW is limited to 2MHz for the system described above. However modified versions of the unit exhibit 7MHz and research is being carried out to examine the feasibility of producing an analyzer with 25MHz bandwidth (which could cover practically the whole of the HF band) using SAW technology.
CHAPTER 5

PROPOSAL FOR AN HF SPREAD-SPECTRUM SYSTEM
5.1 Introduction

It has now been established that the new technologies can be used with advantage to implement matched filter communication links at high frequencies (3 - 30MHz). In recent years, one of the most exciting exploitations of the new matched filter designs has been in the practical realization of efficient and reliable spread-spectrum communication systems.

As their name implies, these systems employ a form of wideband modulation to spread the basic message 'bit bandwidth' over a much larger bandwidth. The ratio of the expanded bandwidth to the original message bandwidth may be expressed as the matched filter processing gain. This implies that in such a system there will be the possibility of saving transmitter power as well as an ability to receive signals reliably in conditions of very low signal-to-noise ratios.

At first sight, spreading the HF signal bandwidth may appear contrary to the general bandwidth conservatism of the HF system philosophy. However the spread-spectrum modulation waveform is in fact a long code which may be altered dynamically through digital programming. Each individual code word then accesses a specific receiver through a front-end matched filter designed to recognize only one of the many separate codes. All other code combinations are non-correlatable in this matched filter and so the idea of a shared spectrum usage by many simultaneous communicators becomes a practical possibility.

This, plus the promised saving in power, enhanced detection capability, and the inherent immunity of the matched filter to multipath and other ionospheric interference makes the spread-spectrum technique a very attractive proposition to the designer and user of the HF band.

The chapter begins by considering the basics of the spread-spectrum...
system. This system is then compared for optimality with the optimum conventional digital system described in chapter 3. Finally a proposal for a novel spread-spectrum system for particular use in the HP band is presented and discussed.
5.2 Spread-Spectrum Communications

Spread-spectrum techniques are used to permit the communication of intelligible messages under difficult channel conditions. Normally they are used to counteract the effects of jamming or to conceal the presence of a communication signal. These uses suggest the potential of such a system to provide

1. resistance to interference,
2. tolerance of multipath signals, and
3. message detection in low SNRs.

This last feature is the principal feature of spread-spectrum modulation techniques.

Spread-spectrum systems are capable of operating at low SNRs, typically -10 to -30dB, because the spread-spectrum modulated signal sent from one transmitter can be made distinct from all other signals in the band, including interference. Since over a time period, $T_1$, $2B_1$, numbers characterize the signal, the dimensionality of the signal can be increased by increasing $B$, the bandwidth of the signal. Spread-spectrum systems use a bandwidth that is wide compared to the bandwidth that would normally be required to communicate the message information. $B$ is increased by coding the signal in a complex fashion such that the code can be recognized and removed only by the intended receiver.

A well-designed spread-spectrum system theoretically achieves a signal processing gain equal to $B_c/B_m$, the ratio of the transmitted bandwidth ($B_c$) to the message bandwidth ($B_m$). This implies that a spread-spectrum receiver can operate in SNRs that are worse by a factor $B_c/B_m$ than those required for transmitting the message in a bandwidth $B_m$. This performance is not achieved without disadvantages however.

The most noteworthy of these are:
1. Increased spectrum-occupancy is in direct proportion to the signal processing gain, compensated in part by the shared use of a common spectrum.

2. Complex signal processing equipment is required in addition to the components of the conventional receiver.

3. The receiver code must be synchronized with the received code. Fortunately however the use of the new technologies discussed in Chapter 4 can partially overcome these disadvantages. This is because they permit the implementation of very complex functions in very small spaces without the need for manual adjustment.

The operation of spread-spectrum systems and various possible configurations have been fully discussed elsewhere and only a brief resume will be given here. They can be implemented in a number of forms involving frequency hopping and/or phase coding, but basic to the system is that the data, or bits, to be transmitted must be multiplied by a suitable fast code known as the transmitter signature. A biphase pseudo-noise sequence is often used which spreads the information over a wide band and subdivides one bit into many 'chips'. Due to common channel usage and other interfering effects the signal is usually buried in noise. A multiply-and-integrate technique must therefore be used for demodulation at the receiver, but only after synchronism has been established.

The multiply-and-integrate procedure provides processing gain for the desired signal and so rejects noise and other interfering signals not processing the relevant fast code format. Similarly unless the code is known, it is difficult to produce the available compression gain and the signal remains practically indetectable. This provides privacy or alternatively prevents any one signal from interfering with
a receiver for which it was not intended. Synchronism is the key to multipath signal rejection since it allows the receiver phase to be adjusted for only the direct signal delay.

In the conventional serial search and integrate receiver, the demodulator code generator is shifted past the code of the received signal. A correlation can only be recognized during a period of time longer than 1/\(B_m\), the reciprocal of the message rate. For a time uncertainty of \(T_u\), there are \(T_u \cdot B_c\) possible synchronism points, each one of which must be tested for a time 1/\(B_m\). The maximum synchronism time, \(T_s\) is thus proportional to

\[
T_s \propto \left(\frac{B_c}{B_m}\right)T_u
\] (5.2.1)

For large values of \(T_u\) and large processing gains, the search time can be excessive (in the order of seconds). A short repetitious code (perhaps 1,023 chips long) may be used but has poor protection against some forms of interference, since interference can then be made to look like the repetitious code. The short code should be no shorter than ten times the processing gain, so for large processing gains and low data rates, as may be required at HF, several hundred seconds may be needed for initial synchronism. Moreover, should synchronism be lost a new search is required. Also it is extremely difficult to provide for 'interrupt' operation since any transmitter must be synchronized so that the signal arrives at the demodulator with the correct timing.

This problem is resolvable by the use of a matched filter as the synchronizing element. This enables a simultaneous search over a large code interval, permitting the receiver to use a much faster search rate. The 'lock-up' time may thus be reduced by a factor equal to the number of chips in the filter, (ie) the TB product or comparison gain of the matched filter. The use of this technique may
also be extended to demodulation of the data. The length of the filter required for this depends on the SNR of the system. This is usually defined in dB/Hz - the SNR in 1 Hz bandwidth of the receiver. Typically this will be in the range 50dB/Hz to 30dB/Hz which will respectively require matched filter periods of 40μs to 4 ms to attain 6dB SNR at the output.

The matched filters used to provide the synchronism at realistic 'lock-up' times may be implemented in any of the technologies discussed in chapter 4.

5.2.1 SAW Technology

With SAW devices, there are two methods of achieving rapid synchronization. In the first, the codeword is stored in the pattern of the IDT deposited on the piezoelectric substrate. This pattern may be either fixed or programmable. In the transmitter, the SAW device is impulsed at its centre frequency.

In the receiver the signal is filtered by an identical SAW device before non-coherent detection. Such devices are capable of codewords approximately 1 Kbit long using single devices and 10 Kbit long using cascaded devices.

In the second, a SAW convolver is used. The codeword is now stored in a binary feedback shift register or a semiconductor memory. The inputs to the convolver are the codeword and the signal. Synchronization occurs in the active region of the convolver and its detection is used to activate the main pseudo-noise codeword generator. Synchronization codewords approximately 1 Kbit long are achievable. A recirculation loop can be included with the convolver to increase the processing gain and length of the codeword. Since SAW convolvers are very lossy and require high input signal levels ($\lambda$), this method is unsuitable for very
compact equipment. However, performance for ground based equipment is very good.

5.2.2 **CCD Technology**

With CCDs the codeword is stored by tap weighting. Samples of the signal are read into the device and weighted by the stored codeword. Synchronism occurs when agreement with the stored codeword is achieved and this activates the main message sequence generator. As noted previously, CCD processing is normally carried out at baseband. Since phase-locked carrier references are not available, detection must be implemented prior to synchronism using in-phase and quadrature (I and Q) channel detection of the synchronizing sequence.

Codeword lengths of approximately 1 Kbit are presently attainable but this should increase in the future with advances in the technology. At present, the codeword may be increased by using a second unweighted CCD in conjunction with the synchronism CCD.

The advantages of the CCD implementation include small size and low power consumption.

5.2.3 **Digital Device Technology**

Synchronism detection can be implemented with digital devices in several different ways. In one of the simplest, the codeword is stored in a multivalued shift register where each 1-bit is represented by the largest positive binary number and each 0-bit by the largest negative binary number. The signal is sampled at the spread-spectrum chip rate, quantized, and read into a parallel shift register. The weighted sum of the respective shift register outputs is formed and synchronism detection is indicated when the weighted sum exceeds a
predetermined threshold. This threshold is determined by the probability of detecting the fast code contaminated with interference and noise. Since the fast code is normally used to PSK- or FSK-modulate the individual bits, this probability is given by the error probability equations given in Chapter 2 for these techniques. Once synchronism is achieved, the message sequence generator can be activated.

For the same reasons as above in the CCD case, I and Q channel detection must be used and dual correlation channels formed with the overall correlation output obtained by adding the weighted-sum outputs of each channel.

One of the main disadvantages of the above system, as is indeed the case with most digital systems, is the large amount of power required by the TTL used in the digital matched filter implementation. However, new digital processes (TTL, ECL and MOS) are fast improving the situation.

Besides the matched filters used as described above, there are other signal processing functions which, though not unique to spread-spectrum systems, can be more easily implemented with the new technologies in the spread-spectrum case. These include

1. Fast switching, frequency synthesizers,
2. Bandpass filters, and
3. Timing and control circuits.

Spread-spectrum communication systems are now considered to be viable both economically and practically because of the advances in the new technologies. This implies that application of these techniques in HF communication systems may have some advantages. These must now be more fully investigated.
5.3 Spread-Spectrum Potential in HF Communication

5.3.1 Introduction

The main question to be considered here is whether spread-spectrum techniques can be used to advantage in the design and operation of HF communication links. The brief unqualified answer to this is 'yes.' To substantiate this it is necessary to compare the conventional HF system with the spread-spectrum system discussed above.

In Chapter 3 it was concluded that the digital HF system was optimum in certain defined senses when compared to the HF analogue system. For consistency, the comparison proposed here must be based on the criteria defined in Chapter 3. It is also important to assume at this point that a practical spread-spectrum system can be implemented at HF. This assumption will be fully considered in the next section where the problems associated with, and the techniques required for the realization of the HF spread-spectrum communication system will be discussed. It is pertinent to consider first some general aspects of both systems.

In the conventional digital HF system, each communication channel is limited to a narrow bandwidth of about 3kHz by international agreement. This is done to protect the transmission of other simultaneous users of the HF channel. This protection is needed because the messages of every user are transmitted in the same type of code - FSK, PSK or DPSK. Conventional receiver front-ends discriminate between carrier frequencies but not message codes. This means that interchannel interference would greatly reduce the reception efficiency in unprotected channels and would make intelligible reception extremely difficult, if not impossible.
The bandwidth limitation is a severe restriction to the HF communicator especially when the multipath-cum-probabilistic nature of the ionosphere is considered. The disadvantages of narrowband communication in multipath are well-known and have already been discussed (Chapter 3) and need not be repeated here.

In spread-spectrum communications, on the other hand, a receiver is activated only when it locks-on to its own particular code, and does not respond to other codes. On the reception of this code, the message decoder is enabled to recover the transmitted message. The use of matched filters, inherent in this system, enables messages to be recovered at very low SNRs.

Ideally this means that the HF band may be used at any time by any user as long as enough codes, distinct from each other, were available to implement the necessary distinct receivers. Put another way, the 3 to 30MHz band could become the communicating property of all HF communicators who would then be able to operate at a variety of carriers at all times in a wideband message mode with a reasonable expectation of establishing reliable interference-free communication.

5.3.2 Comparison

There are several advantages and some disadvantages in the system just proposed. A comparison of the two systems based on the criteria defined in chapter 3 will help to pinpoint these.

1. **Signal Quality** - The spread-spectrum system is inherently immune to the multipath problems encountered with the conventional HF system. With proper receiver synchronism, made possible by the use of matched filters, the receiver phase is adjusted only to the direct signal delay. Moreover, the use of chirped signals for the individual coded bits
(Chapter 4) will be an added aid to reducing the effects of multipath. In addition to this, the use of matched filters, both at the front-end and at the decoding stage, will enable messages to be received in very low channel SNR. This will obviously act to reduce greatly the effects of interference and noise due to natural ionospheric phenomena.

From the standpoint of signal quality, therefore, the spread-spectrum system will have all the advantages of the conventional HF digital system plus those due to the use of spread-spectrum techniques.

2. Reliability -

With better signal quality and immunity to multipath, it is reasonable to expect that the spread-spectrum system will be inherently more reliable, once synchronism is established, than the conventional digital system. In addition to this, since signals can be received in much lower SNR, the power margin in the given link may be reduced thus improving the reliability of the hardware used for communication.

3. Complexity -

From the description of the basic system given in section (5.2), the spread-spectrum will be much more complex than its conventional counterpart. At the present state-of-the-art this point may not be argued. However with the expected advances in the technology upon which spread spectrum systems are based, it is not too difficult to envisage a
time in the near future when spread-spectrum systems may be no more difficult to implement than present conventional digital systems. Of course it may be argued that these latter may be correspondingly simpler due to technological advances of a similar nature. However if the superior performance of the spread-spectrum system is maintained, then the more complex system may be more practically acceptable than the simpler but inferior system. Indeed this decision is often made today in choosing between analogue and digital communication systems when reliability and signal quality are important.

As in the case of system complexity, the cost of the spread-spectrum system will at present be considerably greater than that of the conventional system. However for the same reasons given above this cost will greatly decrease in the near future. Besides this it is probable that for similar performance the spread-spectrum system will be able to operate at much lower power levels than the conventional digital HF system. Since power is a major cost consideration in most HF communication links, the saving with the spread-spectrum should go some way to compensate for the extra cost involved in greater complexity.

To achieve its superior wideband performance, the spread-spectrum system depends on an efficient use of a shared communication band. As discussed
in the introduction to this section, this simultaneous use of a common spectrum is possible because of the use of a distinct code for each user in the channel.

Ideally this would provide unlimited access to all carrier frequencies in the band to all users as long as enough distinct codes could be generated to accommodate each individual communicator.

The generation of sufficient codes is presently one of the main factors preventing the achievement of the truly random access wideband channel. However with technological advancement, the situation is expected to improve rapidly in the near future.

At the moment the number of users practicable in a shared spectrum is limited. Since this spectrum is greater in bandwidth than the conventional system, there is presently no advantage in bandwidth usage with a spread-spectrum system rather than a conventional digital system.

However when other advantages are taken into account, it may be argued that with a spread-spectrum system, the bandwidth available is being used more efficiently than with the conventional system.

Modern spread-spectrum systems are in their infancy and they utilize technology which is ever-advancing. In terms of future investment, their potential is very great and certainly superior to that of the conventional system. It may be stated
that spread-spectrum systems will employ most, if not all, of the technology used by conventional systems, now and in the future, in addition to the technology needed to implement the spread-spectrum concepts. The opposite is not true.

From the above, spread-spectrum systems are seen to have several advantages over the conventional HF communication system. Indeed except for the present cost and complexity of the spread-spectrum system, it seems conclusive that their use may be highly recommended in most communication links.

Before this can be stated categorically, the practical implementation of a spread-spectrum system at HF must first be investigated.
5.4 Proposal for HF Spread-Spectrum System Implementation

5.4.1 System General

As stated in section (5.2), conventional HF digital communication is generally confined to very narrow message bandwidths to protect the transmissions of simultaneous channel users. The intention with a spread-spectrum link is to effect this protection by using specific signature codes produced by spread-spectrum modulation of the digitally encoded message. The relevant question now is how to determine the message bandwidth of an HF spread-spectrum communication link. It is reasonable to assume that this must be set by the frequencies that can be supported by the Ionosphere at any instant - (ie) by the available HF carrier frequencies between 3 to 30MHz.

At the high-frequency end, using an SSB transmitter with the lower sideband on a carrier at 30MHz, there could conceivably be some 27MHz of usable bandwidth. However at the low-frequency end, there would only be 2 to 3MHz of bandwidth. It will therefore be deducible that a suitable standard spread-spectrum bandwidth, usable with all carriers will be about 2.5MHz.

At present an HF channel teletype terminal operates at a 50 baud rate into a 3kHz voice channel for FSK carrier modulation. It is also recommended that a single channel digital data link be operated at no more than 100 baud for reliable operation. This enables a code bit length of 10 ms or more in presently used modulators.

Normally several of these low-speed, low-bandwidth data channels are multiplexed and sent on a single carrier using modems such as Kineplex. This is a multi-channel phase modulated system which can accommodate up to 40 channels at a speed of 75 baud/channel in the normal 3kHz voice channel. This produces an information rate of 3 Kbit
This is thought to be the fastest presently available with HF digital data modems.

For compatibility and easy interfacing with these single-channel encoders, the proposed spread-spectrum system must be designed to cope with the basic information rates and bit lengths discussed above. In other words, it must be at least comparable in capacity and speed with a system like Kinesplex.

Furthermore, the HF spread-spectrum facility may be required to modulate basic channels using the chirped bit described in chapter 4. Each bit in these channels occupies a bandwidth of approximately 2.5kHz to 5kHz in a 10 ms to 20 ms time slot. These signals have a high immunity to multipath and interference due firstly to their chirped nature and secondly to the matched filter system in which they operate. It may be noted here that, in a very basic sense, use of these signals gives a second level of spread-spectrum protection since they can only be properly received by the appropriate matched filter. In fact systems based on the chirped bit have been shown experimentally to give superior performance to the conventional digital system. It is thus reasonable to expect that when the chirp signal is further protected by the fast spread-spectrum code signature, it will perform even more reliably in multipath and other ionospheric interference.

Since the basic chirp occupies between 2.5kHz and 5kHz of bandwidth and has a time duration between 10 ms and 20 ms, the spread-spectrum modulator must be capable of spreading these bandwidths over the 2.5MHz spread-spectrum bandwidth derived in previous discussion.

As will be seen later, the spread-spectrum modulator will consist of a fixed bandwidth SAW of CCD device which will be signature programmable. This implies that to cope with variable-bit bandwidths from the single channel encoder, some form of signal parameter equal-
ization will be needed between the encoder and the spread-spectrum modulator.

This equalization is possible with some form of programmable bandwidth expansion technique implemented with CCDs or digital devices. With either technology expansion ratios of up to 1000:1 are readily available. Besides equalization, this facility will reduce the bandwidth spreading demands of the spread-spectrum modulator. However, since the system performance is measured by the ratio of the spread-spectrum bandwidth to the encoded signal bandwidth at the modulator input, some form of compromise must be reached concerning the proper amount of bandwidth expansion used in any given application. It may also be noted that bandwidth expansion is equivalent to time compression (chapter 4). Hence a 2.5kHz signal occupying 20 ms will be transformed to a 250kHz signal in 200μs with 100:1 time compression. It may also be noted here that this operation will also convert the CF of the signal to a SAW compatible centre frequency.

After the spread-spectrum modulator, the signal may be used to modulate the appropriate carrier from an SSB transmitter and after amplification can then be transmitted. Reception is primarily the opposite of the transmission process except that a given receiver will only accept messages modulated with the spread-spectrum signature particular to its front-end matched filter. This characteristic gives the spread-spectrum system its great advantage over the conventional digital system in that signals buried in noise are retrievable because of the processing gain available in the spread-spectrum demodulator matched filter unique to a given receiver.

A block diagram of the system discussed above is given in figure 5.4.1. Here the large boxes depict the basic blocks of the system. As discussed above, these boxes may be implemented in a variety of ways.
SYSTEM FOR H.T.
GENERALIZED SPECIFIED SPECIFICATIONS

Diagram of system components involving user, decoder, equalizer, modem, equalizer, demux, equalizer, modem, and user.
some of which are combinable to form the complete system. The useful combinations should be clear from the text. In the figure, another equalizer has been included after the spread-spectrum modulator. This gives the option of reducing the bandwidth of the modulated signal before transmission. In general, this should be unnecessary, but if there is a need, the option will be available.

The most important sections of the spread-spectrum system are the modulator and demodulator - the spread-spectrum modem - and this must be discussed in more detail.

5.4.2 The Spread-Spectrum Modem

Wideband spread-spectrum modulation may be effected in any of three ways:

1. Direct Pseudo-Noise (PN)
2. Frequency Hopping (FH)
3. Time Hopping (TH)

Specific combinations of the properties of these allow an optimum trade-off between the desirable characteristics of the spread-spectrum technique.

1. Co-channel multiple access
2. Channel utilization
3. Low detectability by unauthorized users
4. Jamming immunity
5. Multipath resolution
6. Range and velocity measurement.

Success in achieving the above functions depends on the receiver. Here the incoming signal is cross-correlated, with integration over each data bit, against the known spreading waveform of the transmitter. When accurate synchronism exists between the transmitter and receiver, data
recovery is possible despite the linearly superimposed multiple access interference and thermal noise. Spread-spectrum demodulation can be carried out either with an active demodulator or a matched filter. Identical peak output SNR results at the end of the data bit duration ($T$). The ratio of peak output to average input SNR defines the processing gain, which for stationary band-limited white noise equals $2B_sT$ where $B_s$ is the spread-spectrum signal bandwidth.

PN sequences are fundamentally important to the practical implementation of the spread-spectrum system. This is due to the fact that they may be used directly to spread-spectrum modulate the encoded signal, as well as to control the other two techniques mentioned above - PN and TH. PN sequences have many of the properties of random numbers but they are in fact entirely predictable. This predictability enables both a receiver and transmitter to generate the same sequence actively and in synchronism thus allowing the receiver to decode the transmitted message.

The shift register used in the code generation is of fundamental importance since PN codes are distinguished by their repeat length which is fixed by the shift register capacity.

SAW devices have several basic properties which make them a logical choice for application as shift registers. These include high information capacity, low power consumption per bit of stored information and potential high operating speed.

A SAW tapped delay line (TDL) is equivalent to a shift register operating at a fixed clock rate defined by the bit length and the acoustic velocity (ie) the SAW device has a fixed bit rate or bandwidth. The tap positions for a particular code are chosen simply according to the code generating rules (ie) the device is signature programmable.
Fortunately many PN codes can be accessed with a relatively small number of tapping points on the line. This reduces the spurious signal levels caused by tap reflections in the delay path and simplifies the fabrication requirements.

