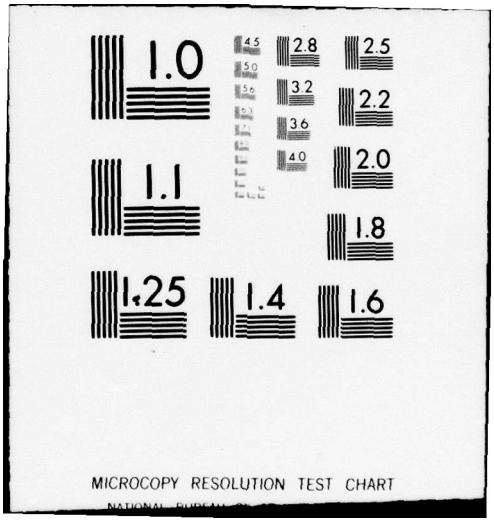
AD-A073 599 ADVISORY GROUP FOR AEROSPACE RESEARCH AND DEVELOPMENT--ETC F/G 17/2
DIGITAL COMMUNICATIONS IN AVIONICS.(U)
JUN 79 H LUEG

UNCLASSIFIED AGARD-CP-239 NL

1 OF 5
ADA 073599




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
7.1
8.0
9.0
MICROCOPY RESOLUTION TEST CHART

CP 239
DIGITAL COMMUNICATIONS IN AVIONICS
AVP

AGARD-CP-239

AGARD-CP-239

AGARD

ADVISORY GROUP FOR AEROSPACE RESEARCH & DEVELOPMENT

7 RUE ANCELLE 92200 NEUILLY SUR SEINE FRANCE

AD A073599

LEVEL II

DDC FILE COPY

AGARD CONFERENCE PROCEEDINGS No. 239

Digital Communications in Avionics

Edited by

Prof.Dr rer.nat.H.Lueg

DDC
RECEIVED
SEP 11 1979
A

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

NORTH ATLANTIC TREATY ORGANIZATION

DISTRIBUTION AND AVAILABILITY
ON BACK COVER

79 09 10 103

AGARD-CP-239

NORTH ATLANTIC TREATY ORGANIZATION

ADVISORY GROUP FOR AEROSPACE RESEARCH AND DEVELOPMENT

(ORGANISATION DU TRAITE DE L'ATLANTIQUE NORD)

AGARD Conference Proceedings

DIGITAL COMMUNICATIONS IN AVIONICS

Edited by

Prof. Dr rer. nat. H.Lueg
Institut für Technische
Elektronik der Rein-Westf.
Technischen Hochschule Aachen
51 Aachen
Templergraben
Germany

Jun 79

468 p.

Copies of papers and discussions presented at the Avionics Panel Symposium held in Munich, Germany, 5–9 June 1978.

THE MISSION OF AGARD

The mission of AGARD is to bring together the leading personalities of the NATO nations in the fields of science and technology relating to aerospace for the following purposes:

- Exchanging of scientific and technical information;
- Continuously stimulating advances in the aerospace sciences relevant to strengthening the common defence posture;
- Improving the co-operation among member nations in aerospace research and development;
- Providing scientific and technical advice and assistance to the North Atlantic Military Committee in the field of aerospace research and development;
- Rendering scientific and technical assistance, as requested, to other NATO bodies and to member nations in connection with research and development problems in the aerospace field;
- Providing assistance to member nations for the purpose of increasing their scientific and technical potential;
- Recommending effective ways for the member nations to use their research and development capabilities for the common benefit of the NATO community.

The highest authority within AGARD is the National Delegates Board consisting of officially appointed senior representatives from each member nation. The mission of AGARD is carried out through the Panels which are composed of experts appointed by the National Delegates, the Consultant and Exchange Programme and the Aerospace Applications Studies Programme. The results of AGARD work are reported to the member nations and the NATO Authorities through the AGARD series of publications of which this is one.

Participation in AGARD activities is by invitation only and is normally limited to citizens of the NATO nations.

The content of this publication has been reproduced
directly from material supplied by AGARD or the authors.

Published June 1979

Copyright © AGARD 1979
All Rights Reserved

ISBN 92-835-0242-6



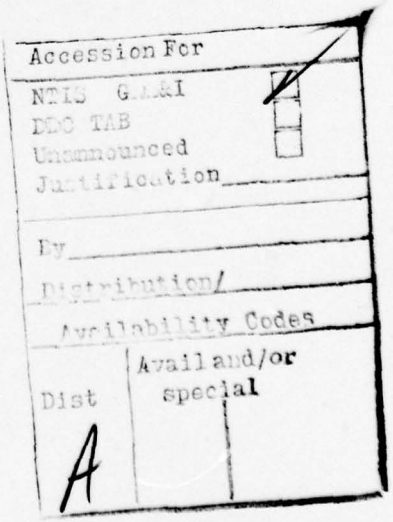
Printed by Technical Editing and Reproduction Ltd
Harford House, 7–9 Charlotte St, London, W1P 1HD

THEME

In recent years, the technology of digital communications has expanded rapidly, and systems embodying these techniques are being implemented. Consequently, the AGARD Avionics Panel considered a symposium on this topic timely, in order to exchange information on new systems being developed and on new applicable technology.

Applications include communications between air and/or ground terminals and relay communications, both satellite-borne and airborne. Improvement of communication in case of multiple path transmission by means of spectrum techniques, echo cancelling or pulse compression and improvement of jam resistance by means of power balance, spread spectrum techniques or error protecting coding were also of interest.

Emphasis was on cost-effective improvements in performance in the presence of noise and interference, both natural and man-made. The utilization of new devices and their impact on cost, size, weight and performance were also considered.



PROGRAM AND MEETING OFFICIALS

CHAIRMAN: Prof. Dr rer. nat. H.Lueg
Institut für Technische
Elektronik der Rein-Westf.
Technischen Hochschule Aachen
51 Aachen
Templergraben
Germany

MEMBERS

Dr Ing. H.J.Albrecht (EPP)
FGAN
5307 Wachtberg-Werthhoven
Königstrasse 2
Germany

Mr l'Ingénieur en Chef M.Carlier
Centre d'Essais en Vol
B.P. No.2
91220 Brétigny-sur-Orge
France

Mr J.P.Andersen
Chief, Aeronautical Systems
Programs Division
Transportation System Center
Kendall Square
Cambridge, Mass 02142, USA

Dr F.I.Diamond
Technical Director
Communications and Division
Rome Air Development Center
Griffiss AFB, N.Y. 13441
USA

AVIONICS PANEL

CHAIRMAN: Ir H.A.T.Timmers, Netherlands

DEPUTY CHAIRMAN: Dr M.Vogel, Germany

HOST COORDINATOR

Dr M.Vogel
DFVLR e.v.
8031 Oberpfaffenhofen
Post Wessling/obb
Germany

PANEL EXECUTIVE

Cdr D.G.Carruthers, USN
AGARD-NATO
7, rue Ancelle
92200, Neuilly-sur-Seine
Tel: 745-08-10

or from USA and Canada only:
APO New York 09777

PREFACE

NATO communication requirements are continually expanding and are approaching the limits of existing telecommunications systems. The purpose of this meeting is to show a small but most significant portion of the entire area of communications with good future prospects, namely the "Digital Communications in Avionics".

Why, particularly, did the digital communication techniques progress so significantly during the last few years? What advantages may be envisaged with respect to classical techniques using analog transmission, such as the modulation of amplitude, frequency, or phase? Perhaps three reasons may be responsible:

(1) The possibility to convert analog values to digital ones, to regenerate signals in relay terminals and thus to reduce the influence of interference which may have originated on interim links. PCM-transmission links tested in the United States display – apart from quantization noise – an absence of all other interference and may transmit practically noise-free signals over distances of several thousand kilometers as a consequence of interference reduction by regeneration.

(2) Modern and fast computers, all using digital methods, permit an adequately fast performance with signal processing. Analog computers are too slow for this task; they have mostly been replaced by digital ones.

In addition, only digital storage facilities allow to store signals in such a way that modifications due to drift or similar effects are reduced to a minimum, and that, in other words, the entire information is maintained.

(3) The semi-conductor technology has been implemented to such an extent that, already today, fast digital units are available at low price, resulting in far more economical digital communication than with analog systems.

Since the first digital transmission link had been introduced about 20 years ago, considerable development has taken place; digital transmission of light pulses in fibre glass is an excellent example. Perhaps only the forthcoming decade will indicate technical limitations related to the number of gigabits per second which may reliably and economically be transmitted.

I have just tried to sketch the enormously steep developmental trend which we presently experience and during which our Symposium on Digital Communications in Avionics is now taking place; the meeting is intended to cover several relevant problem areas.

Each session begins with a review on the state of the art and on problems yet unsolved. Well-known experts in their respective fields have undertaken this task. I would like to thank them for their assistance and cooperation.

Session I covers "digital communication concepts and systems", it represents an introduction to digital communication in general.

The symposium continues with *Session II* on "error-correction coding", which is an important subject, especially with regard to signal regeneration.

Session III deals with "propagation effects including channel modelling and simulation"; related parameters are discussed.

Session IV treats "applications and special devices for digital communication systems" followed by *Session V* on "source encoding and data compression". The unclassified character of the symposium does not allow detailed presentations and discussion in several sessions of the symposium, particularly in the last-mentioned one.

Sessions VIa and *VIb* refer to satellite communications, subdivided into "multiple access" and "modulation and multiplexing", respectively.

I would like to especially express my appreciation to Dr Diamond for his contribution to the programme in its final version and for soliciting and providing a large number of papers authored by colleagues from the United States. In addition, I would like to thank Dr Albrecht for his many valuable suggestions and for his assistance in organizing the symposium.

H. Lueg

HEINZ LUEG

CONTENTS

	Page
THEME	iii
PROGRAM AND MEETING OFFICIALS	iv
PREFACE by Heinz Lueg	v
	Reference
SESSION I – DIGITAL COMMUNICATIONS CONCEPTS AND SYSTEMS	
THE IMPACT OF DIGITIZATION ON MILITARY COMMUNICATIONS by I.L.Lebow	1
LES ASPECTS TECHNIQUES ET OPERATIONNELS DES TELECOMMUNICATIONS EN AERONAUTIQUE par M.Carlier	2
A NOVEL APPROACH TO THE DESIGN ON AN ALL DIGITAL AERONAUTICAL SATELLITE COMMUNICATION SYSTEM by M.E.Ulug	3
CENSAR TDMA CENTRALIZED SYNCHRONIZATION AND RANGING FOR TIME-DIVISION MULTIPLE ACCESS by P.P.Nuspl	4
Paper 5 Cancelled	
A DIGITAL COMMUNICATION SYSTEM AS GATEWAY BETWEEN ADJACENT COMPUTERIZED AIR TRAFFIC CONTROL CENTRES by M.Baum	6
A MARKOV MODEL FOR NONLINEAR CHANNELS WITH MEMORY AND SOME APPLICATIONS by E.Biglieri	7
SESSION II – ERROR CORRECTION CODING	
STATE OF THE ART OF ERROR CONTROL TECHNIQUES by J.K.Wolf	8
FORWARD ERROR-CORRECTION FOR THE AERONAUTICAL SATELLITE COMMUNICATIONS CHANNEL by A.Sewards, L.Beaudet and H.Ahmed	9
AN EXPERIMENTAL EVALUATION OF INTERLEAVED BLOCK CODING IN AERONAUTICAL HF CHANNELS by B.Hillam and G.F.Gott	10
Paper 11 Cancelled	
AN ASYNCHRONOUS DATA TRANSMISSION SYSTEM WITH LOW ERROR PROBABILITY FOR THE SETAC LANDING AID by W.Beier	12
ON THE PERFORMANCE OF A MAXIMUM LIKELIHOOD DECODER FOR CONVOLUTIONAL CODES by J.P.M.Schalkwijk	13
DIGITAL COMMUNICATIONS USING SOFT-DECISION DETECTION TECHNIQUES by P.G.Farrell, E.Munday and N.Kalligeros	14

Reference

THEORETICAL LIMITS ON CHANNEL CODING UNDER VARIOUS CONSTRAINTS
by B.G.Dorsch and F.Dolainsky 15

AN ERROR-RATE MEASUREMENT SET-UP OPERATING AT 1 GBITS/S
by U.Wellens 16

SESSION III – PROPAGATION EFFECTS INCLUDING CHANNEL MODELING AND SIMULATION

INTRODUCTORY NOTES ON PROPAGATION EFFECTS AND RELATED ASPECTS
by H.J.Albrecht 17

PROPAGATION EFFECTS ON DIGITAL COMMUNICATION IN AVIONICS
by E.Lampert 18

MODELLING OF PROPAGATION ASPECTS OF DIGITAL COMMUNICATION SYSTEMS
by H.R.Raemer 19

PERFORMANCE PREDICTIONS AND TRIALS OF A HELICOPTER UHF DATA LINK
by R.M.Harris 20

NEW INSIGHT INTO IONOSPHERIC IRREGULARITIES AND ASSOCIATED VHF/UHF SCINTILLATIONS
by J.Buchau, E.J.Weber and H.E.Whitney 21

MULTIPATH PROPAGATION MEASUREMENTS BY DOPPLER TECHNIQUE
by P.Form and R.Springer 22

A CHANNEL SIMULATOR FOR L-BAND SATELLITE-MOBILE COMMUNICATIONS
by P.D.Engels 23

INVESTIGATION ON INFORMATION ERROR CAUSED BY TRAFFIC LOADING IN APPROACH AND LANDING SYSTEMS
by W.Skupin 24

Paper 25 not Available

SESSION IV – SPECIAL DEVICES FOR DIGITAL COMMUNICATIONS SYSTEMS

NEW DEVICES FOR DIGITAL COMMUNICATIONS IN AVIONICS
by F.I.Diamond, H.J.Bush and J.Graniero 26

Paper 27 Cancelled

TRANSFORM DOMAIN PROCESSING FOR DIGITAL COMMUNICATION SYSTEMS USING SURFACE ACOUSTIC WAVE DEVICES
by L.P.Milstein, D.R.Arsenault and P.Das 28

AN ANALYSIS OF THE ERROR PROBABILITY OF AN ALL DIGITAL DETECTOR
by S.Reisenfeld and K.Yao 29

SESSION V – SOURCE ENCODING AND DATA COMPRESSION

ASPECTS OF SOURCE ENCODING
by D.Wolf 30

PROBLEMS IN COMBINING SOURCE AND CHANNEL CODING
by H.J.Matt 31

SEGMENTATION OF PICTURES INTO CHANGING AND MOVING PARTS FOR FRAME REPLENISHMENT CODING TECHNIQUES
by J.Klie 32

Reference

SESSION VI – MULTIPLE ACCESS

STATE OF THE ART IN DIGITAL SIGNAL PROCESSING WITH APPLICATIONS TO MULTIPLE ACCESS SYSTEMS
by L.A.Gerhardt 33

MODEM TELEGRAPHIQUE A ETALEMENT DE SPECTRE
par D.Brisset et G.Auger 34

THE PERFORMANCE OF CODE DIVISION MULTIPLEXING WITH PULSE POSITION MODULATION
by J.Lindner 35

A TERMINAL ACCESS CONTROL SYSTEM FOR FLEETSAT
by S.L.Bernstein 36

IMPLEMENTING JTIDS IN TACTICAL AIRCRAFT
by D.R.McMillan 37

TDMA FOR RELAYED COMMUNICATIONS
by D.L.Baerwald 38

Paper 39 not Available

RESEAU DE RADIOCOMMUNICATION NUMERIQUE EN DUPLEX TEMPOREL
par J.Lautier 40

A MULTI-GBIT/S RZ-FORMAT DIODE MULTIPLEXER
by U.Barabas 41

A 16 KB/S MODEM FOR SECURE VOICE SERVICE OVER NARROWBAND ANALOG CHANNELS
by R.A.Northrup, T.R.Losson, D.D.McRae and F.A.Perkins 42

DOUBLE DIFFERENTIAL PSK SCHEME IN THE PRESENCE OF DOPPLER SHIFT
by M.Pent 43

THE IMPACT OF DIGITIZATION ON MILITARY COMMUNICATIONS

Irwin L. Lebow
Chief Scientist - Associate Director, Technology
Defense Communications Agency
Washington, D. C. 20305

SUMMARY

All communications systems are going digital largely to achieve reduced investment and operations and maintenance costs. This trend toward digitization will permit the evolution of military communications so as to provide more military attributes such as security, robustness and interoperability than have been affordable in the past.