PN sequences can also be generated using CCDs and digital devices. However, for the same bandwidth capability the SAW device has the potential of producing many more distinct addresses than these devices. Hence these other technologies are useful at low bit rates when the user density is not very great.12

Typical SAW device characteristics for the spread-spectrum modem may be given as:

- Centre frequencies 10 MHz to 30 MHz
- Bandwidths 2.5 MHz to 5 MHz
- Spread-spectrum fast code duration up to 50µs

For the particular application described, a centre frequency of about 20 MHz with a 2.5 MHz bandwidth occupying 20µs would be adequate. This would allow a code length of 200 bits specified by taps spaced by 100 ns. Performance equivalent to that reported by several workers2,3,12,13 would be expected to give adequate performance to a large number of users.

There is a large amount of data on these modems (used at higher frequencies) reported in the literature already referenced and it is not intended to go into the subject more deeply here.
5.5 Discussion

From the foregoing treatment, it may be suggested that spread spectrum techniques have much to offer the HF communicator. There are difficulties in implementing the shared channel concept but nowhere has it been suggested that these are insurmountable.

Most of the work in spread-spectrum has been done at frequencies above those in the HF band, but as this investigation has attempted to show a programme of work to fully investigate the use of spread-spectrum in the HF band would be very profitable, and could go a long way to the realization of an HF system which was very efficient in power consumption and reliable in continuous reception being, as it would be, relatively insensitive to multipath and other ionospheric interference.

Such a system would incorporate all the features proven in this study to be optimum as regards signal quality in the HF channel. It would also form the basis for providing much more widespread and varied use of the large amounts of effort and capital invested in HF communications over the past fifty years.

This investigation will now be concluded since it has usefully gone as far as possible, at present, into a full examination of the HF communication system and the signal quality associated with it. Several important area have been discussed and some important recommendations for improving the HF communication facility. Finally it is thought that this brief look at the requirements of an HF spread-spectrum system features high in the list of priority recommendations.
CHAPTER 6

SUMMARY AND CONCLUSIONS
6.1 Summary

The aim of this investigation was to produce design principles for HF communication systems in which signal quality was optimised. To achieve this, a general study was carried out to determine which aspects of conventional HF communication systems most merited consideration with respect to optimum signal quality. At this early stage, signal quality was defined as the degree of absolute correspondence between the received signal and that transmitted. Later, it was specifically defined with reference to the Signal-to-Noise Ratio (SNR) at relevant parts of the system.

HF system designers are faced with two major types of constraints regarding signal quality. Firstly, there are those due to the ionosphere. In practice, these are not under the engineer's control and must be compensated for at the design stage. The state-of-the-art in the technology is the second type of constraint.

It may be argued that conventional HF system design is very nearly optimized and that the main performance improvements over the last ten years, say, have come from the availability of more efficient devices for use in the manufacture of the standard HF hardware.

Assuming these conditions, it was demonstrated in Chapter 3 that HF digital communication systems, particularly those based on some form of biphase coded waveform (PSK, DPSK) offered the best method of obtaining optimum signal quality. This deduction was based on a comparison between both analogue and digital HF systems. The points of comparison, or criteria, were derived from the initial studies on ionospheric radio propagation and were discussed in Chapter 3.

These systems and their performance in a Rayleigh fading multipath channel were then examined to establish if and where possible improvements
could be made in their design. The Rayleigh model was chosen as being the closest statistical (and thus predictable) model to the actual ionosphere. This study indicated that major improvements might be possible in the implementation of the matched filtering techniques used at the moment in the conventional HF communication systems.

At this stage work was being done in designing and developing a radar pulse compression sub-system based on Surface Acoustic Wave (SAW) devices. This work also involved, but to a lesser extent, the use of digital devices and charge-coupled devices (CCDs).

Essentially, conventional HF digital systems use dual-frequency signalling methods. These signals are very prone to degradation from frequency selective fading and other related phenomena often found in multipath channels such as the ionosphere. The pulse compression technique largely overcomes this problem by the use of the chirp signal as described in Chapter 4. Because the chirp contains a band of frequencies rather than just a single frequency, it is much less prone to frequency selective fading. Also because a given pulse compression filter will only effectively receive the signal transmitted from its matched counterpart, noise and other additive spurious which do not have the required compression code ultimately have much less effect on the received signal than in the conventional HF digital system.

Gott et al (Chapter 4) had demonstrated the use of lumped element matched filters at HF and had shown that performance was indeed improved. However low frequency pulse compression, as described by Gott, has one great drawback. Because only a few cycles are used in the compression, arbitrary phase angles cause the compressed peak to be poorly defined. Better results are attainable at higher frequencies and more efficient matched filters can be used. These would be based on the SAWs, CCDs and digital devices.
Matched filters using these technologies were then discussed and then applied to the design of an HF pulse compression system. A direct use of the radar sub-system developed was also described at the end of Chapter 4.

SAW filters operate at IFs between 10 MHz and 1 GHz. These are normally too high for HF communication; however by using CCD time compression it is possible to bring the relatively low IFs and small bandwidths used at HF up to the frequencies where best performance can be obtained from the SAW filters. This time compression technique is considered to be a basic design tool in the use of SAW devices in HF system design.

The use of the new technologies was taken a step further in Chapter 5 where the techniques of spread spectrum communications were used to propose the design of an HF spread spectrum system. The exciting aspect of this system is the possibility of using the HF channel as a shared spectrum (no channelization) with relatively large message bandwidths (approximately 2.5 MHz). Users of the channel would be distinguished by a fast spread spectrum code which could be recognized only by a receiver containing the relevant decoder. The effect on other transmissions would be minimal. Ultimately it is hoped that this type of design will revolutionize the HF ionospheric channel and make it a much more reliable and efficient channel than it is at present.

Of course there are problems to be solved before a system like this could be practically implemented. Among these the dispersive multipath and dynamic fading characteristics of the ionosphere must be considered of primary importance. A study of these will help to determine several parameters essential to the proper functioning of the proposed system. Some of these are:-
1. The spread spectrum bandwidth which will be the most resistant to ionospheric distortion.
2. The best coding for the basic message bit.
3. The best range of centre frequencies (carriers) to be used.
4. The amount of intersymbol interference to be expected and also how much could be tolerated before the channel became unusable.

Another important area for study will be the inter-user interference which may occur between the spread spectrum coded signals in the proposed common channel. The results of this will set limits to the number of simultaneous users before the channel becomes too corrupted for reliable communication.

These studies will also indicate the need for techniques such as diversity and adaptivity in an operational system to enable reliable communication under all ionospheric conditions.
6.2 Conclusions

The aims of this investigation were to understand and overcome by any technological means the fundamental weaknesses of HF ionospheric radio communication as they have been experienced over the past fifty years. The basic parameter used in this study was signal quality defined with respect to the universal quality variable in communication systems – the Signal-to-Noise Ratio (SNR).

It is considered that these aims have been achieved both qualitatively and quantitatively in theory and design. In terms of design implementation it is admitted that not much has been done except for the development of a radar subsystem based on SAW devices which was shown to have application in HF receiver design. However the practical bases for these designs do exist and they do operate at frequencies above the HF band. Their adaptation for operation at HF is not considered difficult. In addition to this the cost of one-off pairs of SAW matched filters was prohibitive in the context of this investigation. For these reasons it was adjudged sufficient to only propose the design objectives.

The findings of this investigation suggest that there are definite improvements to be gained in applying the new technologies of SAW, CCD and digital devices to HF system design. In this way, it is very probable that the advantages of the HF channel may be more fully exploited than they are at present.

The concept of ionospheric radio propagation is simple to implement practically compared to satellites and undersea cables. It is also relatively cheaper and, probably most important of all, its strategic importance is undeniable. Satellites and undersea cables are obvious military targets at times of unrest. Should they be destroyed, then the only fallback position for long-distance radio communication would be the ionosphere, and HF radio communication.
APPENDIX I

SUMMARY OF IONOSPHERIC RADIO PROPAGATION FUNDAMENTALS
The main properties of the earth’s atmosphere in the ionospheric region are shown in Figure (Al.1,1) where it may be noted that the molecular density decreases rapidly with height.

The D, E and F layers also all vary in altitude and density according to the solar cycles (daily, annually and 11-year), but these changes do not have the same sense in the different layers.

The composition of the three layers is not very different from that of the atmosphere near the earth’s surface. The oxygen \( \text{O}_2 \) and nitrogen \( \text{N}_2 \) molecules are progressively replaced by their respective atoms at increasing altitudes and there is also a small amount of NO due to the combination of nitrogen and oxygen. This small proportion of NO however plays a very important part in ionization processes in the upper atmosphere due to its small ionization potential of 9.4eV.
Fig. (A1.1.1) MAIN UPPER ATMOSPHERE CHARACTERISTICS.
AI.1.1 The D-Region

This, the lowest ionospheric region, is usually defined by the layers between 60 and 100 Km. In some instances, the layers below 64 Km, down to about 50 Km, are designated by a C-region due to the noticeable differences between these layers and those above 64 Km.

The main characteristics of the D-region may be summarized as:

1. A high molecular concentration of the order of $10^{20}$ molecules $\cdot m^{-3}$.

2. A small electron concentration of between $10^8$ and $10^{10}$ electrons $\cdot m^{-3}$.

3. A very high collision frequency between electrons and molecules, in the range $5 \times 10^5$ and $5 \times 10^6$ $m^{-3} \cdot sec^{-1}$.

This region is therefore highly absorbent to radio waves.

4. A high combination coefficient of about $10^{-12}$ to $10^{-13} m^2 \cdot sec^{-1}$.

The equation $T = \left(\frac{2}{3}\alpha\right) \times N_{\text{max}}$, where $\alpha$ is the recombination coefficient and $N_{\text{max}}$ is the maximum value of electron density in a given layer, gives the D-region time constant $T_D$ a value of about 60 seconds. This implies that this region rapidly follows variations in solar illumination.

The above indicates that the D-region is mainly an absorber of radio waves. The attenuation it produces is one of the key variables used to evaluate the quality of a radio signal in the ionosphere and as such will be discussed in a separate section. (AI.1.2)

---

AI.1.2 The E-Region

This region may be treated as if composed of three layers between 100 Km and 150 Km.

The first of these is made up of the so-called sporadic E or $E_s$ 'clouds'. These occur at low and average heights of about 100 Km.
They are about 2 Km thick, are discontinuous, and suffer displacement at great velocity. They are also capable of reflecting HF radio waves. No-one has yet established the exact nature of the temporary 'clouds'. It is generally assumed that they are strongly ionized clouds in the lowest part of the E-region. This theory is supported by rocket soundings indicating the existence of ionized 'clouds' corresponding to $E_2^2$. The $E_2$ 'layer' is usually continuous in the vicinity of the magnetic poles and often shields the higher layers because of its heavy absorption.

The second layer is the $E_1$-layer. This is the best example of the ideal Chapman layer. Its thickness varies between 20 and 40 Km and maximum ionization is found at approximately 120 Km. The $N_{max}$ exceeds $10^{11} \text{m}^{-3}$ at noon; (eg) it can attain $2.10^{11} \text{m}^{-3}$ when the sun is at its zenith which corresponds to a critical frequency of 4MHz. This critical frequency, $f_{OE}$, is linked with solar activity by the following equation, for a given location at noon.

$$(f_{OE_1})^2 = A_n (1 + 0.004R) \quad (AI.1.1)$$

where $f_{OE_1} =$ critical frequency of $E_1$-layer

$A_n =$ coefficient corresponding to the number $n$ of the month

$R =$ sunspot number

The following expression may also be used as an approximation which is valid for any time of day and season.

$$f_{OE} = 0.9((810 + 1.44R) \cos X)^{0.25} \quad (AI.1.2)$$

where $X$ is the sun's zenith angle.

Hence during the day the $E_1$-layer is capable of reflecting radio waves up to its critical frequency, and of absorbing those that pass completely through it.

At night a certain amount of ionization persists due to imperfect
ion-recombination and to meteoric ionization, but for E is low -0.25MHz during solar minimum and 0.5MHz during solar maximum. This latter corresponds to an Ne of about $10^{19} \text{m}^{-3}$. The E₁-layer therefore does not reflect normal HF radio waves at night but as its collision frequency is still quite large, between $5 \times 10^{3} \text{m}^{-3} \text{sec}^{-1}$ and $2 \times 10^{4} \text{m}^{-3} \text{sec}^{-1}$ it does absorb radio waves transmitted at night. Despite this small ionization persistence, the E₁-layer recombination coefficient is quite large at about $5 \times 10^{-4} \text{m}^{3} \text{sec}^{-1}$ and so the layer rapidly follows variations in solar illumination.

The E₁-layer often splits into two layers giving rise to the third E-region layer - the E₂-layer at about 140 Km and with slightly higher ionization. Its basic properties are however similar to those of the lower E₁-layer.

Because of their large recombination coefficients, the D and E layers rapidly, and almost completely, disappear at dusk leaving the region of main importance to HF communication - the F-region and in particular the F₂-layer.

### AI.1.3 The F-Region

This region extends from about 100 Km to 450 Km and consists of two layers with very different behaviour patterns.

The F₁-layer occurs at a height of about 200 Km and like the E₁-layer is a Chapman layer. It obeys rather exactly the Chapman formula

$$N_{\text{max}} = \left( \frac{q_q \cos X \omega}{\alpha} \right)^{\frac{1}{2}}$$

where $q_q$ is the maximum number of electrons per unit volume in the layer as a function of X. $N_{\text{max}}$ is more simply given by

$$N_{\text{max}} = 2.5 \times 10^{11} \left( 1 + 0.0062R \right)$$
The $F_1$-layer critical frequency is given for any time of the day and any season by

$$foF_1 = (4.3 \cdot 0.01 R) \cos 0.02 \lambda$$

(A.I.1.5)

In this layer absorption is low, the recombination coefficient is approximately $10^{-15} \cdot m^3 \cdot sec^{-1}$ and $\lambda$ is of the order of four minutes.

The second section of this region is the $F_2$-layer which appears at about 300 km. Its behaviour is totally different from that of any other region or layer.

In this layer Chapman's formula is completely invalid and an alternative theory due to Martyn is often used to explain the $F_2$-region phenomena. Here it is assumed that a displacement of ionized plasmas takes place from one point in the ionosphere to another under the influence of:

1. wind - when the transport of the ionized particles is almost parallel to the earth's magnetic field,
2. electric fields - which are due to changes, and which in combination with the earth's magnetic field, submit the electrical particles to electromagnetic forces,
3. gravity,
4. partial pressure of the various components of the atmosphere.

The critical frequency $foF_2$ varies directly with solar activity but no satisfactory formula exists to correlate the two variables. Generally however $foF_2$ is higher by day than by night (with a lag on the Sun's zenith angle).

Unlike any other layer, the $F_2$-layer exhibits a seasonal anomaly with $foF_2$ being usually higher in winter than it is in summer and an annual anomaly, when $foF_2$ is higher for the layer as a whole in the summer of the northern hemisphere. Since these two anomalies are in phase in the northern hemisphere, the variations there are particularly
pronounced. In the southern hemisphere they are in antiphase, which
gives rise to a biannual anomaly. All explanations to these phenomena
are still hypothetical.

Nevertheless, for a given station and month, foF₂ at noon is
found to obey the relation

\[(\text{foF}_2)^2 = (1 + 0.02\alpha)\]  \hspace{1cm} (AI.1.6)

and Martyn \(^8\) has given the recombination coefficient of the F₂-layer
as \(\beta\) in

\[\beta = 10^{-4} e^{(300 - h/50)}\]  \hspace{1cm} (AI.1.7)

Also some general conclusions may be made concerning foF₂ from
ggeomagnetic coordinate charts drawn up during the solar maximum of
1947 and the solar minimum of 1943-44. These are:

1. (foF₂) \(_{\text{min}}\) occurs around 0600h or slightly earlier.
2. The highest values of foF₂ are usually found at about
   1500 or 1600h and not at noon.
3. The geomagnetic equator corresponds to (foF₂) minimum values.
4. foF₂ varies considerably from day to day.
5. F₂-layer electron densities can be as high as \(3 \times 10^{12} \text{ m}^{-3}\).

Finally, table AI.1 shows the altitudes at which maximum ionization
may be expected in the F₂-layer for different times of the year.
### Table Al.1

<table>
<thead>
<tr>
<th></th>
<th>Great or average latitudes</th>
<th>Magnetic Equator</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Noon</td>
<td>Midnight</td>
</tr>
<tr>
<td>R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equinox</td>
<td>220</td>
<td>210</td>
</tr>
<tr>
<td>Winter</td>
<td>250</td>
<td>240</td>
</tr>
</tbody>
</table>
All heights in kilometers.
Al.2 Predicted Ionospheric Parameters

The ionosphere is produced and maintained by solar radiation whose intensity at any given point varies diurnally, seasonally and in accordance with an approximate 11-year cycle. Experimental observation for many years has shown that the definitive parameters of the ionosphere vary in much the same way. Ionospheric predictions are thus largely based upon the forecasting of these variations in solar activity; relating them to geophysical changes in the earth's upper atmosphere and thence to the relevant ionospheric parameters useful, indeed indispensable, to the design of efficient radio communication links.

Predictions are published by a number of services who use complex computer-based empirical methods to analyse the large amounts of data collected at the many measuring stations scattered over the earth's surface. The analysis of the data produces the information needed by the radio engineer to both plan new HF radio links and also to maintain reliable communication over existing channels.

There are three methods in operation at present for ionospheric study. These are:

1. The ionosonde;
2. Incoherent scatter radar;
3. Rockets and Satellites.

The ionosonde has the advantage of supplying continuous information at any observation point about

1. the number of existing layers,
2. their virtual heights,
3. their critical frequencies.

Incoherent scatter is an expensive technique, but it provides accurate data concerning

1. the electron density profile of the ionosphere and so
the same information as the ionosonde,
2. the composition of the layers,
3. the temperature in the ionosphere,
4. the ion and electron drift velocity.

These facilities make it useful in the wider field of detailed ionospheric investigation.

Both of the above methods fail, however, to accurately inform about the characteristics of the D-region due to the low values of the D-region critical frequencies. Rockets and satellites which transmit signals to earth through the ionosphere find use in this context.

The ionospheric parameters of main interest to the HF circuit designer are divided into two classes. These are based on
1. short-term predictions,
2. long-term predictions.

For short-term forecasting, the specified future time is taken to be a few hours to a few days in advance and involves mainly the forecasting of variations in solar activity.

The more general predictions are long-term and give monthly median values of the MUF (sec. (A1.3.1)), and absorption data. These factors are derived from statistical data about the ionosphere and are published 3 to 6 months in advance.

Solar activity is indicated by various indices, the oldest of which is $R$, the Zurich sunspot number, which is often given by

$$ R = K \cdot (10g + f) $$  \hspace{1cm} (A1.2.1)

where $g$ is the number of sunspot groups, $f$ is the number of observable individual spots and $K$ is a correction factor $\approx 1$, used to equalize the results from different observers and equipment.

$R$ varies from year to year according to a cycle whose duration is about 11 years. The values of sunspot number at the maxima are
spread from 46 in 1816 to 175 in 1957 and those at minima from 0 in 1810 to 11 in 1766. The period of growth (average 4 1/2 years) is usually much shorter than the period of decline (average 6 1/2 years).

Naturally, prediction of R is difficult. It is usually quite accurate during the periods of increasing and decreasing solar activity, but the determination of the period and of the values of maxima and minima remain uncertain.

This problem has had much discussion particularly by Halley and Gervaise, who have proposed a new method. They prove that the mean times between minima are given by,

\[ t' = 11.15r + 1774.31 \] (Al.2.2) (a)

and those of the maxima by,

\[ T' = 11.08r + 1749.42 \] (Al.2.2) (b)

where r is the number of the cycle. They then consider a secondary cycle of 170 years, which results in a correction to the mean times in order to obtain times of a defined circle.

Most parameters used in ionospheric study are linked rather simply and accurately to R, which is thus established as the most important variable for producing ionospheric predictions. Predictions of R, three months in advance, are made by the NBS in Washington.

R is subject to wide variations from month to month because of the discontinuities of solar activity. The characteristics of the ionosphere do not accurately follow these variations. In order to ensure better correlation, a sliding 12-month average is generally used for R. This is defined by

\[ R_{12} = 1/12 \{ \frac{1}{3}(R_{n-6} + R_{n+6}) + \sum_{m=7}^{n-5} R_{m} \} \] (Al.2.3)

where n is the number of the month. The value of \( R_{12} \) is by definition only known 7 months after the last recorded observation.
Two other related indices bear mention. These are IF \_2^{13} and \( \phi \).

IF \_2 is published six months in advance by RSMR at Slough and enables prediction of foF\_2 at noon. \( \phi \) is a measure of the radio electric emission from the sun at 2000 MHz. It is expressed in \( 10^{-22} \text{W.m}^{-2}.\text{Hz}^{-1} \) and appears capable of medium term prediction (6 - 9 months) to greater accuracy than either R or IF \_2. It is however related to R\_1.2 by \[ \phi = R\_1.2 + 46 + 23e^{-0.05R\_2} \] (A1.2.4)

As will be seen in the following sections, R plays quite a big part in the evaluation of some of the main factors affecting signal quality in the ionosphere.

Besides R, the other parameters which are usually predicted are,

1. The ionization at each point in the ionosphere, and hence the expected level of absorption, and
2. The monthly median values of the MUF.

The first of these predictions is based partly on theoretical and partly on experimental methods for deducing the probable ionization in the various layers at each point in the ionosphere at any time using the index R.

The second is deduced from the ionization expected at each point of the globe at a given time.

These ionization values have in turn been derived from observations during previous years by stations which sound the ionosphere. It can therefore be seen that ionospheric predictions of this type are not very accurate and at best serve as general guide-lines for the initial stages of designing new links and of day-to-day operation of existing radio links.

The short term predictions are usually more useful for day to day running of a link in any reliable way; and so most continuous
measurements are usually passed on to operating links as soon as possible to enable them to make the relevant adjustments for efficient communication.

Besides monitoring the ionosphere, some stations also keep a watch on the Sun's activity since a sudden change in solar action could cause serious degrading of transmission quality. This will be elaborated upon in section (A1.7)
Al.3 Basic Ionospheric Radio Propagation Theory

Long range HF communication is possible because under certain conditions, the ionosphere is capable of reflecting radio waves with frequencies between 3 and 30 MHz. The reflection is in fact progressive refraction caused by the ionized layers (section Al.1) traversed by the wave in its passage from transmitter to receiver. This refraction is controlled by a set of optically derived laws which are fundamental to the realization of HF radio communication links.

To a first approximation these laws relate to plane ionospheric layers parallel to a flat earth; and the process of progressive refraction resulting in complete reflection is depicted in figure (Al.3.1) for this simplified treatment.

In practice, two major corrections must be made to this basic theory. The first is the inclusion of the effect due to the curvature of both the earth and ionosphere and the second is the correction for the effect of the earth's magnetic field.

It is proposed here to outline the first order theory of ionospheric radio propagation and to briefly mention the secondary effects associated with the corrections applied in practice. Full discussions of this theory may be found in several excellent texts on HF radio propagation 6, 21, 25.
FIG. (A1.3.1) PROGRESSIVE REFRACTION
PROCESS IN IONOSPHERE
Al.3.1 Flat Earth/Ionosphere, No Magnetic Field

To a first approximation, the ionosphere may be described as a stratified medium consisting of a large number of thin strips parallel to a flat earth. Each strip exhibits an ionization density, \( N_m^{-3} \), and a critical frequency, \( f_c \), in MHz related by

\[
f_c = 9n^2 \quad \text{(Al.3.1)}
\]

\( f_c \) increases as \( N \) increases with altitude. Each strip also has a refractive index, \( n \), given by

\[
n = \left(1 - \frac{f_c^2}{f^2}\right)^{1/2} \quad \text{(Al.3.2)}
\]

where \( f \) is the frequency of the incident wave and absorption is neglected. \( n \) decreases with increase in height and it may be proved that a radio wave incident on the ionosphere is completely reflected if \( f < f_c \).

Assuming that a ray is incident upon the ionosphere making an angle \( i \) with the vertical, then Snell's law gives

\[
n \sin i = \text{constant} = \sin i_0 \quad \text{(Al.3.3)}
\]

for a medium having a continuously varying refractive index. The angle of the ray with the vertical before it entered the ionosphere is given by \( i_0 \). If the ray becomes horizontal at any point, then at this point \( \sin i = 1 \), and so from equations (Al.3.2) and (Al.3.3)

\[
f = f_c \sec i_0 \quad \text{(Al.3.4)}
\]

This implies that at a point where the plasma frequency is \( f \), an ionized layer can reflect waves of a frequency higher than \( f_c \) so long as they are incident at an angle and not vertically. This statement is known as Martyn's law and gives the maximum frequency for complete reflection in the ionosphere wherever \( N_{\text{max}} \), the maximum electron density, is known. This frequency is the Maximum Usable Frequency (MUF) - used exhaustively in the design and operation of HF communications links. As the discussion so far assumes a flat earth and ionosphere, equation (Al.3.4) is only approximately true for
distances up to 1000 Km, beyond which corrections must be applied as will be seen later.

The equations above produce the path ABFDE in figure (A1.3.2). This is the real path with a section BFD in the ionosphere and a real height of reflection $h$. The lowest point of the ionosphere is denoted by $h_0$ in the figure. If now the tangents to the path at B and D are drawn to meet at point C, the path ABCDE is obtained. This represents the path taken by a beam of light undergoing mirror reflection from a plane surface. This reflection takes place at a height, $h$, which is defined as the virtual height of reflection. If the path ABCDE is used in radio wave analysis, it is clear that the evaluation of the ionospheric path becomes quite simple, since exact analogies may be taken from the accurately documented science of optics. Two important theorems make this type of analysis possible.

Breit and Tuve$^{26}$ have proved that the time required for transmission along ABFDE is the same as the time taken by the path ABCDE at the speed of light. This means that for a flat ionosphere, the reflection process is equivalent to optical reflection at the virtual height.

Secondly, Martyn$^{17}$ has proved that the virtual height of reflection of an obliquely incident wave is the same as that of the equivalent vertical wave. Thus if $f$ and $f'$ represent the frequencies of waves reflected vertically and obliquely from the same real height, then the vertical signal is equal to the height of the equivalent triangular path ABCDE for the oblique signal.

These theorems are valid for any vertical distribution of ionization, and they permit the replacement of the real ray by a vertical ray that travels in a medium with $n = 1$, and is optically reflected from a plane situated at the vertical height.
FIG. (A1.3.2) REAL AND VERTICAL PATH GEOMETRY
In practice, this simplified theory is modified by the corrections mentioned before for the effects of curvature and the earth's magnetic field.
Al.3.2 Curved Earth/Ionosphere - Magnetic Field

In practice, the existence of the earth's magnetic field and the curvature of both the earth's surface and the ionospheric layers must be allowed for to enable the use of the above theory accurately.

The modification due to curvature is very complicated owing to the fact that it depends on the electron density profile and, therefore, differs from one set of ionospheric conditions to the next. However for most cases of interest it is sufficient to amend the Secant law of equation (Al.3.4) thus by

\[ f = K \cdot f_c \sec(\theta_0) \]  

where \( K \) is a correction factor. \( K \) is a function of the transmission range and the real height of reflection. Practical values of \( K \) range between 1.0 and 1.2 in most instances and the value 1.114 has been in use since 1944 by the Radio Propagation Committee in Washington D.C. for propagation over the important HF distance 3000 Km. This value is only a compromise however and \( K \) is only accurate when it is evaluated at a given point for a known electron density profile. This is a very complicated process involving the use of highly complex ray tracing programmes and large amounts of data.

The effect of the earth's magnetic field is found mainly in the definition of the refractive index, \( n \), of equation (Al.3.2). For longitudinal propagation, this equation is replaced by

\[ n = \left\{ 1 + \frac{f_c^2}{f(f - f_H)} \right\}^{1/2} \]  

in the frequency band of interest, \( f > f_H \); \( f_H \) is the gyrofrequency - the natural rotational frequency of electrons in the earth's magnetic field. Its value in our latitudes in Britain, is approximately \( 3.56H \cdot 10^6 \), where \( H \) is the magnitude of the field at the relevant point.
The main effect of the earth's magnetic field therefore is to split the incident radio wave into two parts corresponding to the two values of \( n \) obtained in the equation (A1.3.6). The plus sign gives the so-called ordinary (\( O \)) ray, while the negative sign is indicative of the extra-ordinary (\( X \)) ray. These two rays travel through the ionosphere by different paths and should recombine on leaving the ionosphere. This occurs in the case of a plane stratified ionosphere, but in practice, they do not, a situation which results in polarization losses.

The correction given in equation (A1.3.6) is in general small and is normally negligible for the higher HF frequencies, except those near the critical frequency. Here it is found that the \( O \)-ray will be reflected at an altitude greater than the \( X \)-ray for both vertical and oblique incidence. This means that the \( O \)-ray is closer to the vertical than the \( X \)-ray and so undergoes less absorption.
Al.4 Ionospheric Path Losses

Typically an HF signal propagates in many modes between transmitter and receiver; a phenomenon known as multipath propagation (section Al.5). The individual waves vary rapidly and randomly in phase relative to each other and so the total signal power, averaged over a period long compared to the fading cycles (section Al.5), is essentially the sum of the powers of the individual waves.

Path loss may therefore be defined as the average power loss suffered by the transmitted wave in its multipath passage through the ionosphere; where the power loss is averaged over a period long compared to the fading cycles.

The main components of path loss for a single ionospheric hop are:

1. $L_d$ - the spatial attenuation due to the inverse dependence of energy on distance
2. $L_a$ - the loss due to ionospheric absorption of radio waves
3. $L_f$ - the loss in signal power caused by the focusing, or defocusing, of the signal at frequencies near the MUF.

The path loss, $L_p$, is given by the sum of these three components in decibels as

$$L_p = L_d + L_a + L_f$$

(Al.4.1)

Al.4.1 Spatial Attenuation Loss - $L_d$

This is defined simply as

$$L_d = 20 \log_{10} S$$

(Al.4.2)

where $S$ is the distance in Km of the wavefront from the transmitting antenna. Equation (Al.4.2) expresses the fact that power disperses as it spreads from its excitation point.
Al.4.2 Ionospheric Absorption Loss - $L_a$

Ionospheric radio waves suffer absorption mainly in the D and E regions. It may be shown with magneto-ionic theory that a wave propagated vertically with frequency $f_v$ will suffer absorption $L(f_v)$ dB given by

$$L(f_v) \propto \frac{N}{\nu} \frac{df}{1 + (f_v + f_L)^2}$$

where $N$ - the electron concentration

$\nu$ - the electron collision frequency

$\mu$ - the refractive index

$f_L$ - the electron gyrofrequency about the vertical component of the earth's magnetic field

Integration limits are from the base of the ionosphere to the height of reflection at the frequency $f_v$. Absorption therefore depends on the height distribution of the electron density.

With reference to (Al.4.3), absorption is defined as deviative when $\mu \ll 1$ and non-deviative when $\mu = 1$. Therefore near the reflection level, or whenever there is pronounced bending of the ray, the radio signal suffers deviative absorption.

At vertical incidence, deviative absorption is large near the height of reflection when this occurs in the E-region. At temperate latitudes in winter, the median non-deviative absorption is greater than that expected under normal solar conditions; the day to day variability is also extensive.

At high latitudes there is additional non-deviative absorption. This is auroral and Polar Cap Absorption (PCA) due to the incidence of solar protons and electrons precipitated from the magnetosphere under disturbed conditions.

Absorption over a particular path may be predicted using one of the programmes supplied by the propagation forecasting services such
In the new CCIR method proposed by George and Bradley\textsuperscript{15}, three types of absorption may be calculated. These are:

1. normal non-deviative absorption
2. normal deviative absorption and
3. median winter anomaly absorption.

The importance of auroral absorption is recognized and it is hoped that this will be included in the programme eventually.

With the absorption factor defined as

\[ A(f_v) = L(f_v)(f_v + f_{\perp})^2 \]  

The form of its variation with frequency is found to depend solely upon \( f_v/f_{\text{OE}} \). Hence defines

\[ A(f_v)/A_T = \beta_n(f_v/f_{\text{OE}}) \]  

where \( A_T \) is the limit value of \( A(f_v) \) and assuming that \( f_v \) is high enough for the signal to traverse the whole of the absorption region without deviation, it is found that \( \beta_n \) is approximately independent of location, season or solar epoch. The function \( \beta_n \) is given in figure (Al.4.1) where it may be seen that \( \beta_n \) increases with \( f_v/f_{\text{OE}} \) as the depth of penetration increases, or that it reaches a maximum around

\[ f_v = f_{\text{OE}} \]  

where there is a large amount of deviative absorption in the E-region.

A world map for \( A_T \) is also available as shown in figure (Al.4.2). This was done by combining \( \beta_n \) from figure (Al.4.1) with empirical formulae derived from measurements of the diurnal and solar cycle variations of absorption. The data used to evaluate \( A_T \) came from absorption measurements at several stations using different frequencies.

The absorption at vertical incidence can be deduced for any frequency, place and time by using this map in conjunction with the \( \beta_n \) curves and the above mentioned empirical equations for diurnal and solar
**Fig. (A1.4.1) The Absorption Factor $\alpha_0 (\delta v/\delta \phi_e)$**

**Fig. (A1.4.2) The Absorption Factor at' for an Overhead Sun and a Smoothed Sunspot - Number of Zero.
cycle variation.

The diurnal variation is given by the formula in general use by CPL:

\[ K = 0.142 + 0.858 \cos X \]  \hspace{1cm} \text{(Al.4.6)}

where \( X \) is the Sun's zenith angle. At noon \( X = 0 \) and \( K = 1 \). Muggleton has proposed another formula which gives practically the same results. Viz

\[ A = A_0 (\cos X)^n \]  \hspace{1cm} \text{(Al.4.7)}

where \( n = 0.8 \) in summer and \( 0.86 \) in winter under calm solar conditions.

The solar cycle variation is given by \( q \) as

\[ q = 1 + 0.00372 \]  \hspace{1cm} \text{(Al.4.8)}

Equations (Al.4.8) and (Al.4.6) may be combined to give a factor called the absorption index \( I \) as

\[ I = (1 + 0.00372)K \]  \hspace{1cm} \text{(Al.4.9)}

So far absorption has been for vertical incidence. That at oblique incidence is of more importance here, and this is given by Martyn's absorption theorem as

\[ A(f_{o b}) = A(f) \sec \theta \]  \hspace{1cm} \text{(Al.4.10)}

Using ray tracing calculations for absorption for a range of model ionospheres, simple expressions have been found for the height at which \( i \) is to be determined and for \( f_v \) in terms of \( f_{o b} \). The complete step-by-step procedure is thus:-

1. \( I \) is first evaluated from (Al.4.9) or more simply from the nomogram in figure (Al.4.3) which incorporates the calculation for various values of \( X \) and \( K \).
2. Figure (Al.4.2) is then used to find \( A_v \) over the path to be used at the time of interest.
3. \( f_{o n} \) is then found from figure (Al.4.1) having calculated \( f_v/f_{o b} \).
4. Equation (Al.4.5) then gives the value of \( A(f) \) which is now
FIG. (A1.4.3) NOMO GRAPH FOR ABSORPTION INDEX I FROM THE SOLAR ZENITH ANGLE (θ) AND THE MEAN SUNSPOT NUMBER (R.S)
5. The loss $L(\phi)$ may now be calculated for vertical incidence from equation (Al.4.4) and thence for oblique incidence from (Al.4.10).

This will give the absorption path loss, $L_a$, in dB for any transmission path using one hop propagation.

**Al.4.3 Focusing Loss - $L_f$**

Focusing is due to the convergence or divergence (defocusing) of originally neighbouring rays. This is over and above the spatial effect discussed in section (Al.4.1).

In practice it is very difficult to take the focusing effect into account, as detailed knowledge of the ionosphere over the intended path is needed as well as an elaborate ray tracing programme. It is however possible to consider the effect of focusing in terms of an effective path length, $S_o$.

In the absence of all other forms of energy loss, the power flux at a distance $S_o$ from an isotropic radiator, which radiates a total power $P_{t_0}$, is given by

$$\phi = P_{t_0}/4 \cdot S_o^2$$  (Al.4.11)

Now consider figure (Al.4.4) where the energy radiated into the same cone by the antenna is deformed. If $S$ is the distance from the antenna at which the area covered by the defocussed rays is the same as that for the original case, then the power flux is the same for both. The defocusing can thus be accounted for by replacing the true distance $S$ of the receiver by an effective distance $S_o$. In this way the spatial loss and the focusing losses may be combined in the expression

$$L_d + L_f = 20 \log_{10} S_o$$  (Al.4.12)
Fig. (a14.4) Illustration of effective path length
It may be noted that $Se > 3$ for defocusing and $Se < 3$ for focusing. Further by considering the geometry of the beam near the transmitter and the horizontal spreading of the beam, it can be shown that

$$Se^2 = a \sin \Theta \tan \Delta \left( \frac{dD/d\Delta}{d\Delta} \right)$$

(A1,4.13)

where $a$ - the radius of the earth

$\Theta$ - the angle subtended at the earth's centre by the path along the earth's surface

$\Delta$ - elevation of the wave

$D$ - the ground range

$d\Delta$ - is the vertical angular width of the beam

For given values of $\Delta$, $a$, and $dD/d\Delta$, $Se^2$ may be calculated analytically in the case of certain model layers or as intimated before by means of ray tracing.

Focusing due to ionospheric distortion can be very important, particularly in trans-equatorial propagation during the afternoon and early evening because of the existence of a relatively stable deformation, in which the ionosphere simulates convex or concave mirrors. Such distortions of the F-layer can produce focusing and defocusing of the reflected radio signal amounting to between 5 and 10d$\text{B}$ variation in the path loss.

Path loss may also be increased when the E-layer exists as it does in daytime. This is due to the increase in $Se$ due to the extra refraction introduced by the underlying ionized layer.
Al. 5 Fading and Multinath

Path loss, as defined in section (Al.4) is the major ionospheric factor in the determination of the average received signal power.

Besides this, the instantaneous field strength may fluctuate wildly about this mean value, which sometimes produces amplitude variations in the order of 10:1 in the space of a few seconds. This phenomenon is known as fading and it may take a variety of forms depending upon the particular ionospheric characteristic producing it. The more important of these are:

1. Interference fading - due to movements in the ionospheric layers
2. Polarization fading - due to rotation of the axis of the polarization ellipses.
3. Absorption fading - caused by time variation in the degree of ionospheric absorption
4. Miscellaneous fading effects due to focusing and skipping of the signal due to MUF failure.

It may also be noted here that in the case of signals being transmitted from a moving source, e.g. a satellite, fading will result from the motion of the source relative to ionospheric irregularities. This is of importance in accurate satellite observation of the ionosphere.

Fading occurs in cycles having periods dependent upon the reason for the fading. For example, interference and polarization fading may vary from a fraction of a second to a few seconds; absorption fading may have a period of an hour or longer; fading due for focusing may last from a quarter to half an hour. 'Fade in' and 'fade out' due to MUF failure skipping is highly irregular in period and may occur in early morning or late afternoon when the ionospheric layers are being
formed (fade in) or dispersed (fade out) respectively.

In general fading is faster for the high frequencies compared to lower frequencies. This is due to the fact that a given ionospheric movement produces a greater phase shift on the shorter wave lengths. This frequency-dependent nature of fading means that different side-lobes in a modulated wave fade differently, giving rise to a distortion of the modulation envelope. This is known as frequency selective fading.

From the above it appears that fading may be split into two categories. First there is the slow fade evidenced by those types having periods in excess of one minute, and second there are the rapid fades with periods of a few seconds maximum. This implies that all fading other than interference and polarization may be defined as slow fade types.

Slow fades are due mainly to random daily and seasonal changes of the absorption of the wave in the D-region. Recommendations based on long-term observations suggest that an allowance of $14\text{dB}$ be added to the monthly median signal intensity for a given circuit\(^{18}\). The monthly median value is here defined as the intensity which provides an acceptable SNR, defined by the service grade requirements, for fifteen days of the month. This solution implies that slow fades are not accurately enough defined for optimum compensation. However by means of an adaptive system\(^{19}\) where the ionosphere is continuously monitored from the transmitter, and the signal intensity and frequency is varied appropriately to accommodate the slow fade mechanisms, it seems that this problem is being tackled in a more realistically optimum way.

Rapid fading is best discussed in relation to multipath propagation. A beam of radio waves incident on the ionosphere is not reflected from
from an extended region. Small irregularities in electron density near the level of reflection give rise to individual reflected wavelets, each of which arrives at the receiver over a different path, and hence at non-coincident times. The time difference between the arrival of the first and last wavelets is called the multipath spread of the medium. This is a gross characteristic of the medium and is nominally defined in the same way as bandwidth for a filter. Movements of the irregularities causing multipath effects produce variations in the relative phases of the individual wavelets and so give rise to interference fading.

The received HF signal may therefore consist of a mixture of high and low angle rays, each having an $\pi$- and $\varphi$- component (section Al.3). Each set of such waves may also be combined with other sets corresponding to modes having a different number of hops. It is reasonable therefore to expect that the resultant amplitude can vary over wide limits with a maximum when all the components are in phase. The RMS value of the fluctuating signal is equal to the steady value of the field that would have existed had the ionosphere not split it into many components. Because it is impossible to determine the resultant amplitude at a given moment, the subject of rapid fading is usually treated statistically.

The distribution of amplitude approximates the Rayleigh law when the various components are of approximately equal amplitude and the relative phases are varying randomly. For the Rayleigh distribution

$$p(A) = \exp(-A^2/A_R^2)$$  \hspace{1cm} (Al.5.1)

gives the percentage of time that the amplitude exceeds $A$. Here $A_R^2$ is the mean square value of $A$. Signals in the ionosphere are often better written in terms of the Rice distribution which adds a
constant amplitude wave to the randomly varying signals. If $A_S$ represents this steady amplitude, then

$$p(A) = \frac{2}{A_R^2} \int_{-\infty}^{\infty} A \exp \left[ - \frac{A^2 + A_S^2}{A_R^2} \right] \cdot I_0 \left( \frac{2AA_S}{A_R^2} \right) dA$$

(A1.5.2)

where $I_0$ is the Bessel function of zero order and imaginary argument.

Equation (A1.5.2) approaches the Rayleigh law for $A_S/A_R << 1$, and for $A_S/A_R >> 1$ it approaches a normal distribution with a mean of $A_S$ and a standard deviation of $0.707 A_R$.

In the case of continuous waves and long trains of pulses, the amplitude distribution is usually near-Rayleigh. Some individual modes produced by short pulses often have shallower fading corresponding to a substantial secular component ($A_S$) (ie) $A_S/A_R = 2$ or more.

The Rayleigh distribution gives a median amplitude of 0.832 times the RMS value, and for such a distribution the lower decile value, or the amplitude exceeded 90% of the time is 0.39 times the median value, whilst the upper decile value, exceeded 10% of the time - is 1.8 times the mean.

Fading occurs at a particular rate definable in terms of either the autocorrelation function of the amplitude or its Fourier transform. This latter is termed the fading frequency power spectrum and its width is equal to the fade rate. This rate is dependent upon ionospheric variations and upon changes in receiver position evidenced by Doppler shifting defined by the Doppler spread factor similarly to the multipath spread factor.

If the multipath spread factor for a channel is $T_M$ and the Doppler spread is $B_D$, then the channel spread factor is given by

$$L = T_M B_D$$

(A1.5.3)
If therefore it is desired to examine the channel conditions by using sounding pulses, then the period between successive pulses must be at least $T_M$ to avoid ambiguities. On the other hand, the medium changes state in a time of the order $1/B_D$. This implies that the instantaneous state of the medium is measurable using periodic pulses only if

$$T_M << \frac{1}{B_D} \quad (A1.5.4)$$

(ie) the spread factor must be well below unity. Furthermore if the signal pulses are of length $T$, then $T >> T_M$ to avoid significant intersymbol interference, since only then will the energy arriving at any instant all primarily relate to one signal pulse. On the other hand, to avoid severe time fading distortion of the pulse, equation (A1.5.4) must be satisfied. These two conditions are met only if

$$T_M << T << \frac{1}{B_D} \quad (A1.5.5)$$

which again requires a spread factor much less than the unity. It is because the ionosphere satisfies this criteria that it may be used for the practical transmission of data.

To minimize the effects of fading on signal quality, use is made of the fact that fading is correlated only over relatively small disturbances and frequency bandwidths. Also signals propagated from antennas of different polarizations may fade differently. These factors lead to the use of diversity reception which uses special antenna system configurations. Diversity reception is discussed in Chapter 3.
AI.6 Ionospheric Radio Noise

As noted in Chapter 2, noise is the greatest limiting factor to efficient communication. There, receiver noise was discussed and the (SNR)\textsubscript{0} which defines the signal quality was evaluated. This discussion will describe and evaluate the processes causing noise in the ionosphere. This combined with the assertions in sections (AI.4) and (AI.5) will provide the necessary means to estimate the ionospheric radio signal quality.