The U.S. Defense Communications System (DCS) is upgrading its several subsystems which provide switched services and dedicated transmission to create a second-generation system in the mid 1980's with the common theme of digitization. One consequence of this upgrading will be a greatly improved secure voice capability. Based upon a new common-user secure voice system, AUTOSEVOCOM II, used in conjunction with dedicated satellite transmission employing transportable terminals for robustness, this system is being configured so as to make it maximally interoperable with tactical, civil and Allied secure voice systems.

This paper describes this second-generation DCS with emphasis on the technological issues underlying the design of the improved secure voice capability.

1. Introduction

All communications systems, civil and military, are going digital. The only issue is how long the transition from analog to digital will take. There are many reasons for this transition; there are for example, performance and flexibility features of digital systems which are unobtainable in analog communications. But the fundamental, overiding issue impelling the transition is simply cost. Digital communications are cheaper both in initial cost and operations and maintenance expenditures due in large measure to the rapid evolution of digital components.

In commercial areas, lower cost translates in one way or another to higher profits either by stimulating sales of given services as a result of the reduced costs or by generating demand for new services which become economical by virtue of the new, cheaper technology. In military communications, we have a different situation. It is a truism to say that military communications are developed to satisfy military requirements. But in the implementation process there is constant tension between the desire to make military communications more responsive to military needs and the need to control costs. Our current military communications systems, for example, have less robustness than they might have, have more clear mode and less secure mode traffic than desirable, are less interoperable with one another than they might be, largely because of the need to control costs. In my view, the greatest benefit of digital communications to the military is the fact that it promises to permit an evolution toward communications systems more responsive to a spectrum of military needs.

These comments are made from the point of view of one concerned with a wide spectrum of military communications extending from systems which are narrow in scope dedicated to very specific military requirements to the very broad Defense Communications System. The DCS, as it is usually abbreviated, provides long haul communications for a wide variety of military functions in situations extending from peacetime through crises to wartime. Many of the communications services provided by the DCS in peacetime bear a strong resemblance to non-military communications, so much so, in fact, that there is strong temptation driven by economic exigencies to forget that the DCS exists for military reasons. The trend to digital communications provides the opportunity to upgrade the DCS to a system with military attributes commensurate with the needs of the 1980's in an economical way.

The DCS was established in 1960 and is currently in the midst of a major transition to a second generation system to reach final operational capability in the mid 1980's. The upgrades associated with this transition have the objective of making the DCS more responsive militarily. While these upgrades are varied, they have the common technical theme of digitization. Without the availability of modern digital technology, the improvements being sought could probably not be achieved and even if achievable would be out of the question economically.

One of the most important components of military communications and one of the most difficult to achieve is flexible secure voice communications. Our current communications are limited in this regard and several of the DCS upgrades together with corresponding upgrades of tactical systems have the goal of improving our secure voice capabilities. The improvements are occuring as result of and in the midst of the rapid advancement of digital technology.

This paper will attempt to describe the kind of improvements in military communication capabilities which digitization facilitates using as an example this improved "Secure Voice Architecture." In Section II we provide a brief overview of the architecture problem and of the DCS and its new subsystems. Section III describes the elements of the new secure voice capability, its attempt to achieve a higher degree of interoperability and robustness, and its dependence upon digital technology. Finally, Section IV presents some concluding remarks.

2. Defense Communications System Overview

2.1 General Comments

The Defense Communications System as the long haul communications system of the Department of Defense serves a multiplicity of functions. It provides switched, common user clear voice, secure voice, message and data service to DoD posts, camps and stations throughout the world. It also provides dedicated point to point or special-purpose, transmission both terrestrial and satellite for connectivity to a wide variety of user communities. Its mission is to provide these communications services over the spectrum of hostilities ranging from day-to-day peacetime to wartime.

An adjective often used to describe the DCS is strategic as a way of differentiating it from the tactical communications carried out by each of the Services according to their missions. The rough definition of strategic in this context designates the communications at high levels of command. At this highest level are the communications serving the so-called World Wide Military Command and Control System (WWMCCS) which connect the National Command Authorities to the several Unified and Specified Commanders in Chief (CINCS) and their subordinate commands. As an example, the DCS provides connectivity from the Pentagon to Army Corps level; communications from Corps to Division and lower are provided by the Army and are designated "tactical." Interfaces at similar levels are provided between the DCS and the tactical communications of the Air Force and Navy.

It should be clear even from this very brief description that certain properties are of prime importance in the configuration of the DCS. One of these properties is robustness or survivability. By this we mean the ability of the system to provide essential communications over a spectrum of hostilities. We do not imply the ability of any individual elements of the system to survive but rather the ability of the system as a whole using redundant and diverse assets to provide the basic connectivity.

A second crucial property demanded of the system is interoperability. Not only does the DCS serve its own "subscribers" but it also serves the subscribers of the several tactical communications systems operated by the Services (Figure 1). For example, an Army Division Headquarters in Europe in peacetime garrison status is a subscriber to the DCS's Automatic Voice Network (AUTOVON) subsystem for clear voice connectivity anywhere in the world. In time of war, this Division headquarters becomes a field unit and connects to AUTOVON via Corps level using tactical transmission. Closely associated with interoperability is flexibility. Under different circumstances commanders may have the need to jump the chain of command, in the extreme from the President to the man in the fox hole. The DCS in providing connectivity to the WWMCCS must provide this level of flexibility under a wide spectrum of conditions.

Indeed, so significant has become the interdependence of the DoD systems (and in some cases, civil and Allied systems as well), that it no longer makes sense to consider upgrading the capabilities of the DCS without concurrent consideration of similar capabilities in the tactical communities. Thus there has arisen the perception of a need for the development of communications "architectures," that is sets of end-to-end communications capabilities for DoD as a whole which translate into specific interoperable capabilities of the several component systems. Thus while there may be good and sufficient reasons for systems to remain distinct in a management sense, they cannot be allowed to evolve separately in a technical sense. For this reason, along with the upgrades associated with the second generation DCS, there have been initiated architectural efforts which extend beyond the DCS boundaries in a technical sense. In most cases the specific subsystem upgrades were conceived before the need for an "architecture" was fully perceived, making the achievement of a logical architecture that much harder.

2.2. The Current DCS

The current first generation DCS has been in a state of evolution since its inception in 1960 as a vehicle for unifying the then separate long-haul facilities of the three Services. It contains Continental U.S. and overseas components, differentiated by the fact that the former are primarily leased and the latter primarily government-owned. The government-owned assets of the system are operated and maintained and hence "owned" by the three Services. Indeed, in any given facility will be found both DCS and non-DCS assets operated and maintained by the Service responsible for the facility. While operational direction and management control of the DCS is the responsibility of the Defense Communications Agency, the DCA "owns" only the operations centers in the U.S. and overseas.

The DCS provides both general-purpose or common user and special purpose or point-to-point services. In the former category are the three switched systems: Automatic Voice Network (AUTOVON) providing clear voice; Automatic Secure Voice Communications (AUTOSEVOCOM) providing secure voice; and Automatic Digital Network (AUTODIN) providing secure data and message service.

Transmission in the DCS is provided by a multiplicity of media both government-owned and leased including several terrestrial media as well as satellite relay. The same transmission serves both the common user and the point-to-point applications. Thus, in any DCS facility the transmission will serve both AUTOVON, AUTODIN, and AUTOSEVOCOM switches as well as individual bases or tactical elements.

Some of the pertinent data for the switched systems and for the transmission are shown in Table 1. The division between leased and government-owned transmission is shown in

Figure 2. Figure 3 shows the geographical distribution for the AUTOVON, the most extensive of the switched systems.

The current DCS is almost exclusively analog, with the basic transmission element the 4 kHz voice frequency (VF) channel. Digital data is carried in quasi-analog form in the channels. Multiplexing is almost exclusively by frequency division in both the satellite and terrestrial components of the system.

2.3 The Second Generation DCS

Almost every subsystem of the DCS will be upgraded in one way or another in the second generation. (LEVINE, R.H.., 1976; SHIMABUKURO, T..,1976; SCHULTZ,D.., 1976; COVIELLO, G..,1976) Table 2 lists the major elements of both the transmission and common-user subsystems showing the transition from the first generation at the left to the second generation at the right. The major features of the upgrades are described as follows:

(a) DIGITAL TRANSMISSION

As we have already point out the government-owned transmission, (SCHULTZ, D.., 1976) both terrestrial and satellite is going digital. The first phase of this has already begun in Europe with highly encouraging results. The so-called Digital European Backbone project will result in the digitization of most of the European DCS trunking facilities by the mid 1980's. A similar program is being scheduled for the Pacific DCS. Figure 4 is our current best estimate of the extent of the digitization of the European terrestrial DCS plant by 1985. The CONUS transmission, being leased, will follow the commercial plant in its conversion from analog to digital. It is expected that substantial digitization will occur in the 1980's.

The satellite transmission subsystem of the DCS is being digitized in a similar fashion beginning next year with the introduction of a digital communication subsystem into the DSCS earth terminals. By the time the DSCS-III space segment succeeds the DSCS-II space segment in the early to mid 1980's, the transmission will be virtually all digital.

The digital transmission design is transitional in that it permits interfacing with either analog or digital terminals, transmission and switches. As shown in Figure 5 it interfaces VF channels directly, digitizing the channel signals with 64 Kb/sec pulse code modulation. It also provides digital interfaces in conformity with US and in part European standards. The VF interface permits the mixing of the new DCS transmission with analog transmission of the U.S. and PTT systems. Even though the DCS will be mostly digital in this time frame, there will be residual analog transmission with which it will have to interface for the foreseeable future.

(b) DSCS-III

The Defense Satellite Communications System provides satellite transmission for the DCS and several other communities. The first R&D launches of the third generation space segment called DSCS-III (SCHEMMER, B.F..,1978) will take place over the next several years. It differs from its predecessor DSCS-II in three major ways:

(1) Anti-jam capability - DSCS III contains the capability of discriminating against hostile emitters.

(2) Control - The communications parameters of the spacecraft can be controlled by communications terminals in addition to the Air Force Satellite Control Facility terminals built for satellite housekeeping control.

(3) Downlink efficiency - through the use of bandwidth channelization and multiple-beam transmit antennas, the available effective radiated power of the satellite is shared efficiently among user communities and directed toward areas on earth in which users are located.

The satellites are currently used almost exclusively for inter-area trunking as shown in Table 2. The ability of the satellite to direct power where most needed as cited above will permit the DSCS-III to serve some intra-area as well as inter-area functions. Some of the implications of this will be shown in the next section.

(c) SWITCHED SYSTEMS

The three switched systems (SHIMABUKURO, T..,1976) in the DCS are evolving in different ways into the second generation. AUTOVON, as a system, is now planned to continue for the time being in much the same way as currently. AUTOSEVOCOM will be upgraded in a major way to serve a much larger subscriber base and it will be treated in more detail in Section III. A second generation data system, AUTODIN II, (GORDON, S.H.., 1977) is now being developed for CONUS as a systems overseas will, of course, use the new digital transmission facilities described above.

A new switch will be introduced into the overseas DCS to serve AUTOVON and AUTOSEVOCOM II. This switch, the AN/TTC-39 (BLACKMAN, J.A..,1976) under development by the Joint Tactical Communications Activity (TRI-TAC) for tri-Service use, has both analog and digital capabilities and thus also serves a transitional role. The DCS switches will use the analog capability for upgrading overseas AUTOVON and the digital capability for AUTOSEVOCOM II. It is expected that in a later time frame, the analog facilities will be replaced by augmented digital facilities as the transition to digital is completed.

No changes are now programmed for the CONUS portion of AUTOVON in this time frame. The issues underlying the design of the CONUS component of AUTOSEVOCOM II will be discussed in detail in the section which follows.

3. Secure Voice Architecture

3.1 Introduction

With the foregoing overview of the second generation DCS as background we are now prepared to consider the specific issues involved in developing a more militarily responsive secure voice capability centered on the upgraded AUTOSEVOCOM common user system augmented by selective use of the upgraded dedicated transmission.

The current AUTOSEVOCOM has only some 1400 subscribers. It has two components: (1) wideband using pulse code modulation at 50 k bits/sec to digitize the voice signals, and (2) narrowband, using mostly the channel vocoder to convert the voice signals to a 2.4 k bit/sec bit stream. The wideband portion yields excellent speech quality, but its use has been limited mostly to local areas by the cost of transmission. The narrowband portion, in contrast, can easily be served by both DCS and commercial VF transmission facilities, but it's use has been limited by generally unsatisfactory voice quality.

Both components of AUTOSEVOCOM I are old. In the years since their initial development great strides have been made in both voice processing and transmission technology to serve as the bases for a new system. But, as we pointed out previously, it is not sufficient to consider the configuration of AUTOSEVOCOM II independently of the other systems with which it must interoperate. Indeed, as we shall show, the major constraint imposed upon the design of AUTOSEVOCOM is the necessity that it interoperate effectively with U.S. tactical communities as well as certain civil and allied communities.

We shall not attempt to be exhaustive in the consideration of interoperability. It will be sufficient to consider the two main U.S. tactical communities with evident need to communicate with the DCS, and the discussion will bring out the technical issues. The largest of these tactical systems is that under development by TRI-TAC for use by all the Services. The TRI-TAC architecture includes the various components needed for a tactical switched system to handle clear and secure voice as well as data. For reasons which will become apparent later, this community is called "Wide-band tactical."

A second tactical community is distinguished from the above by being "narrowband." This community found mostly but not exclusively in the Navy is constrained by the limited bandwidth available on high frequency radio links and by the limited power on UHF satellite communications circuits.

As we shall show, achieving secure voice interoperability with both of these communities is difficult at the current state of technology. We shall discuss the reasons for this as well as the different approaches to interoperability available for consideration.

3.2 Voice Processing

The complexities of the secure voice architecture are directly traceable to complexities in voice processing and the adaptability of voice processors to the various classes of communications channels. In this section we present a brief summary of the voice processing state of the art (MIT LINCOLN LABORATORY, 1976) and in the following section we shall consider communications media. There are three voice processing categories of interest operating in different ranges of digitization rate, R, described as follows:

(a) Simple Wave Form Processors. $9.6 \text{ Kb/Sec} \leq R \leq 64 \text{ Kb/Sec}$

This class includes the straightforward pulse-code modulation (PCM) processors at rates in excess of 48 Kb/sec and the slightly more sophisticated adaptive PCM and adaptive delta-modulation schemes nearer the lower end of the range. All devices in this category are simple enough to be realized in light-weight, low-power packages. The quality varies from excellent at the high end to highly intelligible, if slightly noisy near 16 Kb/sec, to unpleasantly noisy but still fairly intelligible at the low end. Figure 6 shows this graphically. In the figure the quality scale is intended to be qualitative and thus to indicate trends and relative performance.

(b) Complex Wave -Form Processors. $6 \text{ Kb/sec} \leq R \leq 16 \text{ Kb/sec}$

This class includes schemes such as Adaptive Predictive Coding (APC), Residual Excited Predictive Coding (RELP) and the Voice Excited Vocoder (VEV). They are similar in form to some of the analysis - synthesis systems in category 3 below but do not require the derivation of vocal chord or pitch information. These processors are considerably more complex than those in class 1, but the additional complexity is shown in Figure 6 to buy improved quality at the lower rates. Their quality can be very good

at 16, good at 9.6 degrading near the lower end of the range. In fact, categories 1 and 2 can be thought of as bounds for the whole class of waveform processors.

(c) Analysis - Synthesis Processors. 2.4 Kb/sec $\leq$ R $\leq$ 6Kb/sec

This class includes the venerable channel vocoder and the more recently developed linear predictive coding (LPC) vocoder. Both model the vocal tract (mouth, tongue, etc.) with a time-varying filter excited by either a pulse source modeling the vocal chords or a noise source for unvoiced sounds. Their complexities are somewhat greater than in the previous class with the channel vocoder somewhat more complex than the linear predictive. The quality of the synthesized voice is intelligible, but artificial or synthetic sounding. These systems are the least robust in that they are the most talker dependent and deteriorate most rapidly in the presence of both background and channel noise. An indication of their quality vs. rate dependency is also shown in Figure 6.

3.3 Communications Channels for Secure Voice

The communications channels over which digital voice signals are transmitted can also be divided into three categories:

(a) Narrowband Channels - R $\approx$ 2.4 Kb/sec

This rate is generally characteristic of that supportable in an HF channel or a low powered tactical satellite link. This channel can accommodate only class 3 voice processors, as indicated by the HF limit range in Figure 6.

(b) Analog Voice Frequency Channels - R < 16 Kb/sec

The ubiquitous 4 kHz VF channel comes in all quality ranges. Almost all can accommodate 2.4 Kb/sec data rates with error probabilities sufficiently low for vocoders. As we move up in data rate, a substantial fraction of the channels will accommodate 9.6 Kb/sec with the use of contemporary modems. A recently developed highly sophisticated modem (NORTHRUP, R.A.., 1978) will permit speech transmission at rates as high as 16 Kb/sec with error rates $\sim 10^{-2}$ over some VF channels, but it is too early to state the extent to which this is achievable in the commercial telephone plant. Thus, as we see in Figure 6, the VF channel can accommodate all vocoders in category 3 of the previous section, almost all class 2 systems and, to some perhaps limited extent, the lower rate simple waveforem processors.