Ionospheric noise power is measured in terms of \( F_a \), which may be defined as the external noise power available from a lossless antenna. It is expressed in dB above \( kT_0 \), where \( k \) = Boltzmann's constant. \( T_0 \) is a reference temperature (288.39\,K) which is the noise generated in a unit bandwidth by a thermal source of temperature \( T_0 \).

If it is assumed that the noise is uniformly incident on the antenna from every direction and, that its power is proportional to the bandwidth \( B \) in Hz, then the total power \( P_n \) in dB above 1 watt, available at the terminals of a lossless antenna is given by

\[
P_n = (F_a + BL - 204) \, \text{dB/watt}
\]

where \( BL = 10 \log_{10} B \) and \( 10 \log_{10} kT_0 = -204 \)

Ionospheric noise affecting HF radio communication is of two types which are

1. Atmospheric noise.
2. Cosmic noise.

AI.6.1 Atmospheric Noise

This is caused by local and distant thunderstorms and is propagated via the ionosphere similarly to radio waves. The noise from distant storms is propagated with great power, measured typically in terms of gigawatts, via several modes, most of which would completely
attenuate even a powerful radio wave. It therefore arrives at the receiving antenna at many different instants. This process produces essentially white noise with a continuous spectrum, which varies greatly in intensity over the entire electromagnetic spectrum. In the case of local storms the noise is of the 'shot' variety and results in impulse-like excitations of the radio circuit.

The atmospheric noise field depends on
1. The geographic location of the receiver. The maximum noise will occur in the vicinity of the thunderstorm centre.
2. The season - the field is stronger during local summer.
3. The time - the field is stronger during the night.
4. The frequency of operation. The field is very strong for frequencies less than 1MHz; it increases to about 9MHz and then decreases becoming negligible between 15 and 20MHz.
5. The bandwidth of receiver - it is proportional to the square root of the passband.

The dependencies imply that noise may be predicted in the same way as the ionospheric variables already discussed. Here however the predictions will be much less accurate due to the large spread in amplitude values; errors typically are in the order of 5dB, but sometimes may be as much as 20dB which clearly limits the usefulness of predictions about atmospheric noise. Despite these inherent errors, the CPRL world maps for atmospheric noise prediction have been widely used for a long time, as has the CCIR system, which though more complicated is more accurate.

These systems give for each location, season and time of day
1. The field value requires in the presence of noise for a particular type of service (CPRL)
2. The field value equivalent to the noise (CCIR), or
3. The values in dB of the ratio of noise powers to thermal noise of an elemental antenna (CCIR).

Apart from regular variations that are quite accurately predicted, atmospheric noise varies considerably from day to day, even in the absence of local thunderstorms. In the presence of local storms, the noise power can be extremely high - a situation that renders reception impossible. It may also be noted that atmospheric noise often has a preferred direction - this being the direction of the centre of thunderstorm activity.

The mean atmospheric noise may be evaluated by the CCIR method given in report 3222k. In this method, the seasons are divided into months as shown in Table (A1.6.1). Figure (A1.6.1) shows a sample of the curves given for each season at the local time of interest. These curves give the mean value of the noise $F_{am}$, which is defined as the median value of hourly values of $F_a$ (equation A1.6.1) over a period of time as a function of

1. The geographic location of the receiver
2. The season (Table A1.6.1)
3. The time of day in the intervals 00-04, 04-08, 08-12, 12-16, 16-20, 20-24, hours local time.

$F_{am}$ is given in dB above thermal noise at a frequency of 1MHz. Figure (A1.6.2) (a) will now give the value of $F_{am}$ for the frequency of interest by the use of $F_{am}$ obtained from (A1.6.1). This new value of $F_{am}$ is now converted to the median noise field $E_n$ for the frequency of interest by using the nomogram in figure (A1.6.3), which gives $E_n$ in dB above $\mu W/m$. Alternatively, the equation

$$E_n = F_a + 20 \log_{10} (f MHz) - 65.5$$  \hspace{1cm} (A1.6.2)

may be used.

Atmospheric noise obeys a normal logarithmic distribution.
<table>
<thead>
<tr>
<th>Month</th>
<th>Local Season</th>
<th>Northern Hemisphere</th>
<th>Southern Hemisphere</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dec, Jan, Feb</td>
<td>Winter</td>
<td></td>
<td>Summer</td>
</tr>
<tr>
<td>Mar, Apr, May</td>
<td>Spring</td>
<td></td>
<td>Autumn</td>
</tr>
<tr>
<td>Jan, Jul, Aug</td>
<td>Summer</td>
<td></td>
<td>Winter</td>
</tr>
<tr>
<td>Sept, Oct, Nov</td>
<td>Autumn</td>
<td></td>
<td>Spring</td>
</tr>
</tbody>
</table>

Table A1.6.1
FIG(A1.6.1) SPRING—BETWEEN 0000 AND 0400HR
MEAN NOISE VALUES
**FIG. A1.63** NOMOGRAPh FOR CONVERTING $F_a$ TO $E_n$ AS A FUNCTION OF FREQUENCY.
similarly to the signal median values (section Al.5). In the case of ionospheric waves, the instantaneous signal values vary around the hourly median in a Rayleigh distribution. The fluctuation margin is defined as the change in signal strength required to maintain the SNR above the service grade requirement for a certain percentage of time, the time availability. This margin may be deduced from figure (Al.6.2) (b). The value \( D_u \) of the maximum deviation of \( F_{am} \) for 90% time availability is read and \( \sigma \), the standard deviation of the ratio of the hourly median values of SNR is given by

\[
\sigma = \left( \frac{D_u^2 + D_s^2}{1.3} \right)^{1/2}
\]

(Al.6.3)

where \( D_s = 0 \) for a ground wave and 7dB for the sky wave. Using figure (Al.6.4) for single reception, or figure (Al.6.5) for double diversity reception, \( \sigma \), and the complement to unity of the time availability \( q \) will give the fluctuation margin expressed in dB from the ordinate.

\( F_{am} \) and the fluctuation margin are both mean values. In view of the unavoidable uncertainty of the data used to calculate them, only half of the circuits realized by these data will produce the expected result. If the service probability must exceed 50%, then a power margin must be added. This is normally referred to as the uncertainty margin.

From figure (Al.6.2) (b), \( \sigma_{Dv} \) is read and used to give \( \sigma_{cu} \) as

\[
\sigma_{cu} = \left( \sigma_{Dv}^2 + \sigma_{Ds}^2 \right)^{1/2}
\]

(Al.6.4)

where \( \sigma_{cu} \) is the standard deviation of the distribution of the ratio of the hourly median values of SNR for 90% time availability. \( \sigma_{Dv} \) is zero for the ground wave and 1.5 for the sky wave. Finally figure (Al.6.6) is used to calculate the uncertainty margin. \( \sigma_c \) is read on the central scale as the point where the line joining \( P \), the percentage of time availability, on the LHS scale, and \( \sigma_{cu} \) from the equation (Al.6.4) on the RHS scale, meets the central scale. This
FIG. (A1.6.4) CONVOLUTION OF RAYLEIGH DISTRIBUTION WITH LOG NORMAL DISTRIBUTION SINGLE RECEIVER.

FIG. (A1.6.5) FADE DISTRIBUTION FOR DIVERSITY RECEPTION.
Fig. (A1.6.6) Uncertainty Margin.
is then used to calculate $\sigma_T$ as

$$\sigma_T = \left( \sigma_c^2 + \sigma_{\text{Fam}}^2 + \sigma_p^2 + \ldots \right)^{\frac{1}{2}} \tag{A16.5}$$

where $\sigma_{\text{Fam}}$ is given in figure (A16.2) (b), $\sigma_p$ is the quadratic average error in the calculation of the received power ($2 - 5\text{dB}$), and all other possible errors being added in as they become relevant. These latter are usually characteristic of a particular path and include errors due to several factors based on local knowledge of noise sources in the atmosphere.

Having calculated $\sigma_T$, the uncertainty margin may be read from the central scale of figure (A16.6).

The values obtained for $E_n$, the fluctuation margin and the uncertainty margin must all be added to the SNR required for service grade under stable conditions, to evaluate the field required in the presence of atmospherics for service grade operation of the circuit. Some correction must usually be made for the discrimination gain of the antenna.

Atmospheric noise is by far the most important noise of ionospheric origin affecting communication reliability. This is because of its great unpredictability due particularly to the shot noise nature of local thunderstorms. Cosmic noise is much more predictable and can therefore almost always be compensated for in system design.

A.16.2 Cosmic Noise

The origin of this extra-terrestrial type of noise is still not absolutely certain but because the direction of its maximum intensity is fixed in relation to the stars, it is almost certainly a combination of intense point-source radio stars at large distance and the integrated noise, of the same kind, from the collected assembly of stars within our galaxy. It is identical to white noise and varies slowly in
intensity with frequency.

Cosmic noise is expressed in $\mu V.m^{-1}$ for a given transmission band. However as with other types of noise, the terminology 'signal reduced in the presence of noise' is often used in the case of HF waves. Since the noise field varies only marginally, a small margin or none at all is taken for fading, and the following expression may be used for converting from one passband to another

$$En = E_{on} + 10 \log B$$

(Al.6.6)

where $E_{on}$ = noise field in dB relative to $1\mu V.m^{-1}$ for a passband of 1Hz

$B$ = passband in Hz.

The value of cosmic noise to be used in system calculations will depend greatly upon the directivity of receiving antennas. Here two cases must be considered.

1. Antenna with little directivity

2. Antenna with high directivity

In the former case the noise is quite constant, and the sum of all measurements, which are normally coherent, gives a mean value of the cosmic noise as

$$En = -47 + 10 \log(B) -1.5 \log f$$

(Al.6.7)

in dB referred to $1\mu V.m^{-1}$. The operating frequency is $f$MHz and $B$ is the receiver passband in Hz.

In the latter case, the maximum value of the noise field is the most important parameter. This may be estimated approximately from figure (Al.6.7)

Using the left-hand side of this figure, the intersection of the latitude curve, for the receiver and the azimuth angle towards which the receiving antenna is directed (counting from the north in a direction to give $180^\circ$) is determined. The intersection of a horizontal line drawn through this point with the curve at the relevant frequency gives
\[ \sin D = \cos \lambda \cos A \]

**Fig (A1:6-7)**
the value of $10 \log(E_{\text{max}})$ which can be read from the horizontal scale on the right-hand side.

In this case, the cosmic noise will pass through a maximum each day at the same sidereal time as the earth spins, and the position of the antenna changes relative to the stellar sphere.
Al.7 Ionospheric Disturbances

The ionospheric variables that affect signal quality have so far been discussed with respect to mainly average conditions. At these times, even with bad fading, it is usually possible to establish reliable HF communication links with the aid of ionospheric prediction techniques. Often however, the sudden onset of severe ionospheric disturbances causes departures from this norm. This results in the severe degradation of signal quality, frequently to the extent of total communication 'black-out'. These disturbances are all associated in one way or another with solar flares, and they may be classified as

1. Sudden ionospheric disturbances (SIDS)
2. Ionospheric storms
3. Polar Cap Absorption events

When a solar flare occurs two major changes take place in the ionosphere.

1. D-region absorption is so greatly enhanced that intelligible radio communication may be impossible for periods ranging from a few minutes to several days on particular paths.
2. The critical frequencies of the F2 layer are depressed by ionospheric storms. This produces a loss of signal due to MUF failure.

The first of these changes occurs simultaneously with the onset of an optical solar flare, the second is often delayed by a day or more. Besides these two large scale disturbances, there are a variety of lesser effects which also degrade signal quality. These are depicted in figure (Al.7.1) together with the particular electromagnetic emissions from a solar flare that produce them. Not all the effects shown in the figure are associated with every flare event. This implies that different flares will affect the same path differently.
It is very difficult to predict the onset of an ionospheric disturbance, and since ionospheric conditions can change very rapidly from very quiet to highly disturbed, the idea of average conditions tends to be of limited usefulness. It is possible however to predict the probable occurrence of an ionospheric storm, which usually occurs a relatively long time after the start of an ionospheric disturbance. In this way, the effect of these variations may be to some extent compensated for in practice.

The forecasts are given for a specified 6-, 12- or 24-hour period in universal time, by CPRL and the North Atlantic Radio Warning Service. Using these predictions, an operator using a link likely to be affected by the storm can usually transmit essential messages before his circuit is 'shut down' by the storm. Alternatively, he may be able to set up various relay links at other usable frequencies of he has this facility available.

In terms of an optimised radio link, these facilities, in some form or other, will be essential and so, in the context of this thesis, the existence of ionospheric disturbances is an important factor in cost and complexity of equipment and in total system design for optimum flexibility.
APPENDIX 2

COMPUTER AIDED DESIGN OF HF ANTENNA SYSTEMS
A2.1 The Horizontal Dipole Array (HDA)

The HDA consists of half-wave (\(\lambda/2\)) dipoles arranged so that one or more dipoles appear in each row and column of the array. Such an arrangement is shown in figure (A2.1.1) for \(\lambda/2\) dipoles/row and \(\lambda/2\) dipoles/column in an HDA designated \(H\lambda/3h\) — where \(h\) defines the height in wavelengths of the lowest row of dipoles above ground.

In the HDA, as the name implies, the rows of dipoles are parallel to the ground making each dipole in the array horizontal with respect to the ground. Such an array thus radiates horizontally polarized waves.

This arrangement of simple dipole antennas produces increased gain and directivity over the single dipole antenna. In the vertical plane, directivity is increased by the stacking of dipoles in columns or bays, while in the horizontal plane, it is increased by adding more bays with the elements of adjacent bays collinear.

The HDA as shown in figure (A2.1.1) radiates bidirectionally, but the addition of a reflector curtain behind the radiator produces unidirectionality with a 3dB increase in gain and minimal change of the forward radiation pattern. This reflector may be made of tuned elements operated in a parasitic mode or by the use of a screen made of closely spaced horizontal wires.

Dipole arrays are used when relatively cheap antenna systems which occupy relatively little area are needed for intermediate range communication requiring modest amounts of gain. They are not widely used in modern point-to-point circuits where rhombic antennas and log periodic dipole arrays (LPDAs) are preferred. In certain cases however the HDA is the only viable means for providing economic and extensive frequency-handling capability in the HF band. Such is the situation in the case of the Cape D'Aguiar transmitter. Here, because the station is
FIG. (A2.1.1)

HORIZONTAL DIPOLE ARRAY CONFIGURATION.

\( \frac{\lambda}{2} \)

\( \frac{\lambda}{2} \)
situated on a high peninsula which slopes steeply to the sea, space is restricted. There are however good Fresnel reflection zones. This makes it possible to support HDAs at optimum heights from only two masts each placed along suitable contours parallel to the coast.

Table (A2.1.2) lists the antennas used on the site, and figure (A2.1.2) shows the layout of the antennas listed as well as geographic information about the peninsula. From these data it can be seen that of the forty-seven antennas used all but four are HDAs. Of these four, three are rhombics and one is an LPDA.

The HDA is discussed in many texts. Each of these gives an equation or equations for the field strength of the HDA in terms of various physical parameters of the antenna. In this investigation the set of equations used by CCIR to produce their antenna polar plots will be used as a starting point to produce an algorithm for computer use. The equations are

A. For horizontally polarized arrays one dipole wide

Type H1/n/h

\[
\frac{E}{E_{\text{max}}} = \frac{\cos(\pi/2 \cos \theta \sin \phi)}{r = n-l} \cdot \frac{\sin \left(2 \left(h + 0.5r\right) \sin \theta\right)}{r = 0} \cdot \left(1 - \cos^2 \theta \sin^2 \phi\right) \sum_{r = 0}^{r = n-l} \sin \left(2 \left(h + 0.5r\right) \sin \phi_{\text{max}}\right)
\]

(A2.1.1)

B. For horizontally polarized arrays two dipoles wide

Type H2/n/h

\[
\frac{E}{E_{\text{max}}} = \cos^2 \left(\pi/2 \cos \theta \sin \phi\right) \cdot \frac{\sin \left(2 \left(h + 0.5r\right) \sin \theta\right)}{r = 0} \cdot \left(1 - \cos^2 \theta \sin^2 \phi\right) \sum_{r = 0}^{r = n-l} \sin \left(2 \left(h + 0.5r\right) \sin \phi_{\text{max}}\right)
\]

(A2.1.2)
<table>
<thead>
<tr>
<th>Aerial No.</th>
<th>Circuit</th>
<th>Type</th>
<th>Freq. m/c</th>
<th>Aerial No.</th>
<th>Circuit</th>
<th>Type</th>
<th>Freq. m/c</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bangkok</td>
<td>HR2/2/06</td>
<td>14</td>
<td>26</td>
<td>Bandong/</td>
<td>Djakarta</td>
<td>HR2/2/05</td>
</tr>
<tr>
<td>2</td>
<td>&quot;</td>
<td>HR2/2/06</td>
<td>12</td>
<td>27</td>
<td>Saigon</td>
<td>HR2/2/065</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>&quot;</td>
<td>HR2/2/08</td>
<td>8</td>
<td>28</td>
<td>Taipei</td>
<td>HR2/2/05</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>&quot;</td>
<td>HR4/3/03</td>
<td>15</td>
<td>29</td>
<td>Japan</td>
<td>HR2/2/05</td>
<td>14</td>
</tr>
<tr>
<td>5</td>
<td>Colombo</td>
<td>HR2/3/23</td>
<td>20</td>
<td>30</td>
<td>&quot;</td>
<td>HR2/3/05</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>Jesselson</td>
<td>HR2/2/05</td>
<td>15</td>
<td>31</td>
<td>Saigon</td>
<td>HR2/1/25</td>
<td>14</td>
</tr>
<tr>
<td>7</td>
<td>Rangoon</td>
<td>HR2/1/09</td>
<td>17</td>
<td>32</td>
<td>Taipei</td>
<td>HR2/1/25</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>Calcutta</td>
<td>HR2/2/07</td>
<td>14</td>
<td>33</td>
<td>&quot;</td>
<td>H4/1/09</td>
<td>3-5</td>
</tr>
<tr>
<td>9</td>
<td>Taipei</td>
<td>HR4/1/06</td>
<td>6</td>
<td>34</td>
<td>&quot;</td>
<td>H4/1/15</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>&quot;</td>
<td>HR4/1/05</td>
<td>10</td>
<td>35</td>
<td>Sydney</td>
<td>Rhombics</td>
<td>8-21</td>
</tr>
<tr>
<td>11</td>
<td>Manila</td>
<td>HR4/1/12</td>
<td>15</td>
<td>36</td>
<td>Saigon</td>
<td>HR2/2/04</td>
<td>7</td>
</tr>
<tr>
<td>12</td>
<td>Laos</td>
<td>Cage Quad</td>
<td>15</td>
<td>37</td>
<td>Singapore</td>
<td>Rhombics</td>
<td>6-18</td>
</tr>
<tr>
<td>13</td>
<td>Ship Shore</td>
<td>&quot;</td>
<td>10</td>
<td>38</td>
<td>Log-Periodic</td>
<td></td>
<td>6-30</td>
</tr>
<tr>
<td>14</td>
<td>Telecast</td>
<td>&quot;</td>
<td>17</td>
<td>39</td>
<td>Shanghai</td>
<td>H2/1/1-4</td>
<td>7</td>
</tr>
<tr>
<td>15</td>
<td>Hanoi</td>
<td>&quot;</td>
<td>10</td>
<td>40</td>
<td>San</td>
<td>Rhombics</td>
<td>7-20</td>
</tr>
<tr>
<td>16</td>
<td>Manila</td>
<td>HR4/1/07</td>
<td>6</td>
<td>41</td>
<td>Francisco</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>&quot;</td>
<td>HR2/2/045</td>
<td>10</td>
<td>42</td>
<td>Shinghui</td>
<td>HR2/2/04</td>
<td>11</td>
</tr>
<tr>
<td>18</td>
<td>&quot;</td>
<td>HR2/1/45</td>
<td>12</td>
<td>43</td>
<td>Seoul</td>
<td>HR2/2/11</td>
<td>18</td>
</tr>
<tr>
<td>19</td>
<td>&quot;</td>
<td>HR2/2/45</td>
<td>14</td>
<td>44</td>
<td>Japan</td>
<td>HR2/2/11</td>
<td>13</td>
</tr>
<tr>
<td>20</td>
<td>&quot;</td>
<td>HR2/2/45</td>
<td>11</td>
<td>45</td>
<td>&quot;</td>
<td>HR2/2/06</td>
<td>7</td>
</tr>
<tr>
<td>21</td>
<td>Saigon</td>
<td>HR2/2/04</td>
<td>12</td>
<td>46</td>
<td>&quot;</td>
<td>HR2/2/05</td>
<td>19</td>
</tr>
<tr>
<td>22</td>
<td>&quot;</td>
<td>HR2/1/04</td>
<td>8</td>
<td>47</td>
<td>&quot;</td>
<td>HR2/2/30</td>
<td>15</td>
</tr>
<tr>
<td>23</td>
<td>Singapore</td>
<td>HR4/2/11</td>
<td>20</td>
<td>48</td>
<td>&quot;</td>
<td>HR2/2/05</td>
<td>15</td>
</tr>
<tr>
<td>24</td>
<td>Phnom penh</td>
<td>HR2/1/08</td>
<td>14</td>
<td>49</td>
<td>Kamchik</td>
<td>HR4/2/1</td>
<td>12</td>
</tr>
<tr>
<td>25</td>
<td>Manila</td>
<td>HR4/1/07</td>
<td>5</td>
<td>50</td>
<td>&quot;</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table (A2.1.2.) List of antennas used at Cape D'Aguilar.
Figure (A2.1.2.) Layout of antenna system at Cape D'Aguilar.
C. For horizontally polarized arrays four dipoles wide

Type H4/n/h

\[
\frac{E}{E_{\text{max}}} = \frac{\cos (\pi \cos \theta \sin \beta \cos^2 (\phi) \Sigma_{r=0}^{n-1} \sin 2 (h + 0.5r) \sin \theta)}{(1 - \cos^2 \theta \sin^2 \beta) \Sigma_{r=0}^{n-1} \sin 2 (h + 0.5r) \sin (\theta_{\text{max}}) }
\]

(A2.1.3)

The symbols are:-

- \( E \) = field strength due to the array at 1 Km distance (mV.m\(^{-1}\))
- \( E_{\text{max}} \) = the maximum value of \( E \) (mV.m\(^{-1}\))
- \( h \) = the height of the lowest dipole or row of dipoles of the arrays (wavelengths, \( \lambda \))
- \( n \) = number of rows or stacks of dipoles in the array
- \( \theta \) = latitude (degrees) (ie) elevation
- \( \theta_{\text{max}} \) = elevation of main lobe of radiation
- \( \phi \) = longitude or azimuth (degrees)
- \( \lambda \) = wavelength

These equations cover the HDAs of interest in this investigation.

Slight modification of the general form will give expressions for arrays eight dipoles wide and for slewed arrays. Interested readers are referred to the CCIR handbook.\(^{10}\)

The arrays represented above have no reflectors and thus radiate bidirectionally. If reflectors are used, as is generally the case, then they would be placed approximately a quarter of a wavelength away from the driven arrays and would be expected to have currents in them equal in amplitude to, but 90° ahead in phase of the driven arrays. This situation would produce suppression of the unwanted major lobe and the concentration of more energy (3dB) into the desired lobe. In practical working conditions where the operating frequency may differ slightly from that for which the antenna was built, or when the angle of the main beam is slewed, the reflector will not be more than 90°
efficient in the forward direction and the backward radiation may well be of the order of 10% of the forward radiation which implies a front-to-back ratio of only 20dB in field intensity is achievable in general.

Examination of equations (A2.1.1), (A2.1.2) and (A2.1.3) will show that they are in fact very similar. The only differences exist in the expression multiplying the summation from \( r = 0 \) to \( r = (n - 1) \) in the numerator. These 'multipliers' are

- in (A2.1.1) \( \cos \left( \frac{\pi}{2} \cos \theta \sin \phi \right) \)
- in (A2.1.2) \( \cos^2 \left( \frac{\pi}{2} \cos \theta \sin \phi \right) \)
- in (A2.1.3) \( \cos \left( \pi \cos \theta \sin \phi \right) \cos^2 \left( \frac{\pi}{2} \cos \theta \sin \phi \right) \)

In terms of programming for the computer, this arrangement is very fortunate as a very compact programme can be written which will have a speedy execution and which will, when desired, produce sets of data for a given HDA configuration.

The flow chart for such a programme is shown in Figure (A2.1.3). The programme written using this flow chart is in Fortran IV and is shown on the next page.

This programme can be easily adapted for use as a subroutine to a main programme and may then be accessed at will to give any data about the HDA of interest to the main programme which will in general be used for designing the optimum units to be used for the correct operation of an HF radio communication link, as it will be in this investigation.

The next aerial of interest here is the rhombic antenna.
Figure (A2.1.3)
Flow Chart for HDA programme

START

read data as follows
POWER - from transmitter
DELTA - elevation angle of main lobe
BRTTOR - bearing transmitter to receiver
IFLAG - no. of dipoles wide
N - no. of rows of dipoles
H - height of lowest row of dipoles above ground
ABR - bearing of aerial east of north
DELMH - elevation

Evaluate the azimuth angle phi as
PHI = 1 ABR - BRTTOR1

Select correct multiplier
IFLAG = 1  - Equation (3.3.1)
IFLAG = 2  - Equation (3.3.2)
IFLAG = 4  - Equation (3.3.3)

Evaluate value of expressions not contained in summation make equal to ACOT

Calculate quantities in summation signs; make dividend summation = SDIV and quotient
summation = SHUNI
Evaluate \( \text{SUM} = \text{SDIV} / \text{SNUM} \)

Calculate field strength at 1 Km in microvolts 1 m as

\[ \text{FIRAT} = (qcon \times \text{SUM} \times E\text{MAX} \times 10^3) \]

Calculate gain relative to half wave dipole

PRINT

STOP

E\text{MAX} is read from Figures 9-14 in ref 10

Figure (A2.1.3) Flow Chart for NDA programme
PROCEDURE FOR HORIZ DIPOLE ARRAYS

PI = 3.142
RDNS = 57.2957795131
READ, POWER
READ, DELTA, BRITTOR
READ, IFLAG, N, H, ABR, DELTM

IFLAG -- NUMBER OF DIPOLES WIDE
N -- NUMBER OF ROWS OR STACKS OF DIPOLES IN ARRAY.
H -- HEIGHT OF LOWEST DIPOLE OR ROW OF DIPOLES OF THE ARRAY
ABOVE THE GROUND (WAVELENGTHS)
ABR -- BEARING OF AERIAL (E. OF N.)
DELTM -- ANGLE OF ELEVATION OF MAIN LOBE OF RADIATION (DEGREES)

PHI = ABS (ABR - BRITTOR)
IF (IFLAG. EQ. 1) GO TO 200
IF (IFLAG. EQ. 2) GO TO 300
IF (IFLAG. EQ. 4) CONTINUE
L = 1
M = 2
GO TO 400
200 L = 0
M = 1
GO TO 400
300 L = 0
N = 2

CANG = (COS(DELTA/RDNS) * SIN(PHI/RDNS))
C = CANG * PI
DIV = (COS(C)**L) * (COS(0.5*C)**M)
QCON = DIV/((1 - CANG**2)**0.5)
SNUM = 0
SDIV = 0
DO 500 NOD = 1, N
X = 2. * PI * (H + (0.5 * (NOD - 1)))
SDIV = SDIV+SIN (X*SIN(DELTA/RDNS))
500 SNUM = SNUM+SIN (X*SIN(DELTM/RDNS))
SUM = SDIV/SNUM
FIRAT = QCON * SUM * EMAX * (POWER**0.5)
WRITE (6,800) FIRAT

FORMAT ('0', 1X, 'FIELD STRENGTH AT 1KM = ', F10.3, ' MV/M')

DBS RELATIVE TO 1 MICROVOLT/METER.
DBS = 20. * ALOG10(FIRAT * (10.**3))
WRITE (6,900) DBS

FORMAT (1X, 'DBS RELATIVE TO 1 MCVOLT = ', F10.3, 'DBS')
CONTINUE
STOP
END
A2.2 The Rhombic Antenna

The rhombic antenna is of the travelling wave variety. It consists of wires which are long compared to the wavelength, \( \lambda \). These are arranged in the shape of a rhombus as in figure (A2.2.1). One apex of the major axis is connected through an impedance matching network to the transmitter, or receiver. The other end is resistively terminated. The terminating resistance is equal to the rhombic antenna characteristic impedance and serves to dissipate power not radiated by the antenna; this ensures that there are no detrimental reflected waves. The power dissipated in the termination ranges in magnitude from 50% of the total power for small rhombics (side length \( L < 2\lambda \)) to 15% for large rhombics (\( L > 2\lambda \)). This power wastage is one of the main disadvantages of the rhombic and is one of the reasons for ensuring that this type of antenna is usually made as large as is economically viable.

Since the early 1940s the resonant arrays discussed in section (A1.1) have been gradually replaced by the rhombic for HF fixed services. This adjustment has taken place because of the more attractive operational and economic characteristics of the rhombic as compared to the dipole array.

Two parameters are important in investigating the performance of the rhombic. These are the side length \( L \), measured in wavelengths, and the half side angle, \( \gamma \) (see figure (A2.2.1)). The variation of the directivity gain of a free space rhombic with \( L \) and \( \gamma \) is given in figure (A2.2.2) where it is assumed that the elevation angle, \( \theta \) is zero. It can be seen that for high gain, \( L \) and \( \gamma \) must both be large. If, as is usual from the economic viewpoint, the antenna has to perform adequately over a reasonable frequency band, it is necessary to accept
FIG(A2.2.1) HORIZONTAL RHOMBIC ANTENNA WITH SYMBOLS.
a compromise based on the ratio of lowest to highest relevant frequencies. If this ratio is 1:2, a free space gain varying by 3dB from the upper to lower limit is achievable with a penalty of about 1.5dB at the upper frequency. Widebanding is also accompanied by some variation in the radiation characteristics of the aerial when erected above the ground; a change in $\theta$ of the main beam in the order of 2 to 1 being the most significant.

Figure (A2.2.3) shows the dependence upon a third parameter, $\theta$, the angle of elevation chosen for the main beam. It is at once obvious that high gain will only be maintainable at relatively low angles of elevation. If a practical limit to the side length is fixed at $10\lambda$, the maximum possible free space gain cannot exceed 13dB under optimum conditions (ie) $\theta = 5^\circ$ and $\gamma = 2.5^\circ$.

In practice $\theta$ is closely related to $h$, the height of the antenna above the ground reflecting plane. The height factor is given by

$$h_{fac} = 2 \sin (2\pi(h/\lambda) \sin \theta)$$

(A2.2.1)

and it must be chosen to optimize at or near the same value of $\theta$ as the rhombic factor if performance is not to suffer. This factor is plotted in figure (A2.2.4) and it will be seen that a height of $3\lambda$ corresponds to a $\theta$ of $5^\circ$ which is compatible with a high performance rhombic (figure A2.2.3)

When designing an aerial system for a particular circuit it is usual to provide between three and five frequencies to be used throughout the sunspot cycle spanning a frequency interval of about 2 octaves. For example to provide coverage from about 6 to 24Hz in a given direction would require at least two rhombic antennas, the larger of which, in meeting the optimum specification given above would have a side length of 250 m and a height of 75 m.

In estimating the gains of rhombics most authors have made the
**Fig. (A2.2.2)** Rhombic Antenna: Theoretical Free-Space Gain Relative to Half-Wave Dipole

**Fig. (A2.2.3)** Free Space Rhombic, Contours of Maximum Gain.
simplifying assumption that the current distribution along the
antenna is uniform, as does the discussion so far in this section.
It is generally known however that the current suffers progressive
radiation attenuation and at the termination may be 0.7 to 0.5 of the
initial value. This has the effect of reducing the gain relative
to the theoretical value by an amount depending mainly upon the size
of the antenna in terms of wavelengths. Attempts have been made to
calculate this effect with various assumed distributions of current.\textsuperscript{12,13}
That probably nearest to the truth is based upon the measured efficiency
and a linearly tapered current distribution and leads to a correction
which increases with length as shown in figure (A2.2.5). The poor
use of available transmitter power by the rhombic is often criticised.
Admittedly, it would be advantageous if an increase of several dB, in
received signal power could be obtained from a transmitting rhombic
by reducing the current flowing in its terminating impedance to zero.
This is not how the rhombic operates however, and for a point-to-point
service it is not a matter of great importance because of the high
powers normally involved. Indeed for the same gain and application, the
width of the main beam of a rhombic antenna is generally less than
that of most other types of arrays used in HF communications, and the
fact that the unused power is absorbed rather than radiated in unwanted
directions could be construed as a positive advantage. The significant
point is probably only fully appreciated when the economics of providing
a given overall grade of service, using different types of aerials, are
analyzed. Such analyses generally indicate that there is a clear economic
advantage in using rhombic antennas rather than any other for HF fixed
services.\textsuperscript{14}

The rhombic is also used for reception and in general the design
considerations are similar to those involved in the transmitting case.
HEIGHT ABOVE GROUND (W/A)

FIG. (A2.2.4) HORIZONTAL POLARIZATION FACTOR.

CORRECTION (dB)

SIDE LENGTH (W/A)

FIG. (A2.2.5) POWER GAIN CORRECTION FOR INEFFECTIVENESS OF RHOMBIC ANTENNA.
There are however some important differences.

Firstly the directivity gain of the rhombic used as a receiving antenna is higher than the power gain of the same antenna used for transmission. When transmitting, the power delivered to the aerial termination represents inefficiency, whereas when receiving both wanted and unwanted signals are affected similarly. The power of the antenna to discriminate against interference is a property of its polar diagram, and is unaffected by its efficiency. The receiving rhombic is an example of the essential dissimilarity of the current-field relationships in transmitting and receiving antennas. As far as the far-field effect is concerned, the similarity holds for the polar diagrams are the same in both cases.

Because inefficiency does not degrade the SNR, it was common for many years to employ single wire rhombics. The impedance characteristics of these antennas are very erratic and their front-to-back ratios are relatively poor (10-14dB) whereas multiwire antennas readily give ratios of 15-22dB with comparable reduction in response over the whole backward area. They are also claimed to be 9dB less sensitive to static interference. A properly constructed but complex reflector array will have a superior ratio, of some 20-25dB, but such performance is only of importance should a station be sited on the outskirts of a town or other source of radio noise.

If each end of a rhombic is connected to a transmission line, the arrangement is then symmetrical and is capable of simultaneous operation in both directions. The saving in site space is of great importance, especially where spaced aerial diversity is used, which is the invariable practice in modern machine-speed telegraphy systems. The separation between the antennas in a space diversity system must be sufficient to ensure a low phase correlation between the fading envelopes of their
output signals. Certain measurements indicate that there is negligible
phase-coherence between aerials spaced by more than 360 m along the
wave arrival direction and 260 m at right angles to the arrival bearing.

The loss of coherence from which the diversity advantage arises
allows considerable flexibility in the siting of diversity pairs of
aerials. This same mechanism might appear to place a limit on the
size of an individual receiving antenna, since the coherent integration
process upon which the directivity gain depends, becomes less efficient
as its aperture, including that of its earth image, increases. From the
standpoint of mean SNR, this may well be so, but there is little doubt
that a coherent element is present in nearly every signal and that a
large aerial having a polar response matched to this coherent element
will show advantages over a smaller antenna irrespective of signal
gain. Experience to date supports the view that an upper limit of
useful size for an HF receiving aerial has yet to be reached.

Flexible interconnections between aerials and receivers are
easily arranged and various methods are available for feeding several
receivers simultaneously from one aerial. For this reason, complete
world wide coverage has sometimes been specified. A rhombic ring is
an attractive way of meeting such a requirement with the minimum of
space and support cost.

Fifteen pairs of double-ended rhombics in an approximately circular
ring of 600 m mean diameter can provide dual diversity in over 30
directions. Although these may be chosen so as to cover the more
important directions, provided that the horizontal beamwidth of each
antenna is 15° at 25 MHz, then virtually complete world coverage is
secured apart from certain unimportant oceanic regions. Such a system
has in fact been in operation at Bearley radio station since 1952.
The station was rebuilt and reopened in November 1967 and the aerial system was retained. However the 1952 system provided coverage only down to 10MHz. Due to the existence of certain important routes below this frequency; down to about 6MHz or below, additional 150 foot high rhombics have been provided. These new antennas in conjunction with the original ring give good performance between 6 and 25MHz and an acceptable performance over the entire usable HF system. Figure (A2.2.6) shows the antenna system.

The desirability of achieving discrimination against interference in an antenna has been mentioned already. But the general case of an interfering emission is not so easy to deal with as that of interference from a fixed source such as a town. Automatic requests for repetitions systems (ARQ) reduce the effect of such emissions provided that they are sporadic. On the other hand every reasonable effort must be made to improve the performance of both transmitting and receiving antennas in reducing their radiation or sensitivity in unwanted directions.

The rhombic is frequently criticized in this respect, and not without some force. The maxima of the subsidiary lobes of a practical rhombic are some 6dB larger than those of a broadside array of similar gain. On the other hand if interference is considered uniformly distributed in the azimuth, then the rhombic antenna will have the same noise energy pick-up as any other aerial. As both noise and interference will be reduced by any aerial inefficiency, the SNR of the rhombic will be the same as for any aerial of the same directivity gain. Generally speaking the rhombic will pick up more noise in its subsidiary lobes and less from its main lobe than an array.

On the other hand, it will be appreciated that, particularly in telegraphy systems which make up a large percentage of modern HF systems,
FIG. (A2.2.6). RHOMBIC RING AT BEARLEY.
interference occurs when the unwanted signal exceeds a relatively well defined threshold. It follows that the levels of the maxima of the sidelobes are then important, and any reasonable effort to reduce them should be made. Several steps are immediately possible. Careful design can produce total or partial extinction of the sidelobe in the rhombic factor by a null in the height factor, although at the sacrifice of some bandwidth.\(^\text{16}\) The end-fire interlaced pair is the best of several forms of rhombic array which has been proposed.\(^\text{17}\) It reaches a performance equal to that of the largest uniform arrays and considerably reduces the sidelobe level.

Based on the above, it is apparent that the optimum design for the rhombic antenna must be based on fairly clear cut parameters.

Among the most important are:

1. It must be long in terms of wavelength.
2. It must have a large half angle.
3. It must have small angles of fire.

Optimally, also it should have a multiwire construction. The 'curtain' rhombic consists of two or more wires per side and has the advantages of reducing the average single impedance and the termination loss, and in producing a more constant input impedance characteristic over a wide frequency band. This type of construction also reduces the precipitation static when the antenna is used in its receiving mode. The 'curtain' rhombic is used in the large majority of modern day applications because of these advantages over the single wire rhombic. In practice, the curtain rhombic will have an additional 2-3 dB of gain over the single wire case.

Be this as it may, the equations used to define the radiation patterns for the rhombic antenna are all the same. In this investigation those used by CCIR in their publication 'Antenna Diagrams' will be used.
These were published in 1953 and the equation given was incorrect by a multiplicative factor of \(1/(\pi L)^2\). This was corrected by private communication between the author and CCIR in 1973. The corrected equation is

\[
F = \frac{L^2}{(rL)^2} \cdot 2400 \cdot I^2 \cdot L^2 \cdot \cos^2 \gamma \cdot \sin^2 \left( \frac{\pi K_1}{L} \right) \cdot \sin^2 \left( \frac{\pi K_2}{L} \right) \cdot (2\sin(2\pi \sin \theta))^2
\]

(A2.2.2)

which reduces to

\[
F = \frac{2400}{\pi} \cdot \cos^2 \gamma \cdot \sin^2 \left( \frac{\pi K_1}{L} \right) \cdot \sin^2 \left( \frac{\pi K_2}{L} \right) \cdot (2\sin(2\pi \sin \theta))^2
\]

(A2.2.3)

where \(F\) - the radiation intensity at 1 m in W/m²

\(I\) - the transmitting current in amps (rms)

\(L\) - length of one side of the rhombic in wavelengths

\(\gamma\) - one half the interior side angle.

\(\phi\) - the horizontal (azimuth) angle between a direction under consideration and the longer diagonal of the rhombic. \(\phi\) is 0° at the centre of the main lobe and 180° at the centre of the back radiation.

\(\theta\) - the vertical angle (elevation) between a direction under consideration and the plane of the rhombic.

\(H\) - height of the rhombic above a perfectly conducting earth.

\(K_1\) - 1 - \(\cos \theta \sin(\gamma + \phi)\)

\(K_2\) - 1 - \(\cos \theta \sin(\gamma - \phi)\)

This equation takes into account both the horizontal and vertical polarizations. It is based on the assumption of uniform progressive waves and a terminating resistance at the forward end of the antenna equal to its characteristic impedance. The magnitude of the radiation pattern or power gain may be obtained from equation (3.4.3) if the current \(I\) is known. The value of the current can be computed from the radiation resistance by

\[
I = \sqrt{\frac{P}{2\pi R}}
\]

(A2.2.4)
where $Pr$ is the power radiated. An approximate expression for the radiation resistance of an isolated rhombic antenna which provides focal accuracy for side lengths greater than $2A$ is given by Lerwin\textsuperscript{19} as

$$Rr = 240 \left( \log_\text{e} \left( \frac{\pi l}{A} \right) \sin^2 A + 0.577 \right) \Omega \quad (A2.2.5)$$

The exact equation for computing the radiation resistance involving tabulated functions has also been evaluated by Lerwin\textsuperscript{19} and confirmed and presented in a different form by Chancey\textsuperscript{20}.

The flow chart for computer programme based on the above equation is given in figure (A2.2.7) and the programme in Fortran IV is given below. Sample printout from this programme is also given. In this programme the value of 600 $\Omega$ is used for $Rr$ if $L$ is less than $2A$.

The final antenna of interest to this investigation is the log periodic dipole array - a comparative, but very important, newcomer to the field of HF radio communication.
Figure (A2.2.7) Flow Chart for Rhombic Programme

START

read data as follows. POWER - transmitter power
AL - side length of rhombic in wavelengths
GAMMA - angle between minor axis and side (degrees)
H - height of rhombic above ground (wavelengths)
ABR - bearing of aerial E of North (degrees)
BRTTOR - bearing of transmitter to receiver (degrees)

Evaluate the azimuth angle

\[ \Phi = \frac{1}{2} (ABR - BRTTOR) \]

if AL > 2 evaluate radiation resistance by Lerwin's formula

\[ R_r = 600 \Omega \]

Evaluate current \( i \) given in equation (3.4.4)

Set elevation angle \( \theta = 1^\circ \)

Evaluate \( K_1 \) and \( K_2 \) in equation (3.4.3)
Calculate $F$ - the field intensity and PRINT

$\theta = \theta + 1$

1. **Yes** $\theta < \theta_{\text{max}}$  $\theta_{\text{max}}$ is fixed arbitrarily say $60^0$

   **No**

   STOP

*Figure (A2.2.7) Flow Chart for Rhombic Programme*
LE IDENTIFIER: RHOMBIG

PI = 3.142
RDNS=57.2957795131

READ, POWER
10 READ, GAMMA, EMAX, H, R
READ, AL, GAMMA, H, ABR
WRITE (6, 20) AL, H, GAMMA, EMAX
20 FORMAT( '0', 'RHOMBIG ANTENNA - L=', F3.1, ', H=', F3.1, ', GAMMA=', F4.1, ', EMAX=', F7.2)

PHI = ABS (ABR - BRTTOR)

RADIATION RESISTANCE BY LERWIN'S FORMULA - L < 2 * LAMDA.
IF (AL LT 2.) GO TO 30
ANLFAC = ALOG(4. * PI * AL * (SIN((90. - GAMMA) / RDNS) ** 2))
RADR = 240. * (ANLFAC + 0.577)
GO TO 40
30 RADR = 600.
40 CONTINUE

DO GO 60 IDelta = M, N
DELTA = IDelta

C1 = 1 - (COS(DELTA/RDNS) * SIN((GAMMA + PHI)/RDNS))
C2 = 1 - (COS(DELTA/RDNS) * SIN((GAMMA - PHI)/RDNS))
CIANG = ((SIN(PI * AL * C1)**2) / C1
C2ANG = ((SIN(PI * AL * C2)**2) / C2

HFAC = ABS(2. * SIN(2. * PI * H * SIN(DELTA/RDNS)) ** 2)
DIV = (240. / PI) * CIANG * C2ANG * HFAC * (COS(GAMMA/RDNS) ** 2)

FIR = EMAX
RADINT = (FIR ** 2) / (240. * PI)
AMP = (RADINT / DIV) ** 0.5
POURAD = (AMP ** 2) * RADR
FACT = POURAD / (POWER ** 1000.)