(c) Digital Channels

Digital service is now available in many metropolitan areas in the U.S. and is being extended to many more. It is usually in the form of multiples of 1.544 Mb/sec trunking with submultiplexing for individual subscribers. Thus, wherever available, digital service will accommodate all classes of voice processors. Digital service is becoming available in other countries as well, some more and some less rapidly than in the U.S.

Ground tactical forces carry their own transmission with them. This is largely digital now and thus no limiatations on data rate are imposed upon secure voice service for these tactical forces.

3.4 Voice Processors Under Development

The tactical forces have standardized on a waveform processor called Continuously Variable Slope Delta Modulation (CVSD) for use in single channel radios at 16 Kb/sec and in tactical trunked service at 16 and 32 Kb/sec in the TRI-TAC family. As evident from the above discussion, it is compatible with the physical constraints on this community and its quality is good. The TRI-TAC terminal is referred to as the Digital Secure Voice Terminal (DSVT). It is shown for reference on Figure 6.

A second terminal designated the STU-2, is under development for VF channel use and will realize APC at 9.6 Kb/sec and LPC at 2.4 Kb/sec. These two algorithms are sufficiently alike to permit dual-mode implementation in the same microprocessor-based hardware. As noted above, the higher rate, higher quality algorithm should work over most circuits with the lower rate LPC serving a backup role. At the current state of the art in digital hardware, the STU-2 is too large and consumes too much power to be considered for the mobile tactical applications.

A third terminal, designated the Advanced Narrowband Digital Voice Terminal (ANDVT) is being developed for the narrowband tactical community. It will be functionally like the 2.4 Kb/sec mode of the STU-2. It will thus be slightly less complex than the STU-2 but still considerably more complex than the DSVT. It is however compatible with the primarily shipboard application. Both STU-2 and ANDVT are shown in Figure 6.

3.5 Interoperability - General Considerations

Now, having summarized the current state of voice processors, we can address their implications on interoperability. Interoperability is not a binary function. There are many ways of achieving interoperability between communications systems and this section will review some general principles before addressing voice systems in the next section.

We will consider the three generic classes of techniques for achieving interoperability shown in Figure 7. The usual image of interoperable systems is shown in Figure 7a. The two systems designated A and B are functionally sufficiently alike that they can be connected together without requiring an interface for technical reasons. Thus, the indicated connectivity between the systems is a management rather than a technical problem. We call such systems conformable. In the extreme, if a management decision is made to allow arbitrary connectivity between the two conformable systems then the two systems become a single system.

If the systems are not sufficiently alike to be considered conformable then some kind of gateway is required to interface the two disparate systems as shown schematically in Figure 7b. The gateway has the property of modifying the system interface in such a way that connectivity is technically achievable. The number of gateways is, of course, variable depending in general upon their cost and complexity.

When the systems are so unlike that gateways are not feasible, then it is still possible to achieve interoperability as in Figure 7c., by overlays in which selected subscribers of each system are made subscribers of the other system. In a telephone system this form of interoperability is referred to as "two-phone" interoperability.

3.6 Secure Voice Interoperability

When secure voice systems use voice processors which employ the same algorithm for voice digitization then conformable interoperability is possible. Even in this case, however, other system differences such as, for example, in the crypto logic may necessitate a gateway-type interoperability.

However, when the voice processing algorithms used in the two systems are unlike, then, as a minimum, a gateway is required. This gateway includes a tandem connection in which the digitized voice from one system is converted to analog (or equivalently to a digital representation of the analog voice) and then redigitized with the voice processing algorithm used by the second system.

With the exception of PCM used at the highest bit rates ($\approx$50 Kb/sec), all of the voice processing techniques are nonlinear. When these nonlinear operations are cascaded in the tandem connection, the voice quality degrades rapidly. Generally speaking, the better the quality of the individual processors, the better the quality of the tandem. Thus, 50 Kb/sec PCM in tandem with any other process preserves the quality of the other process. However, 32 Kb/sec CVSD, by itself a good quality process, noticeably degrades the quality of other processors with which it is tandemed. At the other extreme, the slightly noisy but otherwise good quality 16 Kb/sec CVSD produces such significant degradation when tandemed with 2.4 Kb/sec LPC as to render the result insufficiently intelligible for many applications.

We are thus led to the conclusion that gateway interoperability may be suitable for 32 Kb/sec CVSD and one of the lower rate systems but that 16 Kb/sec CVSD gateways are not acceptable. Thus, when one of the systems to be interconnected uses 16 Kb/sec CVSD, two-phone or overlay interoperability is required.

3.7 AUTOSEVOCOM II

While the final configuration of AUTOSEVOCOM II has not yet been determined, the foregoing discussion has delineated the piece-parts of the set of possible configurations, and the following gross properties can be stated:

(a) The overseas portion of AUTOSEVOCOM II will be primarily TRI-TAC-like in configuration. The overview in Section II has already noted this fact. Its rationale is the fact that in any conceivable wartime situation the DCS and tactical systems must be treated as virtually a single system. We have already pointed out that when interoperable systems are conformable, the amount of interconnectivity is inhibited only by management not technical considerations. In fact, present doctrine states that during military crises the overseas tactical and strategic systems shall be treated as a single entity with single management provided by a theater commander. With conformable systems using, in fact, the same equipment this unified management can be realized. This configuration is technically and economically viable because of the TRI-TAC switch development and the DCS digital transmission upgrades both terrestrial and satellite.

(b) The overseas portion will contain a probably thin overlay of either STU-2 or ANDVT equipments to permit two-phone interoperability of the DCS with the narrowband tactical community primarily in the Navy.

(c) The CONUS portion of the system is the least defined. Under any circumstances it will contain two components, a 16 Kb/sec TRI-TAC-like segment to permit conformable interoperability with the overseas DCS and with the TRI-TAC community, and a segment using the STU-2 terminal which will provide conformable interoperability with the tactical narrowband community and with a civil secure voice community. What is at issue are the relative sizes of the two components.

Two polar configuration alternatives can be cited:

(1) The CONUS configuration is similar to the overseas - i.e., a

digital system is created in the CONUS functionally similar to the TRI-TAC-like overseas portion. The digital system consists of several AUTOVON switches augmented with digital facilities similar to those of the TTC-39 and internetted with digital transmission. Access trunking is provided digitally where such exists and analog elsewhere using the new 16 Kb/sec modem. Limited narrowband service is provided as an overlay.

(2) The CONUS configuration is primarily narrowband analog with a TRI-TAC-like overlay for the most important command and control and related users.

The first configuration (Figure 8a) results in a homogeneous worldwide digital system conformable with the TRI-TAC-like tactical systems and interoperable by overlay with narrowband tactical and other narrowband systems. The second configuration (Figure 8b) results in a hybrid, digital overseas but analog in CONUS, with only the CONUS overlay providing conformable connectivity to overseas systems. The first is conceptually simpler and clearly superior in performance. It appears to be more expensive however, given the fact that CONUS DCS is leased and dependent in its costs on the developments in the commercial world. But since we are developing the military system in the midst of great changes in the digital world, the relative costs of digital or analog implementations, wideband or narrowband, depend crucially on the rate of digitization of commercial communications and the rate at which digital componentry continues to cheapen over the life cycle of the new system.

However the CONUS configuration is ultimately decided, either alternative, each in its own way, will provide an improved level of interoperability contributing to achieving our major objectives in the second generation architecture.

It follows from all of the above discussion that interoperability between AUTOSEVOCOM II and the older systems must be by overlay. As more of the older systems, U.S. or Allied, are phased out and replaced by systems conformable with the AUTOSEVOCOM elements, especially the DSVT, then the overall interoperability picture will be improved.

3.8 Transmission Redundancy for Robustness

Up to this point, the dominant feature of the discussion has been interoperability and its cruciality in determining the AUTOSEVOCOM II configuration. The other military issue to be confronted is robustness or survivability.

AUTOSEVOCOM by its very definition is a common-user system designed to serve a variety of users in a variety of situations. Within AUTOSEVOCOM, as within the other DCS switched subsystems, recognition is given to the fact that all users are not equal in importance. Higher importance users are given higher precedence and preemption rights at switches than others, thus guaranteeing service in times of crisis when the traffic level tends to escalate. In this regard, it is important to note that congestion can occur in switched systems for several reasons having nothing to do with military situations, e.g., Mothers' Day, power outages, severe storms.

A minimum of survivability is additionally provided to the more important users by multiple homing them to more than a single switch to provide protection against a debilitating failure at a single point due to natural causes or enemy action. In addition, a switched system provides robustness by virtue of its multiplicity of routing and connectivity.

Another approach to robustness in a general-purpose system is the use of mixed media. As we have pointed out, the DCS today is largely terrestrial-transmission oriented in its intra-area components with a satellite cable mix for inter-area trunking. With the maturing of satellite communications technology, more satellite communications will undoubtedly be used for intra-area trunking as well as inter-area, thus increasing the redundancy of the system and hence its survivability. General-purpose peacetime demands for capacity as well as flexibility force the DCS to employ the larger fixed terminals for these applications which in turn limits the survivability improvements obtainable in an escalated military situation.

An additional survivability improvement for important users can be made by overlaying a dedicated satellite transmission capability on top of AUTOSEVOCOM and its companion switched systems using small transportable terminals. This satellite overlay using both the DSCS III satellites and, if needed, UHF connectivity can add a dimension of robustness to that achievable by the interoperable switched general-purpose systems by themselves. The mobility of the small terminals and the anti-jam features of the DSCS III add significantly to this capability.

3.9 Secure Voice Conferencing

Perhaps the single most important command and control requirement is the secure voice conference. Given the interoperable switched and dedicated secure voice capabilities described above, the essential ingredients for this capability are present.

An additional complexity must be noted, however. The standard technique for conferencing is the so-called full-duplex conference bridge shown in Figure 9. The essential feature of this conference is the fact that the outputs of all voice transmitters are summed and then transmitted to all receivers. In this scheme, all parties can hear all other parties and any talker can interrupt at will.

The voice processing state of the art introduces a technical difficulty in applying the straightforward full-duplex conferencing technique to the secure voice case. First make the simplifying assumption that all conferees use the same voice processor. As shown in Figure 10, the full-duplex conference requires the conversion of each speech signal to analog before the summing process can take place with a redigitization before transmission to the receivers. This process, of course, introduces the tandem operation which has severe degrading effects even when identical processors are used. In the class of processors we have discussed, this technique introduces significant degradations except at rates in excess of 32 Kb/sec.

The solution to this problem is furnished by the so-called digital conference bridge. This is a name given to a class of techniques which replaces the summing and redigitizing operation in the full duplex bridge with a switching operation which selects one and only one of the signals for transmission to the receivers at a time (Figure 11); the switch, in essence, fools the conference participants into believing that they are in a full-duplex conference. The switch position may be controlled in a number of ways: e.g., manually by a chairman, or automatically by who talks first, who talks loudest, etc Several schemes have been tested in the laboratory and some in the field as well.

The difficulties are compounded when the talkers use disparate processors. One source of difficulty is the fact that the switching algorithm has to deal with more than a single format. A second source of difficulty occurs when the interoperability is by the two-phone technique in which case the participants with two-phones have two parallel channels into the conferencing bridges. All of these complexities can, however, be coped with.

4. CONCLUSIONS

We have thus shown that the next generation military communications will have an upgraded secure voice capability obtained by attacking the interoperability and survivability issues on a systems architecture basis making maximum use of the advancing secure voice and digitization technologies.

As can be seen, the situation is one where considerable improvements in capabilities are promised at considerable developmental pain. It also represents an improvement based upon a technology in the midst of rapid advances. It is very easy for the technologist to plead for a delay in system implementation because of the inevitable technological advances which would make later implementations better and cheaper. But it is just as hard to delay implementations when the existing capabilities are so antiquated and inadequate to the operational needs.

The burgeoning technologies that make the improvements possible lie in many areas: satellite technology, modulation technology, voice processing technology and others. But underlying them all is the revolution in digital component technology, for this affects every aspect of the communications system: switching, transmission and terminals.

Digital technology has already reached the point where, transition issues aside, it is cheaper than analog in almost all aspects. This is shown most eloquently in the commercial world where Canada in making massive increases to its telephone plant is using only digital switching and transmission at an anticipated savings of ½ billion dollars in investment and 1/3 billion dollars in operations and maintenance over a twenty year period. (HOUGHTON, R.N.E..,1976) In a country like the U.S. where a high investment in analog plant already exists, the pace of digitzation is slower although even here substantial digitization is underway in major population centers both in digital transmission and in digital switching through introduction of the 4-ESS switch. The independent telephone companies are rapidly replacing old plant with switches which implement all their internal functions digitally while maintaining VF channel interfaces externally.

The initial experience of the DCS with digital transmission has been similarly salutary both with regard to investment and operations and maintenance costs. The transmission and multiplex equipment now under development for the terrestrial digitization projects is about 16% cheaper in 1977 dollars than the equivalent analog replacement costs in 1974 dollars. Planned technical control facility upgrades should cost close to $1M less because of the simplifications brought about by transmission digitization alone. When, in the future, the switching can also be all digital, then savings in tech controls alone increase by a factor of 10.

Indeed the makeup of the CONUS portion of AUTOSEVOCOM II is, in its essence, an analog vs digital question. Were digitization in the CONUS commercial plant farther along than it is, then economies as well as performance would favor a digital solution for the majority of the subscribers regardless of data rate.

The secure voice terminal area is probably most influenced by technology evolution. In the narrowband area we have seen a reduction in costs by at least a factor of two from the old vocoders now in use to the developmental STU-2 due mostly to new components but also in part due to new speech compression algorithms. The technological optimists tell us that in just a few years these narrowband processors will be reduced in size and power consumption to the point where they can be used in mobile tactical applications. Some also predict that the voice quality produced will improve to the point where the 2.4 Kb/sec rate will be truly universal ending thereby all interoperability problems.

I believe that the size, weight optimism is justified but I wouldn't care to predict when it will happen. I am not among the optimists on the quality issue. For this to happen, a fundamental breakthrough in speech processing is necessary.

Even if the terminal size and hence cost reductions occur, it is still not obvious how secure voice will be handled in the third generation. We may well have a single conformable scheme for all but the narrowband tactical community but the economics could drive the rate either to the higher or lower end.

But this issue is not important except to the technological prophets. The important fact is that we are able to take advantage of digital technology today to improve the military properties of military communications. The opportunities in the next generation will be even greater with a more mature digital technology.

5. References

BLACKMAN, JEROME A., 1976, "Integrated Switching of Voice and Non-Voice Traffic," ICC, Vol II, pp (24-1)-(24-6).

COVIELLO, G. J., 1976, "Unifying System Engineering," Conference Record, ICC, ICC, Vol II, pp (33-18)-(33-23).

GORDON, S. H., 1977, "AUTODIN II System Overview," National Telecommunications Conference Record (NTC-77), p. 37:1.

HAUGHTON, R. N. E., 1976, "Digital! The Network of Tomorrow," Canadian Telecommunications Carriers Association, Presentation at Defense Communications Agency, 7 Jul 1976, (Unpublished).

LEVINE, R. H., 1976, "The Evolving Defense Communications System: Introduction and Overview," Conference Record, ICC, Vol II pp (33-2)-(33-4).

MIT Lincoln Laboratory, 1976, Annual Report on Speech Evaluation, ESD-TR-76-382, 30 Sep 1976.

NORTHRUP, R. A., 1978, "A 16 Kb/s Modem for Secure Voice over Narrow Band Analog Channels," AGARD Symposium on Digital Communications in Avionics, (this issue).

SCHEMMER, B. F. Feb 1978, "Strategic C^3: The Satellite Arena - 20 Year After Sputnik," Armed Forces Journal, P 18.

SCHULTZ, D., 1976, "DCS Transmission Network 1980-1982," Conference Record, ICC, Vol II, pp (33-11)-(33-16).

SHIMABUKURO, T., 1976, "The DCS Circa 1980-1982," Conference Record, ICC, Vol II, pp (33-5)-(33-10).