WRITE (6, 70) RADINT, FACT, AMP, RADR, FIR, POURAD
FIR=', F7.1, ' POURAD =', F7.1)

READ, NFLAG
IF (NFLAG.EQ.0) GO TO 80
GO TO 10
80 CONTINUE
STOP
END
<table>
<thead>
<tr>
<th>DATA:</th>
<th>340.0</th>
<th>0.57</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA:</td>
<td>6.60</td>
<td>1.20</td>
</tr>
</tbody>
</table>

**RHOMBIC ANTENNA** - \( L=6.0 \) \( H=1.2 \) \( \text{GAMMA}=60.0 \) \( \text{FMAXIM}=340.00 \)

<table>
<thead>
<tr>
<th>( F )</th>
<th>153.30</th>
<th>( \text{FACT}=1.117 )</th>
<th>( \text{AMP}=1.15 )</th>
<th>( \text{RR}=843.27 )</th>
<th>( \text{FIR}=340.0 )</th>
<th>( \text{POWRAD}=1117.0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F )</td>
<td>153.30</td>
<td>( \text{FACT}=1.016 )</td>
<td>( \text{AMP}=1.10 )</td>
<td>( \text{RR}=843.27 )</td>
<td>( \text{FIR}=340.0 )</td>
<td>( \text{POWRAD}=1016.0 )</td>
</tr>
<tr>
<td>( F )</td>
<td>153.30</td>
<td>( \text{FACT}=1.043 )</td>
<td>( \text{AMP}=1.11 )</td>
<td>( \text{RR}=843.27 )</td>
<td>( \text{FIR}=340.0 )</td>
<td>( \text{POWRAD}=1043.2 )</td>
</tr>
</tbody>
</table>

**DATA:**

<table>
<thead>
<tr>
<th>DATA:</th>
<th>1885.0</th>
<th>0.810</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA:</td>
<td>6.65</td>
<td>1.20</td>
</tr>
</tbody>
</table>

**RHOMBIC ANTENNA** - \( L=6.0 \) \( H=1.2 \) \( \text{GAMMA}=65.0 \) \( \text{FMAXIM}=1885.00 \)

<table>
<thead>
<tr>
<th>( F )</th>
<th>4712.00</th>
<th>( \text{FACT}=1.203 )</th>
<th>( \text{AMP}=1.26 )</th>
<th>( \text{RR}=762.56 )</th>
<th>( \text{FIR}=1885.0 )</th>
<th>( \text{POWRAD}=1292.6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F )</td>
<td>4712.00</td>
<td>( \text{FACT}=1.189 )</td>
<td>( \text{AMP}=1.25 )</td>
<td>( \text{RR}=762.56 )</td>
<td>( \text{FIR}=1885.0 )</td>
<td>( \text{POWRAD}=1188.4 )</td>
</tr>
<tr>
<td>( F )</td>
<td>4712.00</td>
<td>( \text{FACT}=1.252 )</td>
<td>( \text{AMP}=1.28 )</td>
<td>( \text{RR}=762.56 )</td>
<td>( \text{FIR}=1885.0 )</td>
<td>( \text{POWRAD}=1281.4 )</td>
</tr>
</tbody>
</table>

**DATA:**

<table>
<thead>
<tr>
<th>DATA:</th>
<th>2390.0</th>
<th>0.10</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA:</td>
<td>6.70</td>
<td>1.20</td>
<td></td>
</tr>
</tbody>
</table>

**RHOMBIC ANTENNA** - \( L=6.0 \) \( H=1.2 \) \( \text{GAMMA}=70.0 \) \( \text{FMAXIM}=2390.00 \)

<table>
<thead>
<tr>
<th>( F )</th>
<th>7574.93</th>
<th>( \text{FACT}=0.888 )</th>
<th>( \text{AMP}=1.16 )</th>
<th>( \text{RR}=660.99 )</th>
<th>( \text{FIR}=2390.0 )</th>
<th>( \text{POWRAD}=881.8 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F )</td>
<td>7574.93</td>
<td>( \text{FACT}=0.878 )</td>
<td>( \text{AMP}=1.15 )</td>
<td>( \text{RR}=660.99 )</td>
<td>( \text{FIR}=2390.0 )</td>
<td>( \text{POWRAD}=878.3 )</td>
</tr>
<tr>
<td>( F )</td>
<td>7574.93</td>
<td>( \text{FACT}=0.920 )</td>
<td>( \text{AMP}=1.18 )</td>
<td>( \text{RR}=660.99 )</td>
<td>( \text{FIR}=2390.0 )</td>
<td>( \text{POWRAD}=919.6 )</td>
</tr>
</tbody>
</table>

**DATA:**

| DATA: | 0     |      |
A2.3 The Log Periodic Dipole Array (LPDA)

On circuits operating over less than 4000 Km, the angle of elevation for the most favourable propagation path varies considerably according to the distance, time of day and the other vagaries which affect HF communication. A more flexible system than the rhombic antenna is thus advantageous. The required performance has been achieved in the past by use of a number of resonant arrays of dipoles to cover the necessary frequency range and with the aerial height chosen for maximum radiation in a particular direction. In the light of present day practice, the disadvantage of this system is that a considerable number of separate aerials are necessary for complete coverage of the frequency spectrum. This results in a complex aerial switching and tuning arrangement with the transmitter. Ideally a single antenna is required which has wideband characteristics to suit the loading requirements and also with acceptable radiation characteristics. A solution has been found with the development of the LPDA.

The LPDA consists of several connected structures all similar in the geometrical sense, and each bearing a constant scaling ratio to its neighbour. The geometry of this structure thus repeats itself periodically with the logarithm of the distance from the implied vertex. The performance of the aerial varies periodically with the logarithm of frequency. Provided that this variation in performance is small over one period, the aerial behaves as a wideband device. The LPDA is configured as shown in figure (A2.3.1).

Operation of the LPDA is best discussed by considering the propagation of a wave along the feeder starting at the feed point, at a frequency at which the central dipoles are approximately resonant. The smallest dipoles introduce loading on the feeder by interfering
FIG. (A2.3.1)  TYPICAL LOG-PERIODIC ARRAY OF DIPOLES.
destructively with one another to some extent and so radiate little energy. This is the transmission region of the array. Adjacent to this is the active region in which most of the radiation takes place. The reversal of the feeder between successive dipoles produces radiation in the backward direction and the current in the feeder decays rapidly. Finally there is the unexcited region, containing the largest dipoles which are excited neither by the feeder nor the radiated energy. Since the largest dipoles do not contribute to the radiated energy, the truncation of the log-periodic principle inherent in having an aerial of finite size is not fundamentally damaging.

The number of active elements satisfying the resonant requirement for radiation at a given frequency depends upon the taper. The overall bandwidth of the array is set by the lengths of the shortest and longest elements. The gain of the array is also set by these conditions since the active elements generate the components which combine to produce the resultant field. Moreover, since the input power is not necessarily divided up equally between these elements, and since their mutual coupling is great (due to their close proximity) it is a very complex problem to evaluate theoretically the gain and directivity of the LPDA. It is thus normal practice to investigate the characteristics of a small scale model experimentally at the appropriate higher frequency, before building the full scale array.

An LPDA designed for HF communication should operate from 4 to 27.5 MHz and have as high a gain as is economically possible. The elements are normally made of wire due to the large physical size of these arrays. Simple dipoles with high q's and narrow bandwidths are thus unsuitable for HF LPDA's. It is preferable to use wideband folded composite dipoles in order to keep the total number of elements as small as possible for a required gain and given overall bandwidth.
The narrower the response of an individual radiator, the more gradual will be the taper.

A typical array of 47 composite folded dipoles is shown in figure (A2.3.2). The polar diagram (in the vertical plane) shown in figure (A2.3.3) is kept constant over the required frequency range by successively increasing the height of a radiator above the ground according to its resonant length, thereby keeping the electrical height (in wavelengths) constant along the array.

A uniform taper of the element length gives a uniform gain, with respect to frequency of 15dB compared to the isotropic radiator. The main beamwidth in the azimuthal plane is 48° measured at the 3dB points. The side and back lobe level is on average 22dB down on the main lobe over the frequency range. This figure depends heavily upon the matching between the feeder and radiating elements; the better the match, the lower is the radiation from the feeder and consequently the smaller is the sidelobe level.

The mathematical analysis of the LPDA is complex which makes manual solution a very long and tedious operation. Most modern-day analyses are computer-based and require large amounts of space and time on the computer.

Initial studies were carried out by Du Hamel and Isbell and Isbell in 1957 and 1960 respectively. Also in 1960 Carrell produced the first fully comprehensive mathematical analysis and design which has served as a guide to all analyses since then. This analysis took into account the mutual coupling between dipole elements and presented a step-by-step procedure for the design of an LPDA for a wide range of input impedances, gains, bandwidths and sizes, with the facility to varying voltages and currents at the input terminals of the elements and any feed line section. Based on this analysis and
FIG. (A2.3.2) LOG PERIODIC DIPOLE ARRAY

FIG. (A2.3.2) VERTICAL POLAR DIAGRAM
FOR LPDA IN FIG. (A2.3.2)
his own work, C.E. Smith produced a handbook\textsuperscript{25} for the design of LPDAs. Later workers\textsuperscript{26,27,28} have found however that considerable differences exist between the theory of Carrell and Smith and practical log-periodic antennas. For example, gain degradations ranging from 2 to 3 dB, as well as strong deformations of the radiation patterns and of the VSWR have been found in practice. The need for more accurate design procedure, particularly of large high power HF antennas with high impedance feed lines appeared obvious.

The main criticism of Carrell's analysis was his use of sinusoidal current distribution in the dipoles. Although in a single dipole antenna the sinusoidal distribution has a negligible effect on the gain computation, the situation may be substantially different in the LPDA, where many dipoles contribute to the radiation field and where amplitude and phase of currents at the dipole inputs depend on the true impedances presented to the feed line and on the mutual impedance.

With this in mind, King, Cheong and other workers\textsuperscript{26,29,30,31} developed a new current distribution obtained by summing only three sinusoidal terms. This gives accurate results when $2h/a > 1$, where $h$ is the half length of the dipole and 'a' is its radius, the ratio $h/a$ is constant for the dipoles used in LPDA.

The current distribution used recently may be written as, for the $n$\textsuperscript{th} dipole of an LPDA,

$$\text{In}(z) = I_{An} \sin \beta h - (z)0 + I_{Bn} (\cos^2 - \cosh h) + I_{cn} (\cos h - \cosh h)$$

(A2.3.1)

where $\beta$ is the phase propagation constant and the $n$\textsuperscript{th} dipole is assumed to be oriented along the $z$-axis of the $xyz$ coordinate system. Using this equation and the fact that the vector potential, $A_{zn}$, on the $n$\textsuperscript{th} dipole may also be expressed as a function of two sinusoidal terms
of amplitude $A_{1n}$ and $A_{2n}$, the integral equation which relates $I_n(z)$ on the various dipoles and $A_{zn}(z)$, is transferred in a set of algebraic equations. For example, one may equate the $A_{zn}$ values to the values of the integral expression at four different values of $z$. The unknowns are the current amplitudes $I_{1n}$, $I_{Bn}$ and $I_{Cn}$ and the amplitude of $A_{1n}$ of the first term of $A_{zn}$; the amplitude of $A_{zn}$ of the second term may be found as a function of the voltage $V_0$ at the input.

This procedure allows the transformation of the system of $N$ integral equations, which has to be solved in this case, into a system of $4N$ algebraic equations with $4N$ unknown variables. (cf ref; 26)

The method for evaluating the current distribution above is an application of the method of moments described by Harrington in 1966 \textsuperscript{32} and 1967 \textsuperscript{33}. It in effect produces a more accurate impedance matrix for the LPDA elements than did Carrell's analysis. Besides this, the current distribution as a sum of three sinusoidal terms, produced even more accuracy.

As seen above, the mathematics is fairly complex and so most of the methods are set up on digital computers. The programmes however are large and time consuming because of the number of equations to be solved. Recently however a new method of solution has been suggested \textsuperscript{34} by Sinnott which uses the periodicity of the LPDA to enable an iterative matrix solution of the equations when dealing with many frequencies.

Because of the large amounts of computer space and time, it did not appear justifiable to incorporate a full programme for LPDA analysis in this investigation. However a programme was written which produced an analysis of the LPDA using the method of moments to find the impedance matrix of the LPDA, and then used Carrell's method to find the current distribution on the elements. This in effect produced an
analysis similar in type to the more accurate methods described above. As expected however the programme was very expensive on computer space and time and it was finally decided to use published figures for LPDA gain and other relevant parameters as the need arose. The programme is given below.
**E.R.C.C. FORTRAN COMPILER RELEASE 5 VERSION 4 DATED 15/02/74**

1 C THIS PROGRAM PRODUCES AN ANALYSIS OF THE LOG-PERIODIC
2 C DIPOLE ANTENNA USING THE METHOD OF MOMENTS TO FIND
3 C THE IMPEDANCE MATRIX OF THE ANTENNA OF HARRINGTON.
4 C IT THEN USES CARREL'S METHOD FOR FINDING THE CURRENT.
5 C DISTRIBUTION ON THE ELEMENTS AND OTHER PARAMETERS
6 C OF INTEREST, ALSO INCORPORATED IS A MULTIPLE FREQUENCY
7 C OPTION DUE TO SIGNAL AND THE WHOLE ANALYSIS IS DONE IN
8 C DOUBLE PRECISION ARITHMETIC FOR MAXIMUM ACCURACY.
9 C
10 C PARAMETERS USED IN THIS PROGRAM ARE:
11 C
12 C ARRAYS:
13 C
14 C H,A,B,---HEIGHTS, RADIUS, DISTANCES
15 C Z,PHI,---DYNAMIC ARRAYS FOR INTERMEDIATE
16 C VALUES IN MOMENT ANALYSIS.
17 C PHI,---HOLDS DYNAMIC VALUES OF IMPEDANCE
18 C ELEMENTS IN MOMENT ANALYSIS.
19 C ZIM---IMPEDANCE MATRIX OF ELEMENTS FOR
20 C ANY GIVEN FREQUENCY.
21 C FEEDAD---ADMITTANCE MATRIX OF THE FEED LINE
22 C TO THE ANTENNA.
23 C UFELEZ---THE VALUE OF MATRIX T IN CARREL'S
24 C ANALYSIS.
25 C ECHAMP---INPUT CURRENT MATRIX OF COLUMN
26 C FORM.
27 C ELAMP---CURRENT DISTRIBUTION MATRIX FOR
28 C ANTENNA.
29 C ELVOL---VOLTAGE DISTRIBUTION MATRIX FOR
30 C ANTENNA.
31 C HKSP, HKSP---DYNAMIC WORKING ARRAYS TO HOLD
32 C VALUES MAKING UP LARGER EXPS.
33 C
34 C VARIABLES:
HI1, HI2 ——— VALUES OF LENGTHS ON ELEMENT I BEING
CONSIDERED IN MOMENT ANALYSIS.
HK1, HK2 ——— SIMILAR VALUES ON ELEMENT K.
DL1, DLK ——— CONTAIN VALUES OF OPERATIONS ON H VALUES.
PHG, BS0 ——— DYNAMIC VARS WITH VALUES OF 8, 9**2.
MMn, EP ——— MU AND EPSILON CONSTANTS OF PROPAGATION
MEDIUM.
BOH, ONE ——— PROPAGATION CONSTANT.
PI ——— 3.141593
SIG ——— SPACING FACTOR FOR ANTENNA.
TAU ——— SCALE FACTOR FOR ANTENNA.
HAFAC ——— RATIO OF ELEMENT HEIGHT TO RADIUS.
FR ——— FREQUENCY OF OPERATION.
YC ——— CHARACTERISTIC ADMITTANCE OF ANTENNA.
X, Y ——— VALUES OF ZERO AND ONE FOR GENERAL USE.
CRTH1, 2, 3 ——— COMPLEX CONSTANTS FOR GENERAL USE.
DCMPLX ——— COMPLEX OPERATOR IN DOUBLE PREC. MATHS.
N ——— NUMBER OF ELEMENTS ON ANTENNA.
M ——— ELEMENT AT WHICH ANALYSIS BEGINS.
MY ——— EQ. N FOR FIRST FREQUENCY AND THEREAFTER
EQ. 1 CF. SINTON’S PAPER.
NY ——— NO OF FREQUENCIES AT WHICH ANTENNA IS TO
BE ANALYSED.
H(n) ——— VALUE OF SMALLEST ELEMENT OF ANTENNA.
IPJ, IPO ——— NUMBER OF SEGMENTS EACH HALF ELEMENT IS
DIVIDED INTO FOR MOMENT ANALYSIS.

SUBROUTINES USED:
FIFUNC ——— THIS ROUTINE EVALUATES THE VALUES TO BE
USED IN THE ARRAY PHI.
YFEED ——— THIS EVALUATES THE ADMITTANCE MATRIX OF
THE FEED LINE.
UZIFY ——— THIS EVALUATES THE MATRIX I OF CARREL'S
ANALYSIS.
FO4ADP ——— SOLVES THE SIMULTANEOUS EQU.
KOUNT = 1 CALLS WRITE, KOUNT = 0 CALLS LPDFI.
KOUNT = OTHER SUPRESSES ABOVE ROUTINES.

REAL*8 H(12), A(12), B(12, 12), WKS(36), RHG, BS0,
>HI1, HI2, MY1, HK1, JLI, DLK, Z(5), R(5), FR,
>Y, X, BI, PI, UN, EP, SIG, TAU, HAFAC, ONE, X, Y,

COMPLEX*16 PHI(5), ZMP(36, 36), WSPC(5), DCMPLX,
>YFEED(12, 12), UFEED(36, 36), CRTH1, CRTH2, CRTH3, ENAMP(36, 1),
>ELAMP(36, 1), ELVOL(36, 1)

READ, H, MY, X, Y, KOUNT
READ, H(N), HAFAC, TAU, SIG, IPJ, IPO, FR, YC

X = 0.0
Y = 1.0
CRTH1 = DCMPLX(X, Y)
CRTH2 = DCMPLX(Y, X)
CRTH3 = DCMPLX(X, X)
N = N * IPJ
PI = 3.141593
IND = 4.0 * PI * (1./10.**2)
EP = (1./36.0 * PI * 10.**2)
H(1) = H(1)/(TAU**2*(N-1))

DO 5 I = 1, NO
5 ENAMP(I, 1) = CRTH3
ENAMP(NO-7, 1) = CRTH2
PROGRAM IS NOW INITIALIZED.

DO 10 I = 1, N
10 IF(I.EQ.1) GO TO 10

H(I+1) = H(I) * TAU

R(I, (I+1)) = 4.*SIG*N(I)

B((I+1), I) = B(I, (I+1))

10 B(I, 1) = H(I)/HAFAC

HI = N-2
DO 20 I = 1, HI
IM = I+2
DO 20 K = IM, N
B(I, K) = B(I, (K-1)) + B((K-1), K)
20 B(K, I) = B(K, K)

II = IPJ + 1
JK = IPQ + 1

HEIGHTS OF ELEMENTS AND DISTANCES BETWEEN THEM NOW EVALUATE

DO 400 IFR = 1, NY
WRITE(9, 4000) FR
BOM = 2. * PI * FR/300.
ONE = BOM * 300. * 10.**6

DO 40 I = MY
LR = 3 * (I-1)
DO 40 K = 1, N
LC = 3 * (K-1)
RHO = B(I, K)
BSU = RHO ** 2

DO 40 LI = 1, I!
LL = LR + LI
IF(LI.EQ.II) GO TO 40
HII = H(I) *(LI-1.)/IPJ
H12 = H(II) * LI/IPJ
DL1 = (H12 - HII)/2.
400

DO 40 LK = 1, IK
KK = LC + LK
IF(LK.EQ.IK) GO TO 40
HK1 = H(K) *(LK-1.)/IPQ
HK2 = H(K) * LK/IPQ
DLK = (HK2 - HK1)/2.

Z(1) = (HK1 - HII + DK1 - DL1)
Z(2) = (HK1 + HII)
Z(3) = (HK2 - H12)
Z(4) = (HK1 - H12)
Z(5) = (HK2 - H12)

DO 30 IE = 1, 5
30 RM(IE) = (BS3 + Z(IE)**2)**0.5

CALL FIFUNC (Z, PH, RHO, DL1, S, PHI, FR, BOM, PI)

WKSPC(1) = 4.*ONE*SUMUM*DLK*DL1*PHI(1)
WKSPC(2) = (PHI(2)+PHI(3)-PHI(4)-PHI(5))/ONE*EP
WKSPC(3) = WKSPC(1) - WKSPC(2)

ZIMP(LL, KK) = WKSPC(3) * CRTM1
ZIMP(KK, LL) = ZIMP(LL, KK)

CONTINUE
CALL YFEED (FEEDAD, B, N, YC, YC, BON, CRTM1, CRTM2) (d)
CALL UZITE (ZIMP, FEEDAD, N, IO, IPJ, IPQ, UFEEZ, CRTM1) (d)

C  IL = 0
C  IF1 = 1
C
CALL F04ADF (UFEEZ, MO, ENAMP, MO, MO, IF1, ENAMP, MO, WKSP, IL)
C
IF (KOUNT, IE, 0) GO TO 95
C
CALL LDPI (ONE, MM, PI, BON, NO, N, IPJ, H, B, ELAMP)
C
CONTINUE
IF (KOUNT, EQ, 1) CALL WRITEN (MO, N, ZIMP, IPJ, IPQ)
C
DO 200 IR = 1, NO
ELVOL(IR, 1) = CRTM3
DO 100 IC = 1, NO
100 ELVOL(IR, 1) = ZIMP(IR, IC) * ELAMP(IC, 1) + ELVOL(IR, 1)
C
WRITE (9, 3000) ELVOL(IR, 1), ELAMP(IR, 1)
C
WX = NO-IPJ
DO 300 I = 1, MX
MIB = 1 + (WX-1)
DO 300 K = 1, MX
300 ZIMP(MIB+IPJ), (K+IPJ)) = ZIMP(MIB, K)
FR = FR/TAU
400 NY = 1
C
3000 FORMAT (' ', 2D12.4, 2D12.4)
4000 FORMAT (' ', 'OPERATING FREQ. = ', F7.2)
C
STOP
END

SUBROUTINE FIFUNC(Y, RMS, ROW, ALPHA, N, FIE, FR, BO, PI)
C
THIS SUBROUTINE EVALUATES THE PHI(M, N) FUNCTION OF
HARRINGTON'S METHOD OF MOMENTS ANALYSIS.
C
REAL*8 Y(N), RMS(N), ROW, ALPHA, WKSP(20), DCOS, DSIN,
> DLOG, 20, PI, FR
C
COMPLEX*16 BOLR, FIEF, FIE(N), DCMPLX
C
IF (ROW, GE, (10, *ALPHA)) GO TO 200
C
THE FOLLOWING SECTION GIVES ACCURACY OF BETTER THAN 1
FOR RMS LESS THAN 10*ALPHA.
C
WKSP(2) = 2, * ALPHA
WKSP(5) = 8, * PI * ALPHA
WKSP(15) = ROW**2
C
DO 100 I = 1, N
100 IF (Y(I), LT, 0.) Y(I) = -Y(I)
WKSP(14) = Y(I) + ALPHA
WKSP(13) = Y(I) - ALPHA
C
WKSP(12) = (WKSP(15) + WKSP(14)**2)**0.5
WKSP(11) = (WKSP(15) + WKSP(13)**2)**0.5
WKSP(10) = BO + RMD(1)
WKSP(9) = WKSP(14) + WKSP(12)
WKSP(8) = WKSP(13) + WKSP(11)
C
WKSP(1) = DLOG(WKSP(9)/WKSP(3))
SUBROUTINE YFEED(ADIMF, D, N, YD, YT, BO, CRT1, CRT2)

C

C THE FOLLOWING SECTION GIVES ACCURACY OF BETTER THAN 1%
C FOR RMD GREATER THAN TEN TIMES ALPHA.