TABLE 1

DCS Statistics (1977)

TRANSMISSION MEDIA

TYPE MEDIA	NO. TRUNKS
LANDLINE	170
SUBMARINE CABLE	40
MICROWAVE	2,300
TROPOSPHERIC	300
VERY HIGH FREQUENCY	60
HIGH FREQUENCY	70
SATELLITE	100
TOTAL	3,040

SWITCHED NETWORKS

TYPE	NO. SWITCHES	NO. SUBSCRIBERS
AUTODIN	16	1,300
AUTOVON	85	17,000
AUTOSEVOCOM	113	1,400
TOTAL	214	19,700

TABLE 2

DCS Upgrades – Summary

TRANSMISSION	FIRST GENERATION	SECOND GENERATION
OVERSEAS	ANALOG	DIGITAL (HYBRID INTERFACES)
CONUS	MOSTLY ANALOG	PARTIALLY DIGITAL
INTER-AREA MIX	DSCS-II	DSCS-III
	COMMERCIAL SATELLITE, CABLE MICROWAVE	
INTRA-AREA MIX	SOME TROPO	DSCS-III

SWITCHED SYSTEMS		
CLEAR VOICE	AUTOVON	
CONUS SWITCHES	FIRST GENERATION AUTOVON	
OVERSEAS SWITCHES	FIRST GEN	TTC-39 (ANALOG)
SECURE VOICE	AUTOSEVOCOM I	AUTOSEVOCOM II
CONUS SWITCHES	FIRST GENERATION	DIGITIZED AUTOVON
OVERSEAS SWITCHES	FIRST GENERATION	TTC-39 (DIGITAL)
DATA	AUTODIN I	AUTODIN II

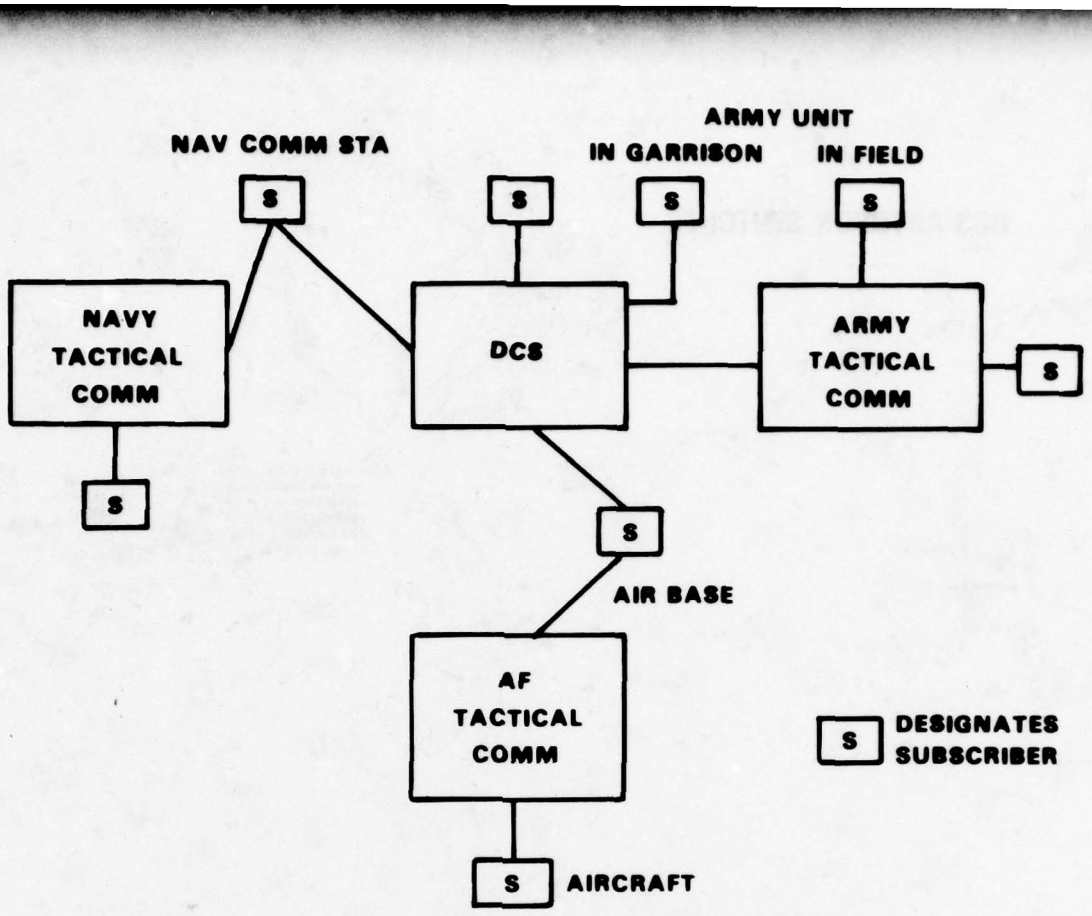


Fig.1 Interoperability of DCS with tactical elements

Defense Communications Agency

DCS TRANSMISSION MEDIA

48,000 CIRCUITS
35,000,000 ONE WAY CHANNEL MILES

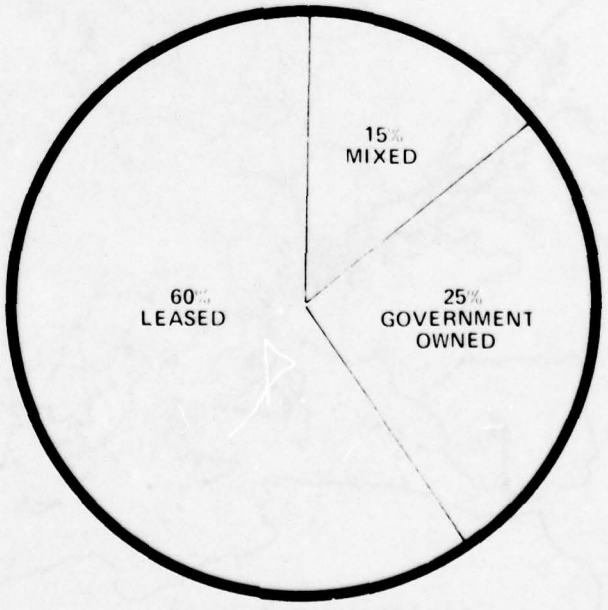


Figure 2

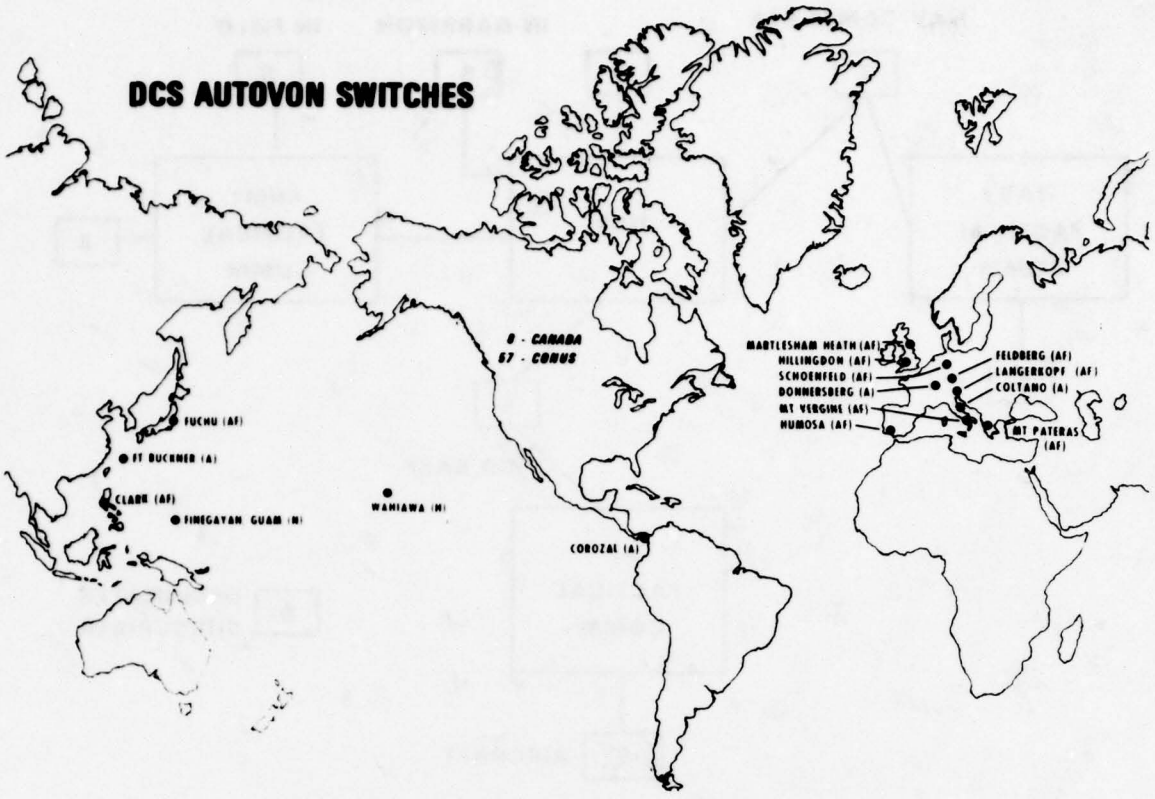


Fig.3 DCS AUTOVON switches

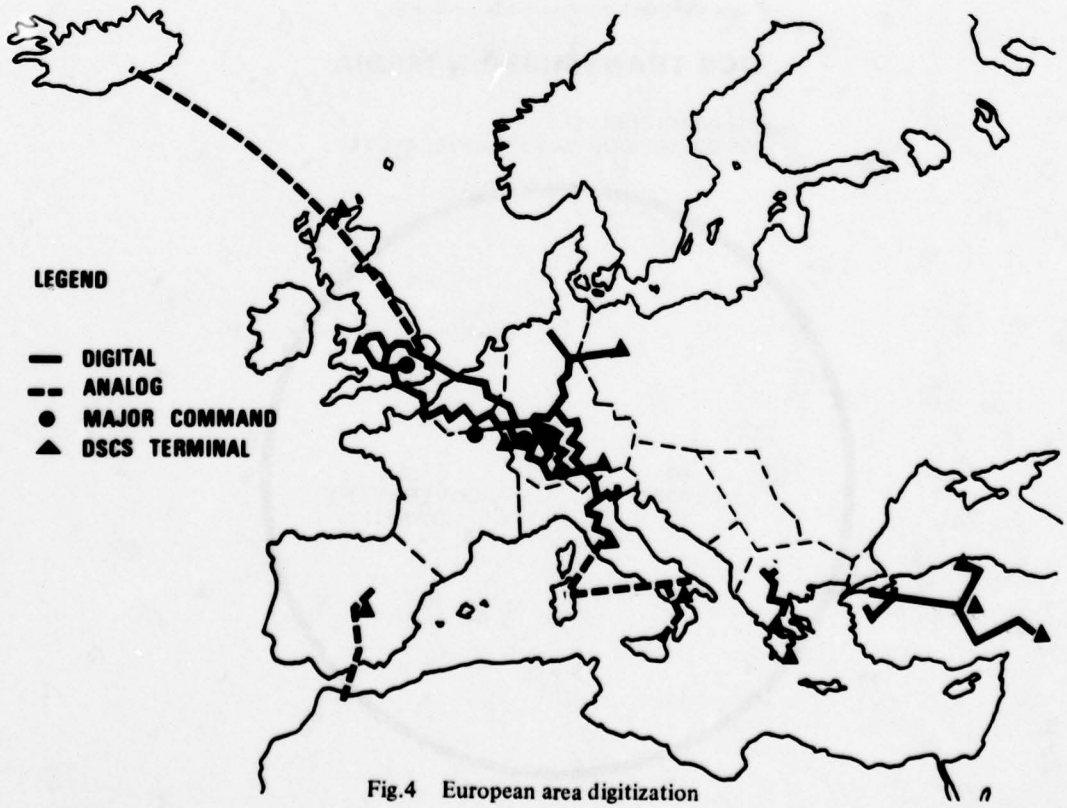


Fig.4 European area digitization

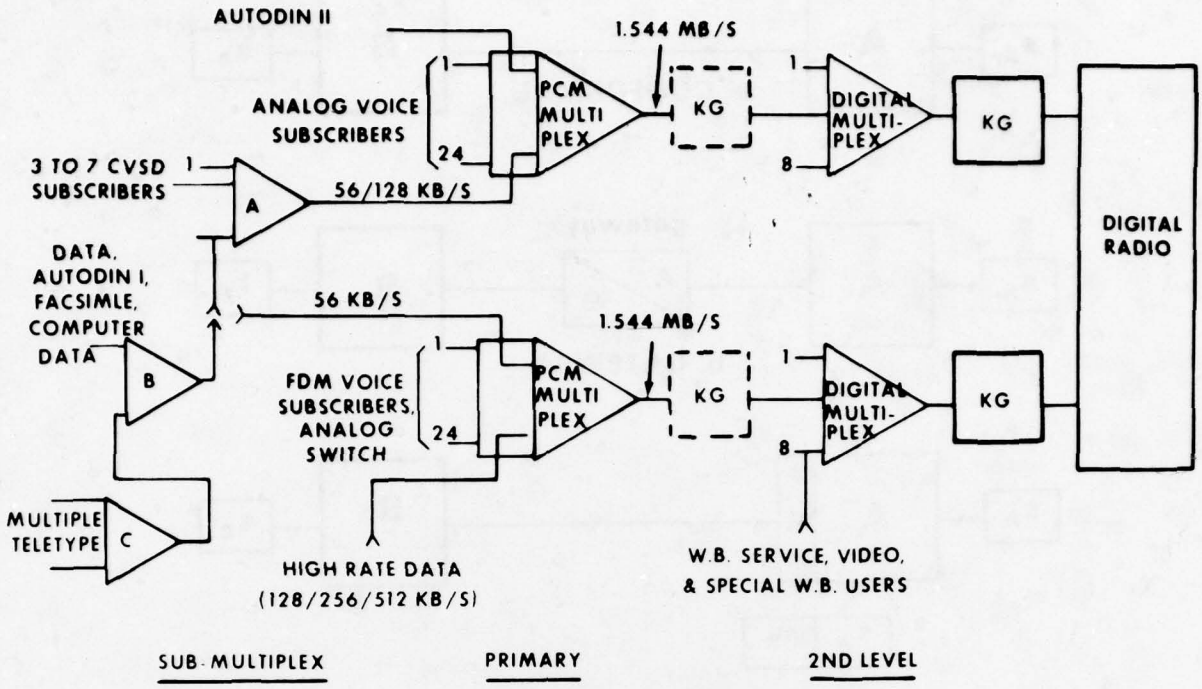


Fig.5 Digital transmission system configuration

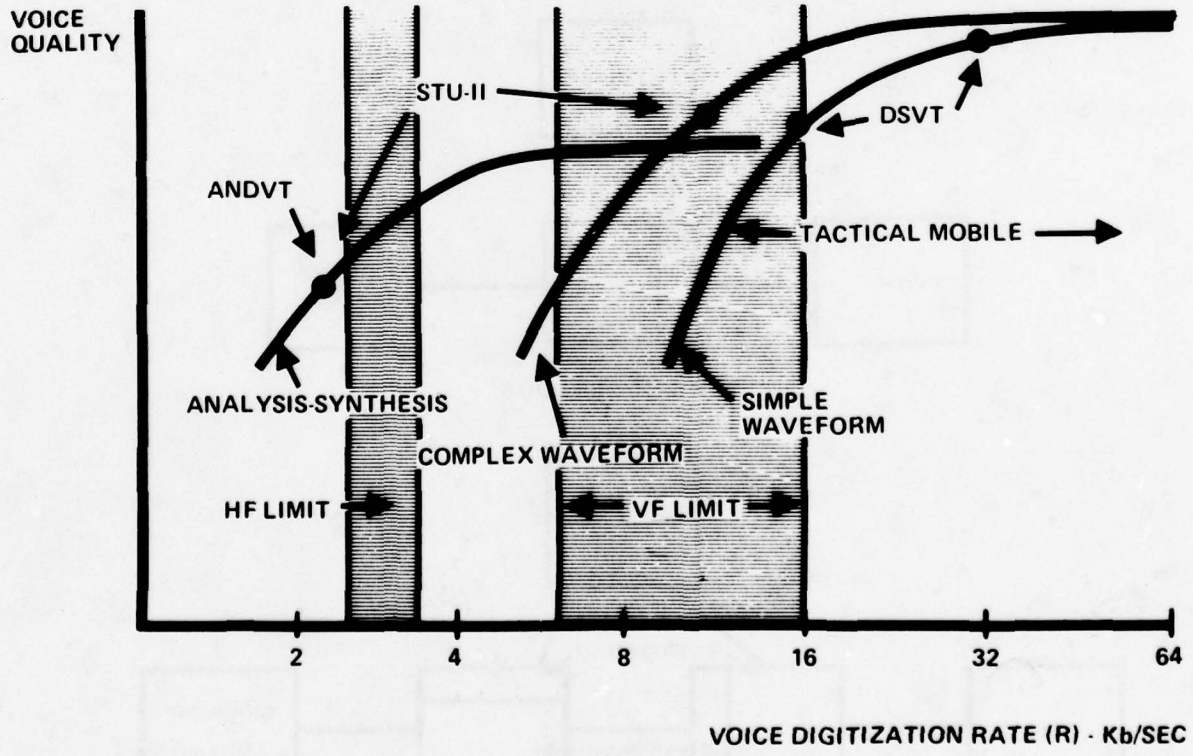


Fig.6 Voice processor quality

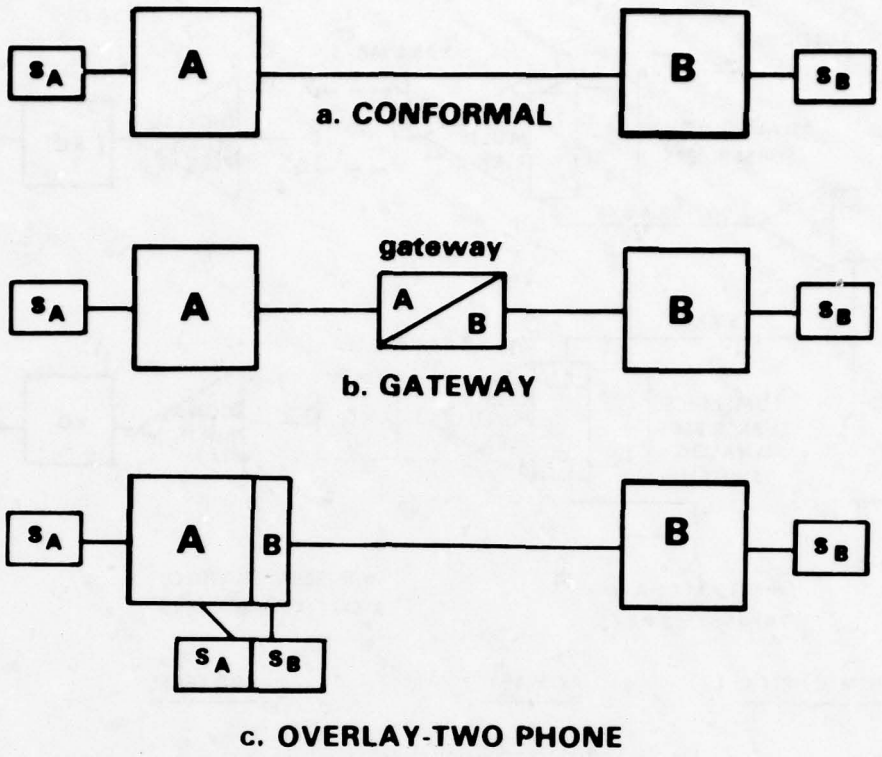


Fig.7 Interoperability

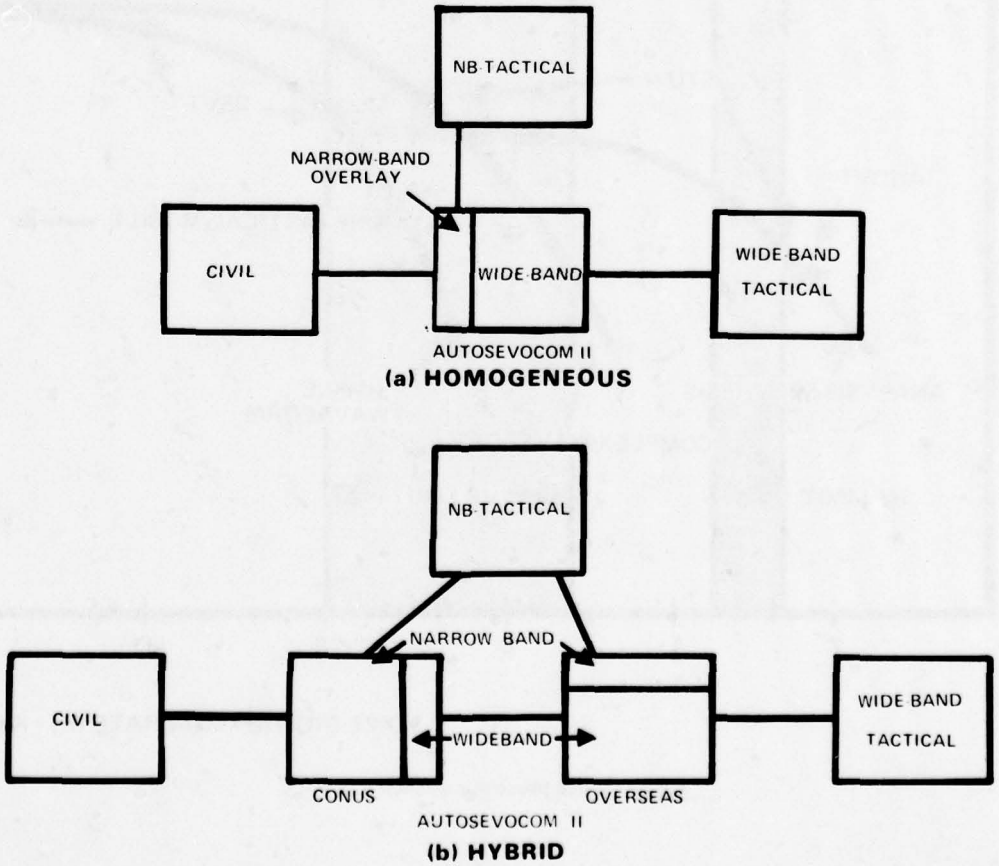


Fig.8 AUTOSEVOCOM II alternatives

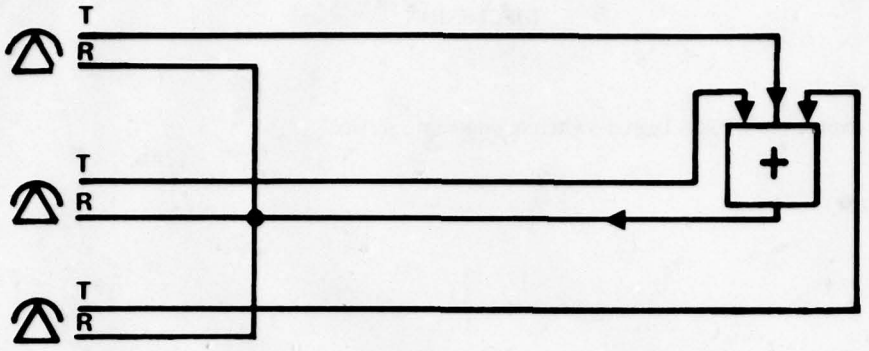


Fig.9 Analog full duplex conference

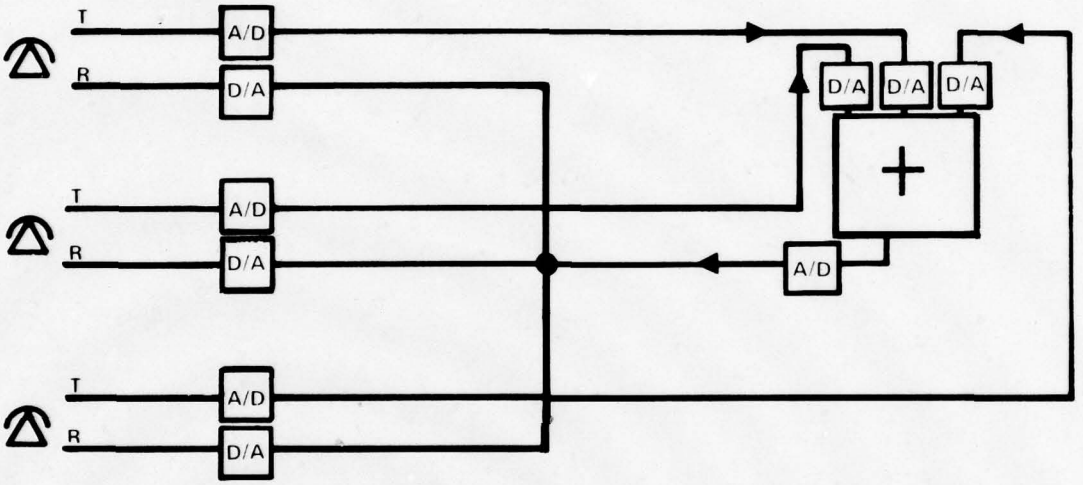


Fig.10 Digital full-duplex conference

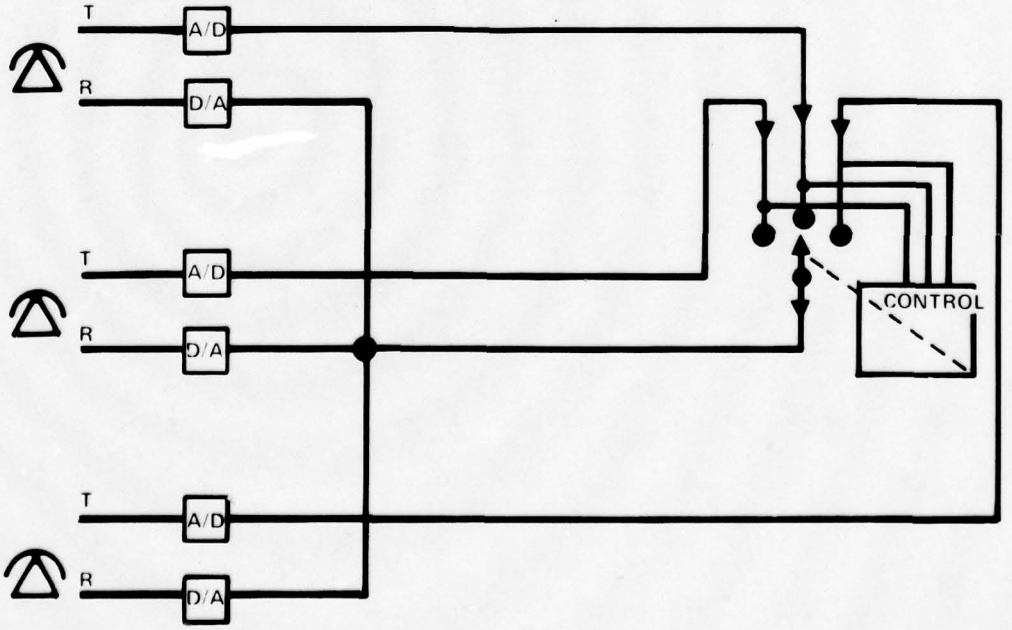


Fig.11 Digital conferencing unit

DISCUSSION

C.E.Tate, UK
When do you expect the TTC-39 Digital Switch to come into service?

Author's Reply
In 1982 or 1983.

LES ASPECTS TECHNIQUES ET OPERATIONNELS DES DES TELECOMMUNICATIONS EN AERONAUTIQUE

INGENIEUR EN CHEF CARLIER MAURICE
CENTRE D'ESSAIS EN VOL - B.P.N°2
9I-220 BRETIGNY SUR ORGE

Le 25 Juillet I909, Louis BLERIOT effectuait la première traversée aérienne de la Manche à bord du monoplan BLERIOT type XI; en Mai 1919 le Commandant READ, aux commandes d'un hydravion CURTISS, reliait Terre-Neuve à Lisbonne après une escale aux Açores, cet hydravion était équipé d'un radiogoniomètre permettant des relèvements jusqu'à 1000 km et d'un poste émetteur-récepteur radio permettant des liaisons jusqu'à 500 km.

Le 20 Mai 1927, Charles LINDBERGH, avec son " Spirit Of St Louis", réussissait la première traversée aérienne de l'Atlantique Nord, New-York - Paris sans escale. Le 2I Juillet I969, l'astronaute ARMSTRONG était le premier homme à poser le pied sur la lune.

Les exploits de BLERIOT et LINDBERGHfurent le résultat des extraordinaires qualités de ces deux pionniers, les premiers pas d'ARMSTRONG, sans minimiser les qualités de l'astronaute, furent le résultat du travail de nombreuses équipes.

A la solitude des premiers aviateurs, dont on était sans nouvelles pendant de longues heures, il faut opposer la transmission vidéo en direct de l'aventure spatiale. En un demi-siècle l'aéronautique a connu un développement étonnant, les télécommunications, pour leur part, ont aussi fait d'énormes progrès. En ce début de Symposium, consacré aux techniques numériques dans les télécommunications en aéronautique, je voudrais d'abord rappeler les principales exigences opérationnelles et les solutions retenues en particulier depuis la fin de la seconde guerre mondiale, ensuite, après avoir dégagé certaines contraintes propres aux télécommunications en aéronautique, je voudrais présenter les tendances qui apparaissent pour l'application des techniques numériques.

x

x x

Sans vouloir prétendre à un examen complet de tous les besoins en télécommunications dans un scénario opérationnel, un avion ou plus exactement l'équipage d'un avion, pour assurer sa mission, doit disposer d'un nombre important d'informations :

- Informations internes (paramètres de vol, consommation de carburants.....)
- Informations externes (position relative par rapport à un point de référence, par rapport à d'autres avions, route à suivre, niveau de vol.....)

La fonction communication permet de disposer d'une grande partie des informations externes, elle est utilisée tout au long de la mission, du décollage à l'atterrissage, pour satisfaire aux règles de la circulation aérienne (autorisation de décollage, ordres de cap et niveau à prendre pour rejoindre une route aérienne, indication de changement de zône de contrôle (FIR) (1) .....).

Elle est le lien indispensable entre la Station de Défense Aérienne et l'avion d'interception; la connaissance de la situation aérienne, autant par la capacité de détection de la station que par la corrélation des plans de vols des avions en circulation aérienne générale permet au contrôleur d'opérations de définir l'objectif (hostile) et de donner les ordres appropriés de guidage au chasseur. Ces ordres tiennent compte, bien entendu des performances de vol du chasseur et de l'hostile, des caractéristiques du radar du chasseur et de l'armement disponible au moment de la phase finale de l'interception. La fonction communication apparait également nécessaire pour les avions d'appui qui doivent entrer en liaison avec les postes de commandement avancés et même avec un simple fantassin qui assure la désignation d'un objectif avec un illuminateur laser. A l'exception de quelques liaisons particulières, transmission de données, transmission d'images télévision prises par avion de surveillance, la fonction communication, en aéronautique, fait appel à la téléphonie A3. Après avoir utilisé toute la gamme VHF (100 à 156 MHZ) avec des canaux espacés à 180 KHZ puis à 100 KHZ, cette gamme VHF, pour les communications, a été restreinte de 118 à 136 MHZ (2) avec des canaux espacés à 50 KHZ puis 25 KHZ pour les besoins de la circulation aérienne générale et en particulier de l'aviation commerciale. La gamme UHF de 225 à 400 MHZ est, depuis vingt-cinq ans environ, réservée aux communications pour l'aéronautique militaire, l'espacement des canaux est passé progressivement de 100 KHZ à 50 KHZ puis 25 KHZ, ce qui permet de disposer de 7000 canaux.

(1) FLIGHT INFORMATION REPORT

.../...

(2) En même temps la gramme VHF, de 108 à 118 MHZ, était réservée au VOR et à l'ILS.

Les applications de transmissions de données en aéronautique sont actuellement peu nombreuses, elles permettent par exemple d'envoyer ou de téléafficher des ordres à un avion d'interception à partir d'une station sol, ces liaisons peuvent être réalisées en modulation FI et nécessiter un encombrement spectral plus important que la radiotéléphonie, elles peuvent être uniquement dans le sens sol-air ou sol-air et air-sol, dans le sens air-sol il s'agit le plus souvent d'un accusé de réception des ordres reçus.

La principale lacune des liaisons vocales est la faible protection vis à vis du brouillage à la fois :

- sur le contenu du message

- sur la sûreté de la transmission.

Il est bien clair que des ordres donnés en temps réel ont une valeur " opérationnelle " qui diminue très vite, toutefois la connaissances des procédures et des types de messages permet à l'ennemi de préparer des actions d'intrusion radioélectrique.

Par ailleurs, les avions ne peuvent mettre en oeuvre des émetteurs de grande puissance et, lors des opérations au-delà de la ligne du front, le rapport puissance d'un brouilleur/puissance de l'émetteur ami est défavorable pour le bilan de liaison.

Les liaisons de transmissions de données citées plus haut se prêtent à une protection du contenu du message et le bilan de liaison peut être amélioré par l'emploi d'antennes directives à grand gain.