C

200 WKS(20) = 60 * ALPHA

DO 300 I = 1, N

WKS(19) = ALPHA/RMD(1)

WKS(18) = WKS(19)**2

WKS(17) = WKS(19)**3

WKS(16) = WKS(15)**2

WKS(15) = (Y(1)/RMD(1))**2

WKS(14) = WKS(15)**2

WKS(13) = (3. - 30.*WKS(15) + 35.*WKS(14))

WKS(12) = (3.*WKS(15) - 1.)

WKS(11) = 1./6.

WKS(1) = 1.+(WKS(11)*WKS(18)+WKS(12)) +

0.025*WKS(16)*WKS(13))

WKS(2) = (WKS(11)*WKS(19)+WKS(12)) +

0.025*WKS(17)*WKS(13))

WKS(3) = 0.9-(WKS(11)*WKS(15)-(0.025*WKS(18)*

1.0-12.*WKS(15)+15.*WKS(14))

WKS(4) = 0.1+WKS(11)*WKS(19)*3.*WKS(15)-5.*WKS(14))

WKS(5) = 0.05*WKS(11)*WKS(14)

WKS(6) = 4.*PI*RMD(1)

WKS(7) = 60 * RMD(1)

BOLR = DCMPLX(DCOS(WKS(7)), -DSIN(WKS(7)))

BOLR = BOLR/WKS(6)

WKS(8) = WKS(1) + ((WKS(2)**2)*WKS(3)) + (WKS(5)*

(WKS(2)**4))

WKS(9) = (WKS(20)*WKS(2)) + (WKS(4)*WKS(20)**3)

FIED = DCMPLX(WKS(6), WKS(7))

100 FIE(1) = BOLR * FIED

RETURN

END

SUBROUTINE YFEED(ADIMF, D, N, YD, YT, BO, CRT1, CRT2)
SUBROUTINE UZIF(Z,Y,N,M,IPR,IPC,UZIF,X)

C THIS SUBROUTINE EVALUATES THE PRODUCT OF
C THE ADMITTANCE AND THE ELEMENT IMPEDANCE
C MATRICES, AND THEN ADDS THE RESULT TO THE UNIT
C MATRIX. THIS PROCEDURE GIVES THE MATRIX T OF
C CARREL'S ANALYSIS OF THE LPDA.

C COMPLEX*16 X,Y(N,N),Z(M,M),UZIF(M,M)

MM = M-IPR
DO 20 I = 1,N
I1 = I+IPR
I2 = I-IPR
K1 = (I-1)/IPR
K2 = K1+I
K3 = K2+I
DO 20 K = 1,N
100 IF(L.EQ.MM) GO TO 200
UZIF(I,K) = Y(I,1)*Z(1,K) + Y(I,2)*Z(1,K)
GO TO 20
200 IF(L.LE.MM) GO TO 250
UZIF(I,K) = Y(I2,12)*Z(1,K) + Y(I2,11)*Z(12,K)
GO TO 25
250 UZIF(I,K) = Y(I2,K1)*Z(12,K) + Y(I2,K2)*Z(1,K)
> + Y(I2,K3)*Z(I1,K)
270 IF(L.EQ.K) UZIF(I,K) = UZIF(I,K) + X
280 RETURN
END
SUBROUTINE LPDFI(OMEGA,PHU,PI,B0,H0E,H0E,PJ,H,ELs,EAMP)

REAL*8 X,Y,OMEGA,PHU,PI,B0,DIST,THETA,PHI,RDNS,S1,Z1,
H(NOE),ELS(HOE,NOE),ARG,ARG1,ARG2,ENZI,EFZI,ENZI,DSIN
>DCOS,XR(5)

COMPLEX*16 CF,CS,CON,EXP1,EXP2,EXP3,SUM1,SUM2,Amp,
>FIELD,EAMP(HOE,1),DCMPLX

INTEGER PJ

RDNS = 57.29578
DIST = 1000.
JDL = PJ + 1
X = 0.0
Y = 1.0
CF = DCMPLX(X,Y)
CS = DCMPLX(X,Y)
CON = (CF*OMEGA*PHU)/(6.*PI)
ARG = B0+DIST
EXP1 = (1./DIST)*DCMPLX(DCOS(ARG),-DSIN(ARG))

DO 30 ITH = 1,181,10
THETA = (ITH-1.)/RDNS
WR(1) = DSIN(THETA)
WR(2) = DCOS(THETA)*BO

DO 30 PHI = 1,181,10
PHI = (ITH-1.)/RDNS
WR(3) = WR(1)*DCOS(PHI)

SUM2 = CS

DO 20 I = 1,NOE
IB = (NOE - I) + 1
IA3 = PJ*IB
IA2 = IA3 - 1
IA1 = IA2 - 1
SI = ELS(NOE,IB)*WR(3)
ARG = SI*B0
EXP2 = DCMPLX(DCOS(ARG),DSIN(ARG))
WR(4) = BO+HOE

SUM1 = CS

DO 10 IDE = 1,IDL
ZI = H(IB)*(IDE-1.)/PJ
ARG = ZI*WR(2)
EXP3 = DCMPLX(DCOS(ARG),DSIN(ARG))
WR(5) = BO+ZI
ENZI = DSIN(WR(4)-WR(5))
EFZI = DCOS(WR(5))-DCOS(WR(4))
ENZI = DCOS(WR(5)/2.)*DCOS(WR(4)/2.)
AMP = ENZI*EAMP(IA1,1)+EFZI*EAMP(IA2,1)+ENZI*EAMP(IA3,1)
IF(IDE,GT,1) AMP = 2.*AMP

10 SUM1 = SUM1 + (AMP*EXP3*WR(1))
20 SUM2 = SUM2 + (SUM1*EXP2)

FIELD = CON*EXP1*SUM2
WRITE(10,100) ITH,PJ,FIELD

CONTINUE

100 FORMAT(' ', 'THETA=',13,'PHI=',13,'FIELD=',13,'2010.3')

C RETURN
END
SUBROUTINE WRITE(M,N,ARRAY,1J,10)

C THIS ROUTINE WILL PRINT OUT THE CALCULATED VALUES
C OF THE IMPEDANCE AND ADMITTANCE MATRICES FOR
C THE LOG-PERIODIC DIPOLE ARRAY WHEN CALLED.

C COMPLEX*16 ARRAY(M,N)

DO 10 I = 1,N
  LR = 3*(I-1)
DO 10 K = 1,N
  LC = 3*(K-1)
WRITE(7,1000) I,K
DO 10 LI = 1,1J
  WRITE(7,2000) (ARRAY((LR+LI),(LC+LK)),LK=1,1PQ)
10 CONTINUE

1000 FORMAT(' ',214)
2000 FORMAT(' ',3D10.3,3D10.3)
RETURN
END

CODE+GLA+LODDATA+SYMTABS+ARRAYS = 10744+ 3483+ 408+ 96 + 47480= 62216 BYTE

COMPILATION SUCCESSFUL
References to Chapter 1

1. **BROWN, G N**  
   'Space Radio Communication'  
   Proc of the 1961 Symposium of the **IUCR** 1961  
   (Elsevier Publishing Company 1962)

2. **HAMSHER, D H**  
   'Communication System Engineering Handbook'  
   (McGraw-Hill Book Company 1967)

3. **BETTS J A**  
   'High Frequency Communications'  
   English Universities Press 1967

4. **SKOLNIK, M.I**  
   'Introduction to Radar Systems'  

5. **DAVIS, N G.**  
   **WOHLBERG, D.B**  
   'Spread Spectrum Techniques' in 'The Impact of New Technologies in Signal Processing'  
   IEE conf publ 144 pp 144-154  
   September 1976
References to Chapter 2


2. INTERNATIONAL RADIO CONSULTATIVE COMMITTEE 'CCIR Document of IXth Plenary Assembly', Vol 1 Recommendations, ITU, Los Angeles, 1959


11. DAVIES, K 'Ionospheric Radio Propagation' NBS Monograph 80 Washington, 1965
References to Chapter 2 continued

12. YEH, K.C., VILLARD, O. G.,

13. HANSHER, D.H.,

14. JASHK, H.,

15. BETTS, J.A.,

16. LAPORT, F.A.,

17. FELDMAN, C.B.,

18. HILDNER, R.C.,

19. BLACKBAND, W.T.,

20. TERNAN, F.N.,

'Fading and Attenuation of HF Radio Waves' JATP Vol 17 No 4 pp 255-270, 1959


'High Frequency Communication' The English Universities Press Ltd. London 1967


'The Choice of Impedance for Coaxial Radio Frequency Cables' JIEE Vol 102 (B) No 6 pp 804-811 Nov 1955

References to Chapter 3

1. STEIN, S
   JONES, J. JAY
   'Modern Communication Principles
(with application to digital
signalling)'
   McGraw-Hill Book Co NY 1967

2. CHESLER, D
   'Optimum Performance of Codes
transmitted at fixed information over
slowly fading channels'
   Applied Research Lab, Sylvania Electronic
Systems, Research Report 367 (Oct 1963)

3. VAN WAMBECK, S H
   ROSS, A H
   'Performance of diversity receiving
systems'
   Proc IRE 39, pp 256-264 (March 1951)

4. GRISDALE, G L
   MORRIS, J C
   PALMER, D S
   'Fading of long distance radio signals
and a comparison of space and polarization
diversity reception in the 6-18 Mc/s range'

5. LACY, R E
   ACKER, M
   GLASER, J L
   'Performance of space and frequency
diversity receiving systems'
   IRE National Convention Record Pt II
   pp 148-152 (1953)

6. STEIN, S
   JOHANSEN, D E
   'A theory of Antenna performance in
scatter-type reception'
   IRE Trans. Antennas and Prop. AP-9
   pp 304-311 (May 1959)

7. SCHWATZ, M
   BENNETT, W R
   STEIN, S
   'Communication systems and techniques'
   McGraw-hill Book Co, NY 1966

8. ALTMAN, F J
   SICHAH, W
   'A simplified diversity communication
system for beyond-the-horizon links'
   IRE Trans PGCS, CS-4 pp 50-55 (March 1956)

9. PIERCE, J N
   STEIN, S
   'Multiple diversity with non-independent
Fading'
   Proc IRE, 48, pp 89-104
   January 1960

10. PIERCE, J N
    'Theoretical diversity improvement
in frequency shift-keying'
    Proc IRE, 46, 903-910 May 1958

11. CARLSON, A BRUCE
    'Communication Systems'
    NY 1975 Chapter 10

12. BRAYER, K
    'The improvement of digital HF communication
through coding I - interleaved cyclic
coding'
    IEEE Trans of Comm. Tech COM-16 No 6
    pp 771-778 December 1968
13. BOSE, R C
RAY-CHAUDHURI, D K
"On a class of error correcting binary group codes" Inform & Control Vol 3 pp 68-79 1960

14. IBID
"Further results on error correcting binary group codes" Inform & Control Vol 3 pp 279-290 1960

15. BERLE KAMP, E
"Algebraic coding theory"
McGraw-Hill Book Co NY 1968

16. BRAYER, K
"The improvement of digital HF communication through Coding II - tandem interleaved cyclic coding"
IEEE Trans on Comm Tech Com-16 (6) pp 779-786 December 1968

17. VITERBI, A J
"Convolutional codes and their performance in communication systems"

18. McMAMON, P
JANC, R V
TSAT, S
"Comparison of diversity, non-diversity and coding on a parallel FSK HF channel"
IEEE Trans Comm Tech COM-17 (3) pp 355-367 June 1969

19. BRAYER, K
"Error Correction Code performance on HF, tramposcatter, and satellite channels"

20. ROBINSON, J P
BERNSTEIN, A J
"A class of binary recurrent codes with limited error propagation"

21. KASAMI, T
"A decoding procedure for multiple-error-correcting cyclic codes"
IEEE Trans Info Theory Vol IT-10 pp 134-138 April 1964

22. BETTS, J A
ELLINGTON, R P
JONES, D R L
"CW sounding and its use for control of HF (3-30MHz) adaptive systems for data transmission"

23. HAMSHER, D M
"Communication System Engineering Handbook"
McGraw-Hill Book Co NY 1967

24. BETTS, J A
"High frequency communications"
The English Universities Press, London 1967
<table>
<thead>
<tr>
<th>Reference</th>
<th>Author(s)</th>
<th>Title</th>
<th>Publisher/Place</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>COOK, C E, BERNFELD, M</td>
<td>'Radar Signals'</td>
<td>Academic Press, New York</td>
<td>1967</td>
</tr>
<tr>
<td>3.</td>
<td>TURIN, G L</td>
<td>'An Introduction to Matched Filters'</td>
<td>IRE Trans on Information Theory</td>
<td>1960</td>
</tr>
<tr>
<td>4.</td>
<td>ALLEN, J</td>
<td>'Digital Device Technology for Signal Processing'</td>
<td>IEE Conf Publ No 144</td>
<td>1976</td>
</tr>
</tbody>
</table>
References to Chapter 4 continued

12. HARTMANN, C S
   CLAIROBII!RE, L T
   BUSS, D D
   STAPLES, E J
   'Programmable Transversal Filters
   Using Surface Waves, Charge Transfer
   Devices and Conventional Digital Approaches'
   International Specialist Seminar on
   Component Performance and Systems
   Application of Surface Acoustic Wave
   Devices, Aviemore September 1973,
   IEE Conference Publication No 109
   pp 102-114

13. TURIN, G L
   'An Introduction to Digital Matched Filters'
   IEE Proceedings Vol 64 No 7 pp 1092-1112
   July 1976

14. SANGSTER, F L J
    TEER, K
   'Bucket-Brigade Electronics - New Possi-
   bilities for Delay, Time-Axis Conversion
   and Summing'
   IEEE J Solid State Circuits SC-4 pp 131-
   136 June 1969

15. SEQUIN, C H
    TOMPSETT, M F
   'Charge Transfer Devices'
   Advances in Electronics and Electron
   Physics, Supplement of Academic Press
   Chapter 6 1975

16. MacLENNAN, D J
   'Charge Coupled Devices 3 - Signal Process-
   ing'
   WIRELESS WORLD February 1975 pp 61-65

17. BUSS, D D
    HEWES, C R
    deWITT, M
   'Charge Coupled Devices for Analogue
   Signal Processing'
   The Impact of New Technologies in Signal
   Processing Aviemore pp 17-26
   IEE Conf Pub 144 1976

18. RABINER, L R
    GOLD, B
   'Theory and Application of Digital
   Signal Processing'
   Prentice-Hall 1975

19. BUSS, D D
    COLLINS, D R
    BAILEY, W H
    REEVES, C R
   'Transversal Filtering Using Charge-
   Transfer Devices'
   IEEE J Solid-State Circuits SC-8
   pp 138-146 April 1973

20. SMITH, W R
    GERARD, H M
    COLLINS, J H
    REEDER, T M
    SHAW, H J
   'Analysis of Interdigital Surface Wave
   Transducers by V of an Equivalent
   Circuit Model'
   IEEE Trans Microwave Technology Vol MTT-17
   pp 856-864 November 1969

21. HARTMANN, C S
    BELL, DeLamer, T Jr
    ROSENFELD, R C
   'Impulse Model Design of Acoustic Surface
   Wave Filters'
   IEEE Trans on Sonics and Ultrasonics
   Vol SU-20 (2) pp 80-93 April 1973
References to Chapter 4 continued

22. ARMSTRONG, A
   'The Design of SAW Dispersive Filters Using Interdigital Transducers'
   Wave Electronics Vol 2 pp 155-176
   July 1976

23. SMITH, W R
    GERARD, H M
    JONES, W J
   'Analysis and Design of Dispersive Interdigital Surface Wave Transducers'
   IEEE Transactions on Microwave Techniques
   MHT-20 1972 p 458

24. HARTMANN, P
   'Examples of Manufactured Dispersive Delay Lines Using Acoustic Surface Waves'
   The Radio and Electronic Engineer Vol 46(5)
   May 1976 pp 229-236

25. ARMSTRONG, G A
    BUTLER, M B N
   'Engineering Design and Evaluation of SAW Pulse Compression Filters with Low Time Sidelobes'
   The Radio and Electronic Engineer Vol 46(5)
   May 1976 pp 221-228

26. OTTO, O W
   'The Chirp Transform Signal Processors'
   1976 Ultrasonics Symposium Proceedings
   September 29 - October 1 1976
   Annapolis Maryland
   IEEE Catalogue no 76 CH 1120-53U
   pp 365-370

27. DARBY, B J
    GRANT, P M
    COLLINS, J H
    CROSBY, I M
   'Performance of Aperiodic Spread Spectrum Modems. Incorporating Surface Acoustic Wave Analogue Matched Filters'

28. LEVER, K V
    PATTERSON, E
    STEVENS, P C
    WILSON, I M
   'Surface Acoustic Wave Matched Filters for Communication Systems'
   The Radio and Electronic Engineer Vol 46(5)
   May 1976 pp 237-246

29. HAGON, J
   'Programmable Analogue Matched Filters' in 'Component Performance and Systems Application of Surface Acoustic Wave Devices'
   IEE Conf Publ 109 September 1973 pp 92-101

30. MITCHELL, R F
   'Basics of SAW Frequency Filter Design: A Review'
   Wave Electronics Vol 2 pp 111 - 132 July 1976

31. SLOBODNIK, A J
   'Surface Acoustic Waves and SAW Materials'
   Proc IEE Vol 64 No 5 pp 581-595 May 1976

32. SMITH, R W
   'Key Trade-offs in SAW Transducer Design and Component Specification'
   1976 Ultrasonics Symposium Proceedings Sept 29 - Oct 1, 1976 Annapolis, Maryland
   IEEE Catalogue no 76 CH 1120-58U pp 547-552
<table>
<thead>
<tr>
<th>Reference Number</th>
<th>Author(s)</th>
<th>Title and Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>35.</td>
<td>-</td>
<td>1976 Ultrasonics Symposium Proceedings, Annapolis, Maryland, Sept 29 - Oct 1, 1976 IEEE Catalogue No 76 CH1120-5SU</td>
</tr>
<tr>
<td>36.</td>
<td>GOTT, C F</td>
<td>'HF Data Transmission Using Chirp Signals' Proc IEE Vol 118 No 9 pp 1162-1166 September 1971</td>
</tr>
<tr>
<td>37.</td>
<td>TANCRELL, R H</td>
<td>'Analytic Design of Surface Wave Bandpass Filters' IEEE Trans on Sonics and Ultrasonics Vol SU-21 No 1 pp 12-22 January 1974</td>
</tr>
<tr>
<td>38.</td>
<td>BRANDON, P S</td>
<td>'The Design of Lump Constant Dispersion Networks' in 'Delay Devices for Pulse Compression Radar' IEE Conf Publ 20 1966 pp 6-13</td>
</tr>
<tr>
<td>40.</td>
<td>JACOBS, W W</td>
<td>'SAW Spectrum Analyzer for Use in Doppler Processor' in 'Conf Publ of the Military Electronics Defense Expo '76, Wiesbaden Germany, October 6-8 1976 pp 153-171</td>
</tr>
<tr>
<td>41.</td>
<td>WINKLER, M R</td>
<td>'Chirp Signals for Communications' Wescon Convention Record p 17 1962</td>
</tr>
</tbody>
</table>
References to Chapter 5

1. COLLINS, J H
   GRANT, P M
   "The Role of Surface Acoustic Wave Technology in Communication Systems"
   Ultrasonics, 10, pp 59-71 1972

2. HUNSINGER, B J

3. DAVIS, N G
   WOHLBERG, D B
   "Surface Wave Devices in Spread-Spectrum Processors"
   Ultrasonics, 11, pp 254-262 1973

4. LÁRGOA, R
   "Impact of New Technologies in Signal Processing"
   IEE conf publ 144 pp 144-154
   September 1976

5. MORGAN, D P
   HANNAH, J M
   COLLINS, J H
   NELC San Diego California March 1973

6. DARBY, B J
   GRANT, P M
   COLLINS, J H
   CROSBY, I M
   "Spread Spectrum Synchronizers using a SAW Convolver and Recirculation Loop"
   Proc IEEE Vol 64. No 5, pp 751-753
   May 1976

7. HAMSHER, D H
   "Performance of Aperiodic Spread Spectrum Moderns Incorporating Surface Acoustic Wave Analogue Matched Filters"
   IEE Conference Publ 109 Aviemore, Scotland 1973

8. MOSIER, R R
   CLABAUGH, R G
   "Communication System Engineering Handbook"
   McGraw-Hill Book Company 1967

9. GOTT, G F
   KARRA, A J
   "Kineplex, A Bandwidth-Efficient Binary Transmission System"
   TRANS A I M E. Comm and Electronics
   Vol 34 pp 723-727 January 1958

10. GOTT, G F
    NEWSOME, J P
    "Differential Phase-Shift Keying Application Applied to Chirp Data Signals"
    Proc IEEE Vol 121 No 9 pp 923-928
    September 1974

11. SMITH, S
    HARP, J G
    "HF Data Transmission Using Chirp Signals"
    Proc IEEE Vol 118 No 9 pp 1162-1166
    September 1971

Private Communication

12. CRISP, J J
    HECKS, J S
    "PH Code Generation Using SAW Devices"
    Proc IEEE 1974 Ultrasonics Symposium Cat no 74 CHO 896 1 SU pp 393-396
References to Chapter 5 continued

13. LEVER, K V
   PATTERSON, E
   STIVENS, P C
   WILSON, I M

'Surface Acoustic Wave Matched Filters for Communications Systems'
The Radio and Electronic Engineer
Vol 46 No 5 pp 237-246 May 1976
References to Appendix I