Pour connaître sa position par rapport à un point de référence et donc assurer sa navigation, l'avion peut utiliser un système autonome ou disposer d'informations données par des systèmes coopératifs (pseudo-communication). Ces systèmes coopératifs sont nombreux : VOR, DME, OMEGA, pour un avion militaire tactique on utilise le TACAN. Le TACAN permet à un mobile d'interroger une balise et de connaître sa position (distance p et relèvement θ) par rapport à cette balise. Si la balise est aéroportée on a une fonction air-air. Le TACAN fonctionne entre 962 et 1213 MHZ et permet d'utiliser 256 canaux (123 canaux X et 123 canaux Y).

La résistance au brouillage du TACAN est difficile à améliorer en raison de sa compatibilité avec les DME de l'aviation civile. On ne peut empêcher l'ennemi équipé de matériels de bord TACAN d'utiliser les balises d'infrastructure dont la position est bien connue; de même, en cas de conflit, pourrait-on l'empêcher d'implanter des balises dans sa zône afin de provoquer des erreurs de navigation ?

Pour identifier les avions, l'utilisation du radar secondaire dans la bande L ou IFF (Identification Foe or Friend) est l'un des moyens importants dont disposent les stations sol de Défense Aérienne et les centres de contrôle de la circulation aérienne civile. On peut dire s'il s'agit d'un système de communications particulier, en effet, l'avion interrogé décode l'interrogation et la transmet après un nouveau codage. Dans ce transcodage l'information d'altitude de l'avion peut être introduite, ce qui permet à la station interrogatrice de disposer des trois coordonnées (p, θ, Z) des avions amis présentés habituellement sur les indicateurs panoramiques, avec indication de l'altitude dans une étiquette voisine de l'écho-radar. Il suffit de noter que les avions ennemis n'envoient pas de réponse, ce qui permet de les distinguer des amis. Cette fonction identification est également indispensable pour la station sol de missiles sol-air, de même que pour l'avion d'appui sol qui doit pouvoir identifier ses objectifs (chars,.....), ce problème de surveillance du champ de bataille conduit à multiplier le nombre d'interrogateurs IFF et complique le travail de séparation des multiples réponses reçues par chaque interrogateur.

La fonction anticollision est une préoccupation importante des Centres de Contrôle de la Circulation Aérienne, elle peut être obtenue par le respect de procédures de séparation en temps et en niveaux de vol comme c'est le cas pour la circulation aérienne générale, elle peut être assurée par les contrôleurs d'opérations pour la circulation opérationnelle militaire et la circulation d'essais réception dans les stations militaires.

x

x x

Au cours des années qui ont suivi la seconde guerre mondiale les matériels de communications, de navigation et d'identification ont été utilisés au sol et à bord des aéronefs. Au fur et à mesure des années, en raison de l'augmentation des performances des avions et de la densité du trafic, les nouvelles exigences opérationnelles ont été satisfaites, soit par des perfectionnements apportés aux matériels existants en profitant des progrès technologiques incessants, soit par adjonctions de matériels dont l'emploi s'imposait à la communauté aéronautique comme le VOR, le DME ou le TACAN.

.../...

Outre les progrès technologiques qui ont permis, soit d'améliorer la fiabilité des matériels supportant des contraintes d'environnement sévères (on peut citer l'importance de la technologie à l'état solide), soit de présenter des performances accrues (stabilité et précision de fréquence, amélioration de la sensibilité des récepteurs, accès simple des canaux.....), des tentatives ont été faites pour éviter certaines duplications de fonctions ce qui a donné naissance à des matériels tels que :

- émetteurs-récepteurs V/UHF assurant la fonction communication en gamme VHF ou UHF à la demande du pilote
- équipement VOR/ILS regroupant les récepteurs VOR et ILS.

A bord des aéronefs l'exploitation des informations internes (paramètres de vol....) et des informations fournies par les différents capteurs (centrale à inertie, radar de navigation Doppler, récepteur OMEGA,.....) a fait des progrès considérables. La charge de travail de l'équipage ne pouvant croître indéfiniment et pour optimiser l'utilisation de toutes les informations disponibles, des développements importants ont été effectués pour présenter, suivant les différentes phases de vol, les éléments essentiels pour ces phases sur des viseurs " tête haute ", ou des tubes à rayons cathodiques en couleurs, en présentation " tête basse ". Les techniques numériques ont favorisé ces développements, les équipements de bord présentent de plus en plus des informations de sortie numérisées que l'on envoie sur une barre-bus, la gestion des échanges d'informations étant effectuée par un calculateur spécialisé. Grâce à ces techniques de dialogue entre les différents capteurs, on améliore la précision des sous-systèmes, comme celui du sous-système de navigation par le couplage inertie - Dopler par exemple. Les systèmes d'armes complexes ne peuvent plus se concevoir sans faire appel à la technique de la barre-bus avec comme support de transmissions la ligne bifilaire blindée, la ligne coaxiale ou, dans l'avenir, les fibres optiques.

Les réseaux de télécommunications au sol, civils ou militaires, font largement appel aux techniques numériques. Alors que ces techniques se sont introduites rapidement à bord d'avion, dans le " réseau de communication entre tous les capteurs ", il peut paraître surprenant que le domaine des télécommunications sol-air n'y fasse pas très largement appel. Certaines particularités des réseaux radio air-sol, particularités que l'on verra plus loin, le souhait d'aller plus loin que le regroupement de fonctions, comme dans les émetteurs-récepteurs V/UHF, conduit différents pays à envisager des systèmes dont les objectifs ambitieux entraînent des délais d'étude et de développement considérables.

Un réseau radio sol-air forme, sur chaque fréquence de travail, et dans la limite de portée optique de chaque participant du réseau, une communauté particulière où le message d'un participant est entendu par tous les autres. Lorsqu'un avion appelle un Centre de Contrôle Aérien sur la fréquence de travail allouée, tous les avions qui utilisent cette fréquence entendent le dialogue entre l'avion et le Centre de Contrôle, ceci est d'ailleurs une source d'informations mutuelles.

Cependant, l'accès au réseau, ou plus exactement l'utilisation de la fréquence par un avion, ne peut se faire que pendant le silence des autres participants, par expérience, l'écoute de certaines fréquences de travail montre qu'il apparait des problèmes de saturation du réseau. Pour éviter cette saturation, en cas d'urgence, les matériels UHF en service ont une fréquence de garde.

Pour changer des paramètres dans des réseaux de télécommunications sol-air, il faut assurer la compatibilité des différents participants. Si une armée de l'Air décide d'équiper un nouvel avion d'armes d'un radar de tir nouveau, les performances de ce nouvel avion seront liées à cet équipement, mais cela n'empêchera pas l'utilisation des avions d'armes déjà en service qui conservent leurs performances propres. Si cette même Armée de l'Air décide d'équiper ce nouvel avion d'armes avec un UHF à modulation numérique, il faudra, pour la mise en service, avoir réalisé des installations au sol à modulation numérique; les avions déjà en service ne pourront, sans nouvel équipement ou modification de l'équipement ancien, entrer dans ce nouveau réseau UHF numérisé. Ce problème de compatibilité peut conduire à un double équipement des avions (matériel VHF et UHF des avions militaires appelés à travailler avec la circulation aérienne générale et la circulation opérationnelle militaire). Il faut d'ailleurs penser que les réseaux en place durent plus longtemps qu'on ne le prévoit à l'origine, pour plusieurs raisons et en particuliers à cause :

- du coût de leur remplacement
- des perfectionnements que l'on peut apporter aux matériels anciens pour améliorer leurs performances.

Quelles sont les techniques numériques appliquées largement aux réseaux de télécommunications sol, qu'il s'agisse de réseau par câble, par faisceaux hertziens ou de réseau radio et qui pourraient être utilisées pour les réseaux sol-air ? En tout premier lieu les développements et réalisations dans le domaine de la numérisation de la voix permettent d'utiliser :

- les techniques PTT de MIC à 64 Kbts
- les techniques de modulation delta à 19,2 Kbts
- les techniques de vocoder (environ IO Kbts).

.../...

Par ailleurs les techniques de commutation de paquets, de commutation temporelle, offrent aux ingénieurs des possibilités et des exemples d'application qu'on ne peut négliger. Cette diversité des techniques disponibles pose dans les réseaux sol le problème de l'intéropérabilité (analogique - numérique, MIC - Delta, fréquentiel - temporel) et incite à une large concertation entre les partenaires du traité de l'Atlantique Nord pour développer et mettre en oeuvre un nouveau système de communication qui assure en même temps les fonctions de navigation, d'identification et si possible d'anticollision (concept Communication, Navigation, Identification).

Ce système temps - fréquence à accès multiple présente les caractéristiques générales suivantes :

- informations numérisées dans des messages de courte durée
- utilisation d'une large bande de fréquence
- synchronisation de tous les participants du réseau
- accès de chaque participant pendant une durée de temps déterminée, à intervalle régulier ou période du réseau.

On peut considérer que chaque utilisateur, au lieu d'utiliser un canal (une fréquence radio) dispose d'un canal temporel pendant lequel il dialogue avec le réseau. En dehors de cette période particulière, l'utilisateur peut utiliser toutes les informations échangées par les autres participants du réseau, quitte à limiter cette utilisation par un filtrage particulier.

L'échange de messages entre des participants ayant la même référence de temps permet de déterminer la distance entre les participants; en outre, si l'on dispose au sol d'un quadrillage de stations, tout aéronef peut se localiser avec les distances par rapport à trois stations.

Connaissant sa position par rapport à un réseau de référence chaque aéronef peut, dans son dialogue avec le réseau, donner les informations relatives à sa position, son altitude, sa vitesse, son cap....., cet envoi d'informations vers un Centre de Contrôle radar permet d'établir des corrélations nécessaires avec les échos radar et assurer ainsi la fonction identification.

Les exposés sur les systèmes qui seront présentés au cours du Symposium donneront des informations bien plus complètes sur les fonctions assurées, la protection contre le brouillage et le chiffrage en ligne étant avec l'intéropérabilité les principales préoccupations des concepteurs.

x

x x

Nous avons vu que, depuis la fin de la seconde guerre mondiale, les matériels de communications, au sens large du terme, se sont développés :

- pour assurer un contrôle de plus en plus rigoureux de l'espace aérien et en particulier des routes aériennes et des zônes terminales pour la circulation aérienne générale.
- pour permettre aux stations de Défense Aérienne de donner des ordres de guidage aux avions d'interception et aux Postes de Commandement de diriger les avions d'appui au sol.

Les progrès technologiques ont été mis à profit pour ces développements, sans toutefois faire appel aussi largement aux techniques numériques que dans les réseaux sol (modulation numérique et commutation électronique). Le concept C.N.I. donne lieu à des études dans différents pays de l'O.T.A.N., les difficultés à résoudre tant au plan technique qu'au plan de l'intéropérabilité, l'importance de la novation apportée par des réseaux de ce type conduiront à des retards de mise en oeuvre par rapport aux délais souhaités des concepteurs. Il ne me parait pas souhaitable d'attendre le développement complet des réseaux du type C.N.I., sans essayer d'utiliser, dès maintenant, et pour des applications partielles, les ressources des techniques numériques. Les développements limités qui apparaissent apportent des solutions à certains problèmes opérationnels, ils ne remettent pas en cause complètement les concepts actuels et permettent d'attendre, outre les délais de développement, d'expérimentation et de mise en oeuvre, les délais importants liés à la nécessaire concertation entre les membres de l'alliance.

A NOVEL APPROACH TO THE DESIGN OF AN ALL DIGITAL AERONAUTICAL SATELLITE COMMUNICATION SYSTEM

M. E. Ulug
Department of Systems Engineering and Computing Science
Carleton University
Ottawa, Canada

ABSTRACT

This paper describes a novel approach to the design of an all digital aeronautical satellite communication system. The basic system design is based on the transparent intelligent network, for short, TI-NET, principles (ULUG, M.E.). The novel features of the system, namely a 9.6 kb/s ground to aircraft link which provides a TDMA operation and the statistical multiplexing of the encoded voice and data have been experimentally tried out using 12-14 GHz satellite link between Carleton University, in Ottawa, Canada and the NASA AMES Research Centre in Palo Alto, California using Hermes (CTS) satellite (ULUG, M.E., WEIR, D.F., MORRIS, L.R., GRUBER, J.G., 1976), (ULUG, M.E., GRUBER, J.G., 1977). Other unique features of the system are the trading off of the satellite up-down delay with the packet formation delay, multi-polling and multi-addressing capability, and the complete transparency to the user's protocol. In addition the system has a turn around time of 22.5 ms. which is most useful in making quick changes in the polling sequence as well as producing fast re-transmissions. The paper also describes a polling algorithm which meets the particular needs of the aircraft in high, medium and low density areas. The terrestrial network connecting the communication centres and earth stations in North America has also been discussed and its associated interfaces and protocols have been defined.

It is believed that the proposed system will result in a more efficient use of the communication channels, greater immunity against noise, and a less complex airborn computer with smaller memory.

Although the system design is based on a set of hypothetical traffic data, the model can be readily modified to perform at a higher or lower traffic level. In this connection another system using a 4.8 kb/s ground-aircraft link has been described and its performance compared with that of the proposed model.

I INTRODUCTION

At the start of this section it should be made quite clear that the proposed system model described in this paper and design philosophy behind it are based on the author's own convictions and are not related in any way to the views of any aeronautical authority.

It should also be made quite clear that the proposed model concerns itself with the use of geosynchronous satellite relays as a means of providing digital communications to the aviation sector. The model is not restricted to the use of one or more satellites and does not concern itself with whether dependent or independent surveillance is the best approach, although both of these have been briefly mentioned in the paper as the possible user requirements.

The author firmly believes that a successful communication system should have marketable features and provide services which can be justified economically to the aviation industry. Moreover such a communication system should be designed from top down starting with satellites and developed from bottom up. If the system design is simply a set of constraints imposed on the specialists developing components, then it is extremely difficult to achieve the optimum solution, although the specialists will undoubtedly meet the challenges presented to them.

This paper describes an all digital communication system assuming that the satellites do not attempt to cover very large areas and can support a 9.6 kb/s ground to aircraft channel with adequate down link power, i.e. $E_b/N_o \approx 4.5$ db corresponding to a 10^{-5} bit error probability at the convolution decoder. Based on this assumption the system uses the available bandwidth very efficiently by statistical multiplexing of encoded voice and data, improves the noise immunity by transporting data in minipackets using TDMA mode of operation, provides multi-polling, multi-addressing capability together with a very fast system turn around time of 22.5 ms.

The critical parts of the system have been tried out experimentally under similar but not exactly the same conditions. The results clearly indicate the feasibility of the proposed system.

II USER REQUIREMENTS

At this time the user requirements are not too well known. Because of this the following types of traffic each with its own set of priorities have been assumed. These assumptions are based on the future projections which involves a number of terminals using the airborn computer as a cluster controller. These terminals may include CRT's, displays, digital facsimiles, printers, vocoders, and different types of inquiry/response terminals.

I. Real Time Poll/Address:

Number of aircraft: 360
Polling cycle in high density areas: 1 minute
Polling interval: 700 ms.
Maximum ground-aircraft (G-A) packet length: 85 ch's
Maximum aircraft-ground (A-G) packet length: 54 ch's

The Poll/Address traffic is used for the following purposes:

a) Aircraft Surveillance:

Using G-A messages having a mean length of 8 ch's in order to

i) conduct dependent surveillance, i.e. to obtain from each aircraft its position as determined by its self-contained inertial navigation system.
ii) conduct independent surveillance, i.e. to transmit multi-tone ranging signals to the aircraft and receive them back via two satellites together with information such as altitude, air temperature etc. so that by measuring the phase shifts and knowing the other details such as those previously mentioned the aircraft's position can be determined within the tolerance of approximately one mile.

b) Inquiry/Response:

In response to an inquiry from the aircraft to address one of the terminals through the airborn computer with messages having a mean length of 35 ch's.

c) Supervisory/Management:

To provide a means of communicating with the aircraft to manage their flight plans and/or to supervise their usage of various data and voice channels using messages having an average length of 20 ch's.

2. Non-Real Time Bulk Data

To transmit to and receive from the aircraft bulk data using packets having a mean length of 80 ch's and maximum length 135 ch's.

Average number of G-A packets per aircraft per hour: 22
Average number of A-G packets per aircraft per hour: 14

3. Encoded Voice

An undetermined number of voice channels are required. The useage of the voice channels at the present time is 27.9×10^{-3} Erlangs (calls seconds/second) per aircraft with typical call duration of 45 seconds. This level of usage is expected to drop considerably depending on how efficiently data channels are used in the proposed system.

Average voice burst: 1.1 seconds
Average idle period: 2 seconds
Number of voice bursts per average call (45 seconds): 14.6
Average number of encoded voice called per aircraft per hour: 0.22 i.e. on the average 80 aircraft per hour

III TRANSPARENT INTELLIGENT NETWORK APPROACH

Transparent Intelligent Network, for short TI-NET, was designed (ULUG, M.E.) and developed (ULUG, M.E., WEIR, D.F., MORRIS, L.R., GRUBER, J.G., 1976), (ULUG, M.E., GRUBER, J.G., 1977) at Carleton University, in Ottawa, to permit network access with minimum standardization, and provide maximum transparency in both time and protocol. Ideally then TI-NET accepts only a bit stream consisting of users' protocol and data and transports it wherever it has to go using all of its intelligence.

3.1 Transparency in Protocol

Transparency in protocol is extremely important since it determines how easily a user can interface to a communication system. Manufacturers have traditionally provided the capability to communicate via leased or switched lines with terminals using protocols such as BSC or SDLC. Terminal communication and handling is based on a communication facility which is transparent in protocol.

In the case of aeronautical communication systems which are not transparent the airborn computer is required to emulate the user's host computer. This results in a more complex and costly computer when using a number of different types of terminals in the aircraft. Therefore it is very desirable to be able to communicate from the ground with these terminals using their own protocol. In TI-NET the interfacing with the user's host computer and the terminals simplifies to EIA RS 232-C standard.

3.2 Transparency in Time

A communication system that is transparent in time is subject to a delay equal to the propagation time in transmitting data from a users' host computer to its terminal. A leased telephone line for example is transparent in time. Circuit switched systems are also transparent after an initial delay for call set up. However, a communication system using packet switching can be very non-transparent in time because of packet formation or entrance link delay as well as processing, queueing, intermediate link and intermediate node delays. From these the packet formation delay, i.e. the time required for sufficient user's data to form a packet or time required to receive and error check user's packet at the entrance to the communication system, is by far the largest one, particularly if the user is transmitting over a low speed line.

In TI-NET although the user's data enters the system at low speed it is transported towards its destination in mini-packets (mp's) at regular time intervals (e.g. 10 ms.). This is in effect equivalent to entering the system over a high speed line. Hence both the packet formation and the storage requirements are minimized at the entrance. To provide sequencing, error detection/correction and routing of mp's multi-user packets (MP's) are used. MP's are generated synchronously by all the nodes at fixed

time intervals. This is called rythmic operation.

3.3 Reduction in Satellite Up-Down Delay

In aeronautical satellite communication systems the use of C band between ground and satellite and L band between satellite and aircraft has been proposed. In general it is very difficult to mount a high gain L band antenna on the aircraft and to keep it directed at two satellites during flight. Typically the gain of such antenna varies between 9 to 4 db. In addition there are strong reflections from the ocean surface. Because of these factors the satellite down link has a very poor signal to noise ratio.

As a result every effort is made to keep the baud rate and hence the bandwidth as low as possible in the system design. When using conventional packet switching techniques the entrance link delay becomes quite high at low baud rates unless the communication centre and the earth station are at the same building and connected to each other with a high speed link such as the 100 kb/s line between a host computer and an IMP (ORNSTEIN, S.M., HEART, F.E., CRAWTHER, W.R., RISING, H.K., RUSSELL, S.B., MITCHELL, A., 1972) in Arpanet. For example to receive over a 1200 baud line a packet 37.5 CH's long takes 250 ms which is approximately equal to the satellite up-down delay. It is clear therefore that if the TI-NET approach is adopted message transmittion delay will be reduced from

$$D_{M.T.} = 0.250 + \frac{l_p}{C}$$

where l_p = packet length in bits

C = baud rate in bits/sec.

to $D_{M.T.} = 0.250 + T_s$

T_s = average service time of $\frac{22.5}{2}$ ms.

As mentioned above at 1.2 kb/s the cross over point is at l_p = 37.5 CH's. After that as l_p increases $D_{M.T.}$ becomes as good or as better than the terrestrial propagation delay when using link control and conventional packets switching.

IV SYSTEM DESIGN

4.1 Architecture of Satellite Subnet

In the proposed model 360 aircraft are divided into 3 sets, i.e. A_1, A_2, A_3. Each set has its own polling cycle and each set is in turn divided into three subsets according to the location of the aircraft in the high, medium or low density areas. The aircraft in each set A_i communicate with the ground using a different channel C_i. Hence the first part of the subnet used for polling looks like three separate starnets using a common central node which is the ground earth station. The earth station will be equipped with two 34' dishes each directed to one of the satellites which are placed 22° apart. (WOODFORD, J.B., 1973). Moreover, for reliability purposes there may be more than one earth station on each side of the ocean, e.g. one in Canada and one in U.S.A. In that case these two stations are connected together using a terrestrial link. This will be further discussed in 4.9.

These 3 starnets operate in TDMA mode from ground to aircraft, G-A, and in FDMA mode A-G. The G-A transmission is at 9.6 kb/s and uses statistical multiplexing of encoded voice and data in MP's. The A-G transmission is at 1.2 kb/s over three FDMA channels using private (conventional) packets, PP's.

In addition, to meet the users' requirements four different A-G channels are provided. These are as follows:

1. Voice Channel:

This channel operates at 2.4 kb/s and uses statistical multiplexing. It is for the transmission of encoded voice (840 bd) and bulk data (1560 bd) in the PP's coming from the same aircraft with the former having the priority over the latter. The transmission of bulk data is not solicited by the ground. However, access to this channel to transmit encoded voice is controlled by the ground.

2. Retransmission Channel:

This channel operates at 1.2 kb/s and is used solely for the retransmission of PP's carrying FDMA messages which are incorrectly received by the ground. In other words polling channels are not used for retransmission. The access to this channel is controlled by the ground by the transmission of a negative acknowledgement (NAK) to the aircraft.

3. Bulk Data Channel:

This channel operates at 1.2 kb/s and is used for the transmission of non-real time bulk data. The access to this channel is controlled by the ground by either authorization or by the transmission of a NAK. The bulk data traffic is also transported in PP's.

4. Emergency Channel:

This channel operates at 1.2 kb/s and used only in case of an emergency. The random access method is chosen to transmit PP's. If there are two emergencies at the same time the channel number 3 will be closed to bulk data and assigned to emergency as long as it is required.

As can be seen from the above, A-G channel capacities add up to 9.6 kb/s. Therefore, a 9.6 kb/s Full Duplex (F.D.) digital transmission link is required between the communication centre and the earth station. All satellite channels are rate 1/2 convolution encoded and Viterbi soft descision

decoded in the earth station as well as in the aircraft.

4.2 Two Types of Protocol

In the proposed model two types of protocol are used, namely inner and outer core protocols. The outer core protocol is basically a modified form of the protocol used in TI-NET. It handles MP's and mp's in order to create a G-A TDMA operation. It also controls the access to all return channels except the emergency channel and acknowledges the PP's by using explicit NAK and implicit ACK technique, as well as carrying out numbers of other supervisory functions. In other words it is the extension of the communication centre into the space. The design of the outer core protocol is described in this paper in detail since it is an integral part of the communication system.

The design of the inner core protocol, on the other hand, is left to the user since it concerns the actual messages transported by mp's G-A and by PP's A-G. The system is transparent to the inner core protocol. There is one area of overlap, however, between the inner and outer core protocols. This area is the acknowledgement of MP's going G-A by PP's returning A-G, and is specified by the outer core protocol as follows: After the usual HDLC flags 10 bits are inserted in all returning PP's except the ones on the emergency channel. The first 2 bits are S, the HDLC supervisory commands/responses (ISO/TC 97/SC 6 (Tokyo-17) 1974), and the remaining 8 bits are N(R), i.e. received MP number. The supervisory commands are interpreted by the system as follows:

S = 00 acknowledges up to N(R) - 1 (Receive Ready)

S = 01 rejects starting with N(R) and acknowledges up to N(R) - 1 (Reject).

S = 10 acknowledges up to N(R) - 1 and informs temporary inability to accept any more (Receive Not Ready)

S = 11 rejects N(R) and acknowledges up to N(R) - 1 (Selective Reject)

This acknowledgement scheme is carried out for all MP's by all aircraft regardless which set A_i they are in. The overhead of the outer core protocol is either 44.44% or 30.55% depending on whether 16 or 20 byte data section is used with 26 ch. long MP. As shown later, on the average 71% of the time only three byte header is required. This results in an average overhead of 34.58%.

4.3 Use of G-A Channel

In this 9.6 kb/s channel an MP is generated at every 22.5 ms. at the communication centre synchronized to the master clock of the digital transmission facility such as Data Route (HORTON, D.J., BOWIE, P.G., 1974). A MP carries 4 or 5 mp's one of which is always for encoded voice processor and/or bulk data terminal operating at 2.4 kb/s. The other 3 or 4 mp's are for the airborn computer and/or the synchronous terminals which are connected to it. All 360 aircraft receive, error check and examine every MP transmitted. To make the examination easier each MP is given a fixed packed length of 26 ch's before HDLC bit insertion. Maximum packet length after bit insertion is 27 ch's (22.5 ms.). When an MP contains few bit errors, it does not necessarily mean that all mp's are damaged. The mp's transport messages under the control of the inner core protocol. The messages are probably in PP's and have their own block check (BC). The only exception to this is the encoded voice messages or bursts. These do not go in PP's and do not have BC's. As the MP's arrive at the airborn computer every 22.5 ms., they go through a CRC16 shift register. The mp's are then sorted out and stored in the memory locations identifiable by MP numbers as they exit from their own CRC16 shift registers. Let us consider a G-A message having an average length of 35 ch's. It will take only 35x6.666 = 233.33 ms. to receive and check the CRC16 of this message. If the BC checks out alright the message will be used immediately. On the other hand if an MP is damaged it will be retransmitted as a result of either the earth station monitoring the L band or the earth station receiving a NAK from one of the FDMA channels. In the former the transmission will be received on the average $250 + 22.5 + \frac{22.5}{2} = 283.75$ later. In the latter it will be received on the average $116.66 + 300 + 26 \times 0.833 + 500 + 22.5 + \frac{22.5}{2} = 972$ ms. later. Here 116.66 ms. represent the average time between returning messages. As will be discussed later, an aircraft is polled approximately every 700 ms. (840 ch's) synchronized to the received 9.6 kb/s clock. Moreover, this polling is staggered among the three sets by 700/3 = 233.33 ms. (280 ch's) in order to obtain a more frequent updating about the status of MP's. The second item, 300 ms., represents the data acquisition time before the information can be transmitted in FDMA channels. In addition before the status of MP's can be learned, 26 bits at 1.2 kb/s must be received. These are flag, flag, S and N(S) bits (see 4.2). Here it has been assumed that there is only one damaged MP in the 233.33 ms. interval. As can be seen time required to check an average length G-A message for CRC16 is 233.33 ms. This is shorter than both the 283.75 ms. (error detection by the earth station) and 972 ms. (error detection by an aircraft). This example is illustrated below:

	MP No	MP-BC	mp Status	Message BC
	1	v	v	
Damaged MP	2	x	?	
	3	v	v	
	4	v	v	
	.	.	.	v
	.	.	.	
	.	.	.	
	10	v	v	
	11	v	v	

Now if PP's at 1.2 kb/s were used instead of MP's at 9.6 kb/s it would have still taken 233.33 ms. to receive and error check the 35 ch. message but if it were found to be damaged its re-transmission would have taken on the average (233.33 - 22.5) x 1.5 = 316.25 ms. longer to receive. Then the delays for the two types of error detection mentioned above would be 600 ms. and 1288.24 ms. respectively.

It is clear therefore that messages transmitted using MP's are less likely to be damaged because they are transported in small blocks called mp's. Moreover if a message is damaged, it is faster to correct it when using MP's. The distribution of errors in a MP's is discussed in Section V.

4.4 M.P. Structure

MP's are generated every 22.5 ms. which corresponds to 27 ch. interval at 9.6 kb/s. MP structure can be designed to give the system a great deal of operating flexibility using activity fields, destination tables, etc. as in the asynchronous MP's used in TI-NET system. The fixed length and completely byte oriented MP structure described here, is designed to produce greater noise immunity rather than flexibility. The 26 byte long MP is divided up into the Header (7 or 3 bytes), Data (16 or 20 bytes) and BC (2 bytes) sections before the HDLC bit insertion and HDLC flag (1 byte). Similarly the data section is divided up into 4 or 5 mp's on a fixed number of bytes basis, with the voice mp always being placed either before or after the BC section and the flag (see 4.5). This way it is quite easy to identify the mp's even though a MP may be damaged by the noise.

The header basically consists of 7 bytes. The first byte is the HDLC flag. The second byte is N(S) or the number of the MP. The first bit of N(S) is used to identify the source of the MP on one side of the ocean, e.g. 0 = Canada, 1 = U.S.A., leaving the remaining 7 bits for the Mod (128) MP number. The third byte is the Priority/Format (P/F) field which will be described later. Next to the P/F field there are 4 bytes taken up by two aircraft addresses. An aircraft address or identification is 13 bits (up to 8192) followed by 3 bits of trailer. The first two bits identify the set A_i or the channel C_i (1.2 kb/s) with which the aircraft is to respond to a poll, or indicate the use of the encoded voice/bulk data channel (2.4 kb/s) as follows:

$$00 = C_1,\ 01 = C_2,\ 10 = C_3,\ 11 = \text{encoded voice/bulk data}$$

The last bit is the extension bit, i.e. 1 = another address is following, 0 = this is the last address. Sometimes only one address is required. In this case the same address is repeated twice but this time the extension bits are used in reverse order, i.e. 0 first and 1 last. This is done to use up the four bytes allocated to the aircraft addresses and to retain rigid byte integrity of the MP. Hence the structure of the header is as follows:

7 byte Header

F	N(S)	P/F	1st ADR 1	2nd ADR 0
1	1	1	2	2 bytes

OR

F	N(S)	P/F	1st ADR 0	1st ADR 1

The data section is also partitioned into compartments, the first 9 bytes contain 3 mp's carrying real time Poll/Address (P/A) messages to the airborn computer. The next 7 bytes is the voice mp which carry either vocoder frames to a 2.4 kb/s speech processor or bulk data to the airborn computer (not necessarily in the same aircraft). The structure of the data section is as follows:

24 bits	24 bits	24 bits	56 bits
P/A mp	P/A mp	P/A mp	Voice/Bulk Data mp
3 bytes	3 bytes	3 bytes	7 bytes

The data section followed by two BC bytes containing CRC16 and a flag. The P/F field, together with the two aircraft addresses following it, executes the outer core protocol. The two addresses, however, are not always required. In this case the P/F field ends with 000 and the next four bytes are used to form a Bulk Data (BD) mp. Hence a 7 byte header is reduced to 3 and the 16 byte data section is increased to 20 keeping the total length of the MP constant at 26 bytes.

Header	Data (3 P/A + 1 Voice mp's	BC	F	MH
← 7 bytes →	← 16 bytes →	← 2 →	← 1 →	

Header	Data (1BD + 3 P/A + 1 Voice mp's)	BC	F	MH
← 3 bytes →	← 20 bytes →	← 2 →	← 1 →	

As mentioned earlier, MP's are generated every 22.5 ms. At 9.6 kb/s this corresponds to 27 bytes. In other words one byte is left for the HDLC bit insertion. One byte is 4.16% of the 24 bytes of information placed between two flags in a 26 ch. fixed length MP. If the packet length does not quite reach 27 bytes after the bit stuffing the remaining space will be filled by 1's, i.e. mark hold (MH) condition will be applied. The P/F field consists of an identification field (3 bits), service field (2 bits) and extension field (3 bits) as follows:

Identification Field:

000: a MP carrying 9 addresses for P/A mp's and 1 address for voice mp.

001: to be assigned

010: normal MP with three byte header

011: normal MP with seven byte header

100: re-transmitted MP with 3 byte header

101: re-transmitted MP with 7 byte header

110: emergency message to one aircraft using all 20 bytes of the data section and reassigning the encoded voice to their aircraft, i.e. MP will include a 13 byte P/A mp and a 7 byte voice mp.

111: a MP broadcasting to all aircraft the occurrence of an emergency condition and requestion immediate clearance of the encoded voice channel. The MP structure is the same as in 110.

Service Field (2 bits):

00 = Normal MP, 01 = Escape, 10 - to be assigned, 11 - loop back

Extension Field (3 bits):

000: the header is 3 bytes

001: cancel the following P/A address or addresses from the original sequence of P/A addresses

P/F EXT. 001	1st ADR. 1	2nd ADR. 0	DATA

or

P/F EXT. 001	1st ADR. 0	1st ADR. 1	DATA

010: Add the following P/A address or addresses at the end of the original sequence of P/A addresses, or specify an address so that in the next MP space reserved for 2 addresses in the header will be transferred to the data section to form a BD mp. This specific address will be repeated twice using the trailer 11 (V/BD). The header structure is the same as the second one given in 001.