1. BAIN, W C. HARRISON, M D.
   'Model Ionosphere for D-Region at summer noon during sunspot maximum'
   Proc IEE, 119, pp 790-796 1972

2. KRASSOUKY, V I.
   'Exploration of the upper atmosphere with the help of the third soviet
   sputnik'
   P. IRE 47 pp 289-296 Feb 1959

3. RENARD, C.
   'Détermination des paramètres constituifs de l'ionosphère entre 78 et 100 Km
   d'altitude par la mesure locale du champ TBF rayonné par un émetteur au sol'
   Ann Téléc 21 pp 117-135 May-June 1966

4. CHAPMAN, S.
   'The absorption and dissociative or ionizing effect of non-chromatic radi-
   ation in an atmosphere on a rotating
   earth'
   Proc Phys Soc 43 pp 26-45 Jan 1931

5. APPLETON, E V
   'The normal E-region of the ionosphere'
   PIRE 47 pp 155-159 1959

6. DAVIES, K
   'Ionospheric Radio Propagation'
   NBS Monograph 80, Washington 1965

7. NICOLET, M
   'Meteor ionization and the night-time E-lyser'
   JATP special supplement on meteor physics 2 pp 99-110 1955

8. MARTYN, D F
   'The normal F-region of the ionosphere'
   PIRE 47 pp 147-155 Feb 1959

9. WRIGHT, J W
   KNECHT, R W
   DAVIES, K
   'Annals of the IGY' Vol III Part 1
   (Pergamon Press, 1957)

10. GORDON, W E
    'Incoherent scattering by radio waves by free electrons with applications to
    space exploration by radar'
    Proc IRE 46 pp 1824-9 1958

11. ARENDT, P R
    PAPAYDANOU, A
    SOICHER, H
    'Determination of the ionospheric electron content utilizing satellite signals'
    PIRE 53 pp 268-277 March 1965

12. HALLEY, P
    GERVAISE, ANNE-MARIE
    'Essai d'une prévision à long term de l'activité solaire paramètres du 20
    cm cycle'
    Ann Téléc 18 pp 94-104 May-June 1963
References to Appendix I continued

13. NAISMITH, R
'A new method of ionospheric forecasting'  

14. JOACHIM, M
'Progrès récents dans la prévision à long terme de la propagation ionosphérique des ondes radio électroniques.  
Conférence prononcée à l'UIT 25 Jan 1967 (HC)

15. GEORGE, P C  
BRADLEY, P A
'A new method of predicting the ionospheric absorption of HF waves at oblique incidence'  
ITU Telecom J 41 pp 307-312 May 1974

16. MUGGLETON, L M
'Dependence of ionospheric absorption on solar zenith angle'  
IEEE 113 pp 1909-1912 Dec 1966

17. MARTYN, D F
'The propagation of medium radio waves in the ionosphere'  
Proc Phys Soc 47 p323, 1955

18. CCIR
'Fading allowances for the various classes of emission - recommendation 340'  
Documents of the Xth plenary assembly vol III pp 25-29

19. BETTS, J A  
ELLINGTON, R P  
DNGES, D R L
'CW sounding and its use for control of HF (3-30MHz) adaptive systems for data transmission'  
Proc IEE 117 No 12 Dec 1970 pp 2209-2215

20. YEH, K C  
VILLARD, O G
'Fading and attenuation of HF radio waves'  
JATP Vol 17 No 4 pp 255-270 1959

21. BUDDEN, K G
'Radio waves in the ionosphere'  
Cambridge University Press 1961

22. NBS
Circular 462 Ionospheric Radio Propagation - government printing office Washington DC

23. CCIR
'Utilisation des données sur les bruits atmosphériques radio électriques'  
AVIS 372 OSLO 1966

24. CCIR
'Repartition mondiale et caractéristiques des bruits atmosphériques radio électriques'  
Rapport 322 Genève 1964

25. HAMSITER, D H
'Communication Systems Engineering Handbook'  

26. BREIT, G  
TUVE, M A
'A test of the existence of the conducting layer'  
Phys Rev 28 p 554 1926
Reference to Appendix 2

1. LAPORT, E.A.
   'Radio Antenna Engineering'
   McGraw-Hill Book Co Inc
   New York 1952

2. FELDMAN, C.B.
   'Deviations of Short Radio Waves from the London-New York Great Circle Path'
   Proc IRE vol 27 p635
   October 1939

3. MILDNER, R.C.
   'The Power Rating (Thermal) of Radio Frequency Cables'
   IEE vol 93 Part III No 24 pp 296-304
   July 1946

4. BLACKBAND, W.T.
   'The Choice of Impedance for Coaxial Radio-Frequency Cables'
   P. LEE Vol 102B No 6 pp 804-81V
   November 1955

5. TERNAN, F.E.
   'Radio Engineers' Handbook'
   pp 678-700 McGraw-Hill Book Co Inc
   New York 1943

6. IEE
   'Convention on HF Communication'
   March 1963

7. JASIK, H.
   'Antenna Engineering Handbook'
   McGraw-Hill Book Co Inc
   New York, London, 1961

8. WILLIAMS, H. PAUL
   'Antenna Theory and Design Vol II The Electrical Design of Antennae'
   SIR ISAAC PITMAN and SONS LTD
   LONDON 1950

9. SCHELKUNOFF, S.A.
    FRTIS, H.T.

10. CCIR

11. CRESS, P.
    DIRKSEN, P.
    GRAHAM, J.W.

12. LERWIN, L.

13. DECJER, R.R.

14. BLACKBAND, W.T.
   'The Influence of Gain and Current Attenuation on the Design of the Rhombic Antenna'
   Trans of the IRE, 1951 AP - 7 p 188

15. BRUNSDON, E.G. 'The New Bearley Receiving Station'


17. CHRISTIANSEN, W.N. 'Rhombic Antenna Arrays' AWA Technical Review 1947 7 p 361

18. JACOBS, W.W. Private communication to CCIR April 1973

19. LERWIN, L. 'Discussion on "Radiation from Rhombic Antennas" by Donald Foster' Proc IRE Vol 29 p523 September 1941


24. CARRELL, R.L. 'Analysis and Design of the Log Periodic Dipole Antenna' Tech Rpt No 52 Antenna Laboratory, University of Illinois, 1960


References to Appendix 2 continued

28. DE VITO, G
STRACCA, G.B.

29. CHEONG, W M

30. CHEONG, W M
KING, R W P

31. KING, R W P
MACK, R B
SANDLER, S S

32. HARRINGTON, R F

33. HARRINGTON, R F

34. SINNOTT, D H

'Comments on the Design of Log-Periodic Dipole Antennas'
IEEE Trans on Antennas and Propagation Vol AP-21 No3 May 1973

'Arrays of unequal and unequally spaced elements'
PhD Dissertation Harvard University Cambridge Mass 1967

'Arrays of unequal and unequally spaced elements'
Radio Sci Vol 2 pp 1301-1314 November 1967

'Arrays of cylindrical dipoles'
New York Cambridge University Press 1968

'Generalized network parameters in field theory'
Proc Symp on Generalized Networks MR15 Series Vol 16 Brooklyn
New York Polytechnic Press 1966

'Matrix methods for field problems'
Proc IEEE Vol 55 No 2 pp 136-149 February 1967

'Multiple Frequency Computer Analysis of the Log-Periodic Dipole Antenna'
IEEE Trans on Antennas and Propagation pp 592-593 July 1974
The application of Surface Acoustic Wave (SAW) technology to Doppler Signal processing permits the realisation of real time spectral analysis. The parameters of SAW filters however limit spectral resolution to around 20 KHz. This limitation can be overcome by interfacing the SAW processor with an analogue time compression store which converts data from an input rate related to the Doppler bandwidth to an output rate compatible with SAW processing.

This form of compressed time spectral analysis is particularly suited to radar signal processing where the required resolution is tens to hundreds of Hz. Also the processing gain associated with pulse compression assists detection and analysis of coherent Doppler signals immersed in noise. In order to realise the full potential of this technique it is essential to optimise the system resolution and provide maximum suppression of spurious signals. This paper describes in detail the design and implementation of a SAW Spectrum Analyser.
1. INTRODUCTION

In Doppler signal processing, the realisation of filter banks for measurement of target velocities at all ranges is difficult with either analogue circuitry or an equivalent digital Fourier Transformer. The advent of surface acoustic wave technology (1) permits the application of pulse compression techniques to fast spectral analysis (2). The parameters of S.A.W. filters, however, limit spectral resolution to approximately 20KHz. These disadvantages can be removed by interfacing an S.A.W. spectrum analyser with a C.C.D. analogue data store to accept data at a rate dictated by the Doppler bandwidth, but outputting it in a time scale compatible with the surface wave processing. This technique of variable spectral resolution dependent on C.C.D. storage time was first described by Roberts (3) and is shown schematically in Fig. 1. Discrete samples of bipolar video corresponding to a particular range bin are clocked into a C.C.D. register at a rate determined by the radar p.r.f. Following acquisition of a burst of data, each sample is shifted to free the input register for the next data sample. Segmenting the video signals in this way is a conceptually simple process if the data store is structured as a rectangular matrix of cells with orthogonal read and write directions. Fast read-out is achieved by suitably arranging the output clock frequency to be a factor of up to 1000 greater than the input clock frequency.

This form of compressed time spectral analysis is particularly attractive for radar signal processing where the required resolution is tens to hundreds of Hz. Fast processing allows the channels to be analysed sequentially using surface wave technology while data ordering can be incorporated in the
design of the C.C.D. In addition, the processing gain associated with pulse compression permits detection and analysis of coherent Doppler signals immersed in noise. In order to realise the full potential of this type of equipment, it is essential that the surface wave spectrum analyser should be designed to give optimum overall system performance in respect of resolution and suppression of spurious signals. This paper describes in detail the design and implementation of such a unit.

2. CHIRP TRANSFORM

The chirp transform is a signal processing algorithm which can perform Fourier Transformation in real time. It relies on structuring the conventional Fourier transform as two complex multiplications separated by a convolution stage as shown in Fig. 2. The output of a linear FM filter with baseband phase response \( \exp(j\frac{\mu t}{2}) \) can be written as

\[
P(t) = \int p(t) \exp\left\{j\mu (t-\tau)^2/2\right\} d\tau
\]

where \( p(t) \) is the input signal and \( \mu \) is the dispersive slope.

If the conventional Fourier transform is given by

\[
R(\omega) = \int r(t) \exp(-j\omega t) \, dt
\]

and both input and output signals are multiplied by chirps so that

\[
p(t) = r(t) \exp(-j\mu t/2)
\]

and

\[
R(\tau) = P(\tau) \exp(-j\mu \tau /2)
\]

then the final output can be written as

\[
R(\mu t) = \int r(t) \exp\left\{-j(\mu t)t\right\} dt
\]

where the transform variable \( \omega = \mu t \).

The / cont...
The system time bandwidth product can be deduced from the condition that the bandwidth, \( B_2 \), of chirp filter \( C_2 \), must be greater than or equal to the bandwidth, \( B_1 \), of the filter \( C_1 \), plus the maximum signal bandwidth \( B_s \) so that

\[ \mu T_1 = B_s + \mu T_2 \]  

(6)

where \( T_1, T_2 \) are the lengths of the impulse responses of the two chirp filters \( C_1, C_2 \).

If the input signal time duration is equal to \( T_1 \), then from (6)

\[ T_2 = 2T_1 \]  

(7)

and hence

\[ T_2 B_2 = 4T_1 B_1 \]  

(8)

The finite input signal duration limits the resolution of the analyser to \( 1/T_1 \). For a single frequency input, the output of the analyser is a pulse defined by the function \( \text{sinc} \left( \frac{\mu T_1}{T_1} \right) \) which has a 4dB width of \( 1/T_1 \) but introduces unwanted sidelobes 13dB down from the main peak. These sidelobes can be reduced at the expense of frequency resolution by spectral weighting as in conventional pulse compression radar (4). There are a number of possible ways in which weighting can be applied:

(a) Integral amplitude weighting in the S.A.W. delay line (1).

(b) Frequency filtering of the output display after post-multiplication.

(c) Amplitude weighting of the input data using a digitally synthesized weighting function prior to multiplication by chirp \( C_1 \).

Integral / cont...
Integral amplitude weighting is only possible in the premultiplying chirp delay line $C_1$, since in $C_2$ different sections of the delay line are activated by chirps corresponding to different input frequencies. It was, however, not possible to find a mixer with sufficient linearity to sustain up to 40dB dynamic range of the input data on the R port, while still providing up to 20dB on the L port, for amplitude weighting of the chirp waveform. On the other hand, controlled frequency filtering of the output would require the use of a multi-element discrete LC filter or else a custom designed S.A.W. bandpass filter. While this approach is distinctly possible, it was decided to provide time domain modulation at the input, as this was considered to give a simpler implementation of the required effect. The input signal is multiplied by a digitally synthesized weighting function using analogue multipliers. The weighting function is stored as digital data in a read-only memory. The output from the memory is synchronised to the input signal by an appropriate timing pulse and converted to the required function by a D to A converter. The use of analogue multipliers permits the modulation of the input waveform to be performed at baseband, greatly enhancing the linearity of this operation. Fine adjustments of the compressed pulse shape can be made by altering the amplitude weighting using the multiplier bias circuits. This gives electronic adjustable pulse shaping with a possible trade-off between resolution and sidelobe level. When integral spectral weighting is employed in an S.A.W. filter, it is not possible to adjust the output pulse shape to the same extent.

3. SYSTEM DESCRIPTION

A block diagram of the spectrum analyser module is shown in Fig. 3(a). The system consists of four main sections.
Sections (a) and (b) feed the R and L ports of mixer M3 respectively while section (c) is fed by the X port of this mixer. Section (d) provides synchronisation of sections (a) and (b).

3.1 S.A.W. Device Design

The system is designed to process and display the line spectrum of 25μsec. baseband data samples from a C.C.D. store with a spectral resolution of better than 60kHz. The frequency range covered is 0 to ±2MHz. Hence the S.A.W. delay lines are designed to have the following specification:

<table>
<thead>
<tr>
<th></th>
<th>Time Duration</th>
<th>Bandwidth</th>
<th>Centre Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₁</td>
<td>25μsec.</td>
<td>4MHz</td>
<td>22MHz</td>
</tr>
<tr>
<td>C₂</td>
<td>50μsec.</td>
<td>8MHz</td>
<td>32MHz</td>
</tr>
</tbody>
</table>

The time duration of C₁ is determined by the length of each data sample. The design of C₂ is determined by the requirement for a dispersive slope equal to that of C₁, while its bandwidth must be capable of accommodating the bandwidth of C₁ plus the maximum range of Doppler signals, i.e. 8MHz. The choice of chirp I.F's are only limited by the constraint their sum must equal the local oscillator I.F, i.e. 54MHz, while the fractional bandwidth of each device should not exceed approximately 30%.
Surface wave filters were designed using the interdigital transducer technology where the dispersion is determined by placement of interleaved metal fingers. The design principles for this type of S.A.W. filter are well established (5) but particular care was taken to using a split-electrode geometry (6) to reduce acoustic reflections at finger edges, and to compensate the electrode overlaps for acoustic diffraction (7). The substrate material used was ST-X cut quartz, and in order to obtain 50μsec. dispersive delay, a 200mm substrate was required.

4. SYSTEM OPERATION

The system has two modes of operation - the 'I' mode and the 'I + Q' mode.

4.1 'I' Mode

In this mode, the system analyses only the in-phase component, while in the latter it operates on a combination of both in-phase and quadrature components and is hence capable of distinguishing approaching and receding targets. In the 'I' mode the signal is processed as described above, except that after mixing the modulated input data with the local oscillator output in M1, the lower sideband is rejected by a surface wave bandpass filter. The bandpass characteristic of this filter is shown in Fig. 4. An S.A.W. filter is used in order to achieve effective image rejection. The remaining sideband is amplified and mixed with chirp C1 in M3. When the output from M3 is passed through the compressor C2, the relevant sections are activated and the resultant output is a series of compressed pulses at a time position determined by the frequency components of the input signal. These pulses are detected using an envelope detector to permit an accurate appraisal of the amplitude spectrum of the radar return.
In the 'I + Q' mode, two inputs are used from distinct C.C.D. stores, as shown in Fig. 3(b). These are modulated in similar input channels and up-converted using a $\pi/2$ phase delay in the 54MHz line to the Q-channel to give I and Q channel outputs in anti-phase. Subsequent summation of the two signals gives either the upper or lower sideband of 54MHz as a result of the Q-channel data leading or lagging the I-channel data in phase. After summation of the Doppler signals in this manner, the sideband indicates directly the direction of motion of the target. This facility imposes two significant constraints on the spectrum analyser:

a) The input processing channels must be set up to closely similar performance if sideband cancellation to better than 30dB is to occur for all input frequencies;

b) The spectrum analyser must be designed to cover a 4MHz bandwidth instead of the 2MHz directly required by the input signal bandwidth.

Following the summation point, the signal is routed to the R port of M3 and processed in the same way as the 'I' mode.

In the system implementation, the post multiplying chirp has been omitted since only the spectral amplitude is required. The amplitude component is obtained by envelope detecting the signal output from the S.A.W. compressor (C₂).

### Spectral Usage

From the above description, it is clear that there are three I.F's in the system. This situation was deliberately permitted to ensure that the three distinct operations, analogue multiplication, chirp generation and acoustic convolution, were carried out at different points in the frequency band.
This ensures that intermodulation distortion and undesirable interactions between different sections of the system would be kept to a minimum. The particular frequencies of 22MHz for chirp C₁ and 32MHz for chirp C₂ were chosen as optimum frequencies compatible with practical realisation of these devices with moderate fractional bandwidth. It should be borne in mind that chirp filter C₂ itself gives very good suppression of out-of-band spurious, as is evident from its bandpass characteristic in Fig. 5.

5. **SYSTEM PERFORMANCE**

The system was designed to offer the following characteristics:

(a) Linear dynamic range of better than 40dB.

(b) Flat amplitude response over the range of input signals (0 to + 2MHz).

(c) Frequency resolution of better than 60KHz.

(d) Spurious signal levels, including noise, of better than 40dB below the maximum output signal.

Beside these general constraints, the sideband suppression must be effective. In the double sideband 'I + Q' mode the unwanted sideband is cancelled by a precise phase match of two equal and opposite signals. This cancellation must be sufficiently well controlled to ensure at least 30dB suppression of the unwanted sideband. In the single sideband mode this level of suppression is provided by the surface wave bandpass filter.
S.A.W. Spectrum Analyser Performance

Fig. 6(a) and Fig. 6(b) indicate the amplitude response of the analyser as a function of frequency. Over its specified bandwidth it is flat to within 0.5dB in either mode. In the double sideband mode the amplitude response has only decreased by 1dB at ± 3MHz. The response in the single sideband mode is determined by the frequency response of the bandpass filter.

The form of I.F. compressed pulse for a single frequency input is shown in Fig. 7. The high degree of symmetry in the virtually theoretical sinc \((\mu T)\) characteristic implies a very low level of amplitude and phase error in the S.A.W. chirp filters. Indeed the measured phase errors were less than 2.5 degrees r.m.s. in both filters and the dispersive slopes were matched to within 1 part in 10\(^3\). The measured 3dB width of the compressed pulse achieves its theoretical value of 36KHz. When the input waveform is modulated by the digitally synthesized weighting function, the output appears as shown in Fig. 8, with -40dB sidelobes shown in Fig. 9. The corresponding photographs of the amplitude detected compressed pulse are shown in Fig. 10. The 3dB width is 52KHz which is predicted theoretically.

The level of suppression of spurious signals is shown in Fig. 11. The main response is a line spectrum at 800KHz. Other spurious signals are suppressed by more than 37dB. The main spurii occur at the zero frequency position and are due to direct electromagnetic leakage. All other spurious signals originate in the S.A.W. filters.
5.2 Performance of S.A.W. and CCD in Combination

The operation of the analyser and CCD store in combination is shown in Fig. 12. In Fig. 12(a), the top trace (5msec./division) shows the summation of two inputs at 1100 and 1200Hz respectively. The centre trace (5μsec./division) indicates the CCD output, while the bottom trace (800KHz/division) illustrates the resolution which is obtained with the spectrum analyser. The way in which the analyser is able to distinguish input signals immersed in noise is shown in Fig. 12(b). The CCD output trace comprises equal amounts of 500Hz signal and noise, while the analyser output shows the display of the spectral line corresponding to the input signal.

6. CONCLUSIONS

The operation of a spectrum analyser for real time data analysis has been demonstrated using surface acoustic wave interdigital transducer technology. When such an analyser is interfaced with a CCD time compressor, one obtains a sensitive, compact, high-speed Doppler frequency analyser. When compared with digital techniques, the CCD/S.A.W. analyser is likely to be cheaper and less power-consuming, but suffers in dynamic range from the high insertion loss of the S.A.W. filters. Since the unit operates at the I.F. of the S.A.W. filters, much greater flexibility in operating speed is obtained.

7. ACKNOWLEDGEMENTS

This work has been carried out with the support of the Ministry of Defence, Procurement Executive, sponsored by C.V.D. The authors are indebted to J.G.B. Roberts and R. Eames of the Royal Signals and Radar Establishment for advice and encouragement and for permission to include Fig. 12.
8. REFERENCES


FIG. 1. MULTIPLEXED CCD STORE

FIG. 2. CHIRP TRANSFORM
FIG. 3(a)  
SYSTEM CONFIGURATION

FIG. 3(b)  
DOUBLE SIDEBAND MODULATION CIRCUIT

FIG. 4.  
S.A.W. BANDPASS FILTER RESPONSE
FIG. 5. CHIRP FILTER RESPONSE

FIG. 6. SPECTRUM ANALYSER RESPONSE
FIG. 7.  I.F. COMPRESSED PULSE WITHOUT SIDELOBE SUPPRESSION
FIG. 8. I.F. COMPRESSED PULSE WITH SIDELOBE SUPPRESSION.
FIG. 9. SIDELOBE DETAIL

0.5\mu sec. / division

Maximum sidelobes

-40dB
FIG. 10 ENVELOPE DETECTED PULSE

(a) 1 μsec. / division

(b) 0.2 μsec. / division
FIG. 11. SPURIOUS SIGNAL SUPPRESSION
FIG. 12. OPERATION WITH CCD STORE

(a) Spectral Resolution

(i) Input signal
(ii) CCD Output
(iii) Spectrum Analyser Output

(b) Recovery of signal from noise

(i) Input signal
(ii) Noise
(iii) CCD Output
   (Signal plus noise)
(iv) Spectrum Analyser Output