011: Cancel the 1st address from the original sequence of addresses and insert in its place the 2nd address.

P/F EXT. 011	1st ADR 1	2nd ADR. 0	DATA
	cancel	add	

100: Authorization of the use of the Encoded Voice/Bulk Data (A-G) Channel operating at 2.4 kb/s by the following addressee. The use of this coding also indicates to which aircraft the voice mp is directed. If the voice mp is carrying bulk data the repeated address is complimented.

P/F EXT 100	1st ADR. 0	1st ADR. or 1st $\overline{\text{ADR}}$. 1	DATA

or request for the retransmission of the bulk data PP number N(R) in the A-G channel operating at 2.4 kb/s by the same addressee.

P/F EXT 100	1st ADR. 0	N(R) 0 N(R) 1	DATA

The 1st 7 bits of N(R) is MOD 128 received MP number the last bit is the extension bit. When the same N(R) is repeated twice the extension bits are used in reverse order, i.e. 0 first and 1 last.

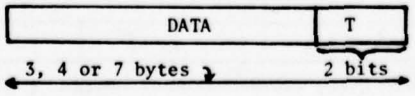
101: Authorization of the use of the Re-transmission channel operating at 1.2 kb/s by the following addressee or request for retransmission of the PP number N(R) in the same channel by the same addressee. The header structures are the same as those shown in 100.

110: The same as 100 except is to be used for the Bulk Data channel operating at 1.2 kb/s.

111: The service function as specified by the service field of the P/F field to be executed by 1 or 2 aircraft with the following address or addresses in specified FDMA channels operating at 1.2 kb/s. The header structures are the same as those shown in 001.

4.5 mp Structure

In this proposed model three types of mp's are used. These all have the same packet structure as shown below:



The 2 bit trailer (T) is coded as follows:

00: this data mp is not the last mp

01: this voice mp is not the last mp

10: this is the last mp of the message but it is not padded

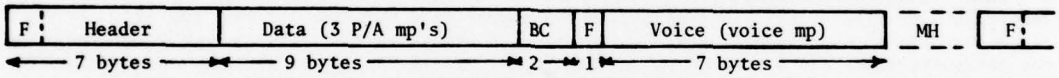
11: this is the last mp of the message and it is padded as per the following padding algorithm:

Padding Algorithm:

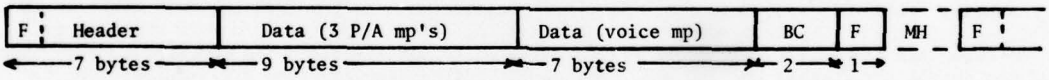
In order to force the transmission of the last mp it may be necessary to pad it with zeros preceeded by a one which flags the end of the data section. If there is only one bit to be padded a 1 is placed in this position.

Following are the three different types of mp's used:

Voice mp's: A vocoder, such as Vadac 5 (FULGHUM, D.P., 1974), operating at 2.4 kb/s requires 54 bits every 22.5. Since MP's are generated every 22.5 ms. a voice mp must have a data section which is 54 bits long. Then with a 2 bit trailer the voice mp is exactly 7 bytes long. In the case of Vadac 5 the choice of 22.5 ms. MP interval is very convenient because this particular vocoder uses 54 bit frames. Hence the first bit in every mp is the frame bit. The voice bursts carried by these mp's do not have BC sections and they are transparent in time. For example during an average voice burst of 1.1 seconds approximately 49 MP are transmitted. Some of these may be re-transmissions. But nevertheless they must all carry a different voice mp so that vocoder frames, which are not stored and error checked in the airborn computer, can be fed contiguously to the speech processor. The voice bursts vary from 0.25 to 7 seconds with mean burst length of 1.1 seconds. The channel is normally idle 60-65% of the time with an average idle period of approximately 2 seconds. During the idle periods non-real time bulk data can be transmitted from the same aircraft in PP's having a maximum packet length of 135 ch's (450 ms.) over the A-G voice channel operating at 2.4 kb/s. During the idle periods in the opposite direction voice mp's can carry bulk data to any aircraft in PP's with the same packet length restriction, i.e. 20 mp's (450 ms.). This is statistical multiplexing of encoded voice and data which is discussed in detail elsewhere (ULUG, M.E., GRUBER, J.G., 1977). When this 2.4 kb/s voice channel is not in use it can be entirely devoted to non-real time bulk data traffic with the understanding that the encoded voice has always priority over bulk data. The coding of the trailer, except in the last mp, identifies whether the voice mp is used for encoded voice (00) or data (01). As mentioned above voice mp carries voice bursts solely relying upon the forward error correction and bulk data using CRC16. Now if the voice mp is placed before the BC and the last flag its contents are also error checked and they can cause an unnecessary retransmission. This happens when the only error or cluster of errors happens to be in the 7 byte voice mp which is not retransmitted anyway. On the other hand while voice mp is carrying bulk data it must be placed before the BC. Therefore the following packet structure is proposed: When voice mp carrying voice burst with an average baud rate 840.



When voice mp is carrying bulk data with an average baud rate of 1560.



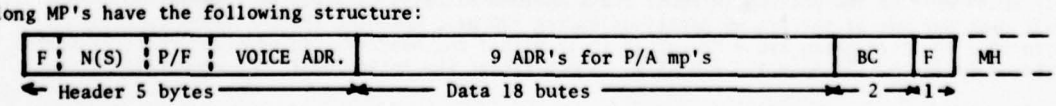
The CRC16 shift register operates between the start flag and the next flag, the former occurring every 22.5 ms. or 27 ch's at 9.6 kb/s. Since the operation is rhythmic this does not present any problems (see 4.7).

Poll/Address (P/A) mp's: These mp's are intended to carry real time polling and/or addressing messages to the aircraft. There are three P/A mp's in each MP and they are placed before the voice mp in the data section. The transmission of P/A messages are not transparent in time. The mp's are stored and error checked by the airborn computer before the message is used either by the computer itself or transmitted to a synchronous terminal say at 1.2 kb/s. The storage time may still become larger as a result of waiting for the re-transmissions of one or more MP's. Therefore unlike in the case of the voice mp, there is no relationship between the size of the data section and the operating speed of the terminal. In other words the airborn computer either uses the message itself or acts as a cluster controller for the terminals. For convenience the data section of the A/P mp is selected to be 22 bits. Then with 2 bits trailer the length of an A/P mp is exactly 3 bytes. The 01 trailer which indicates the voice transmission is not used with A/P mp's.

Bulk Data (BD) mp's: When the last four bytes of the header section reserved for two aircraft addresses are not required they become part of the data section. These four bytes are then used to form BD mp's to transport 30 bits of non-real time bulk data per MP to any aircraft in PP's having a maximum length of 135 ch's. Such a large PP requires the transmission of 36 mp's during 810 ms. It is more than likely that this transmission will be interrupted under the control of the extension field if and when the four byte space is required for the aircraft addresses. Like A/P and voice mp's carrying bulk data BD mp's are not transparent in time. They are stored and error checked by the airborn computer before processing. Also the 01 trailer is not used.

4.6 Destination Table

When G-A transmission starts two identical MP's are broadcasted to all the aircraft. These 26 byte long MP's have the following structure:



The header includes a two byte address for the voice mp and the data section contains a list of 9

addresses for P/A mp's. This list or sequence of addresses is called the destination table (DT). After the broadcast of these two identical MP's the transmission of normal MP's start. The first (left most) three entries in the DT determines the addresses of the three P/A mp's in the first MP. After this the DT is modified by the P/F field by adding and/or cancelling up to 2 addresses at a time. When an address is cancelled the remaining ones shift left by one. The operation of the destination table is as follows:

1. 9 addresses are partitioned into 3 groups each having 3 addresses as shown below

   $A_1^1, A_1^2, A_1^3, - A_2^1, A_2^2, A_2^3 - A_3^1, A_3^2, A_3^3$

2. When A_i^j's last mp is transmitted A_{i+1}^j is automatically assumed to take its place. Similarly A_{i+1}^j is replaced by A_{i+2}^j.

3. Subscript i represents one of the three FDMA channels through which the polled/addressed aircraft is expected to respond. It also corresponds to one of the 3 P/A mp positions in the MP.

4. When an address A_i^j is cancelled, the remaining addresses in A_i group shift left by one.

5. When a new address A_i^j is added to the A_i group it is placed at the last (right most) position.

6. It is also possible to place a new address in any position within the group A_i by using the cancelled routine.

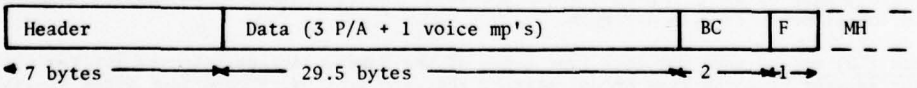
7. The minimum time interval in which a change can be made in the DT, e.g. to make a change in the polling sequence, is 22.5 ms. (a MP interval).

8. If it is necessary to have a time fill between A_i^j and A_i^{j+1} P/A mp's filled with alternate 0 and 1's followed up with T = 01 (voice) will be transmitted.

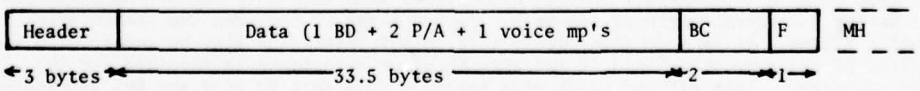
9. If there is no requirement for any transmission using the ith P/A mp position until the next polling message is due, and there is a long queue of addressing messages which are normally sent using the k'th P/A mp position, then an overflow is arranged by cancelling A_k^j and adding A_i^j in DT.

Now each A-G FDMA channel operating at 1.2 kb/s is used to transmit 54 ch's (360 ms.) every 700 ms. This is equivalent to a baud rate of 617. Retransmissions are done on a separate 1.2 kb/s channel. In the opposite direction, P/A mp's transporting 22 bits every 22.5 ms. are used. This is equivalent to 978 bd. However, the retransmissions are also done using these mp's. If it is assumed that on the average 25% of the channel capacity is used for this purpose the actual G-A P/A baud rate reduces to 733.5. Approximately 13.1% of this capacity is expected to be used for polling messages (96 bd), 56.5% for addressing messages (414.5 bd) and 30.4% for supervisory/management messages (223 bd). In other words during a polling interval of 700 ms.on the average 8ch.polling, 35 ch. addressing and 20 supervisory/management messages are transmitted using one P/A mp position to one aircraft. On the average then there is one address change per channel per 700 ms. Another two address declarations are required, one to convert the four bytes into a BD mp and the other for the statistical multiplexing of voice and bulk data on the voice channel per P/A channel per 700 ms. If the polling is equally staggered among the 3 P/A channels then on the average every 233.33 ms. there are three MP's with 7 byte headers. During 700 ms. approximately 31 MP's are transmitted and 9 of these have 7 ch. headers allowing 22 MP to have a BD mp (4 bytes). These assumptions result in a G-A BD channel having an average baud rate of 946.

Based on the traffic figures given in Section II this system model is design with an G-A TDMA link operating at 9.6 kb/s producing 26 ch. fixed length MP's every 22.5 ms. Using this link it is possible to transmit to an aircraft over any one of the three P/A mp positions a maximum of 85.5 ch's during a polling interval of 700 ms. The advantage of having a 9.6 ms G-A link is the ability to take rapid action to control the system, to provide fast re-transmissions, and to have adequate capacity for multi-terminal addressing. The disadvantage of it is the larger power requirement and the lower signal to noise ratio in comparison to a say 4.8 kb/s link. To make a valid comparison let us consider a model with a 4.8 kb/s G-A link. In this case a 39.5 ch. fixed length MP is generated every 67.5 ms. having the following packet structure:



or



leaving only 1 byte for bit insertion or 2.66% of 37.5 bytes of information placed between the flags.

The BD and P/A mp's are 4 and 3 bytes respectively, i.e. the sizes of these mp's and the header structure remain unchanged. However the voice mp is now 3 x 54 + 2 = 164 bits or 20.5 bytes. The MP interval of 67.5 ms. is selected so that the first bit of the mp is always the frame bit. This time there is only 10.37 MP's in the polling interval and a maximum of 28.51 ch's can be transmitted at 326 bd. to an aircraft over any one of the P/A mp positions during 700 ms. Moreover the G-A BD mp capacity is now reduced to 58.7 bd since there are 9 MP out of 10.37 with 7 ch. headers. In addition re-transmission time is increased by (67.5 - 22.5) x 1.5 = 67.5 m. Summarizing, the following is the comparison of the two links:

	9.6 kb/s		4.8 kb/s	
Capacity per P/A Channel	978	bd.	326	bd.
Max. No. of Ch's per 700 ms. per P/A channel	85.5	ch's	28.51	ch's
Max. No. of MP's per 700 ms. per P/A channel	31.11	MP's	10.37	MP's
Average capacity of G-A Bd Channel	946	bd.	58.7	bd.
Capacity of G-A Voice Channel	2400	bd.	2400	bd.
MP time interval	22.5	ms.	67.5	ms.
MP length	26	ch's	39.5	ch's
Re-transmission time (detected by earth station)	283.75	ms.	351.25	ms.
Re-transmission time (detected by an aircraft)	972	ms.	1039.5	ms.
Total mp capacity	6280	bd.	3436	bd.
Average overhead of the outer core protocol	34.58 %		28.41 %	

As can be seen reducing the link capacity by a factor of 2 resulted in a reduction of P/A mp capacity by a factor of 3 and in BD mp capacity by a factor of 16. There is no change in the voice mp capacity since it has to remain at 2.4 kb/s (840 bd. voice + 1560 bd. bulk data). As a result MP's at 4.8 kb/s are longer by a factor of 1.52 and the average overhead of the outer core protocol is reduced to 28.41% from 34.58%. Although at 4.8 kb/s the transmission has higher signal to noise ratio, the longer packets are more likely to result in re-transmissions. From overall system's point of view the biggest disadvantage of going down to 4.8 kb/s is the fact that the turn around time of the communication centre is increased by a factor of 3.

The model using 4.8 kb/s G-A link has been used for the comparison purposes only. In the remaining sections the original model using 9.6 kb/s link will be discussed.

4.7 Polling Algorithm

Let the total number of aircraft in each set A_i be N/m where m is the number of sets. Let each set A_i be divided into high, medium and low density subsets having N_x^i, N_y^i and N_z^i aircraft respectively. Let x, y, z be the polling cycles in seconds associated with these subsets. Then

$$\sum_{i=1}^{m} \left[\frac{N_x^i}{x} + \frac{N_y^i}{y} + \frac{N_z^i}{z} \right] \left[\bar{I}_{P/A} \cdot k_{P/A} \right]$$
$$+ N \cdot \left[\bar{I}_{BD} \cdot k_{BD} \cdot N_{BD} + \bar{I}_{VBD} \cdot k_{VBD} \cdot N_{VBD} + \bar{I}_v N_v \right] \cdot 3600^{-1} = C_{G-A}$$

where the subscripts s's

P/A = poll/address

BD = bulk data

VBD = bulk data in voice mp

V = voice

$\bar{I}_s$ = average message length or voice burst in bits

k_s = retransmission factor ≥ 1

N_s = average number of data messages or voice burst per aircraft per hour

C_{G-A} = actual G-A channel capacity available in bits/second

In the proposed model m = 3 and C_{G-A} = 6280 bd.

Also $\left[\bar{I}_{VBD} \cdot k_{VBD} \cdot N_{VBD} + l_v \cdot N_v \right] \Big/ 3600 = 2400$ bd.

Let us form a polling array having R rows and C columns for each set A_i. Let each row contain

$(N_x^i + \alpha i \,.\, N_y^i + \beta^i \,.\, N_z^i)$ aircraft. If T is the polling interval in seconds and $\alpha^i < 1$, $\beta^i < 1$ then

$$(N_x^i + \alpha^i N_y^i + \beta^i N_z^i)\ T = x$$

At least $(\frac{1}{\beta i})$ rows are required to include N_z^i. Hence $z = Rx = \frac{1}{\beta i}\ x$.

In R rows there are

$$\alpha^i \,.\, N_y^i \,.\, R/N_y^i = \alpha^i \,.\, R = \frac{\alpha^i}{i} \quad \text{sets of } N_y^i\text{'s}$$

Hence $y = \frac{x\ R}{\alpha i \beta i} = x\ \frac{1}{\alpha^i}$ and $c = \frac{x}{T}$

Let us apply this algorithm to the proposed model with the following assumptions:

$x = 60$ s., $y = 90$ s., $z = 120$ s., $N = 360$, $N_x^i = 30$, $N_y^i = 40$ and $N_z^i = 50$ for all i

Then

$T = 60/(30 + \alpha.40 + \beta.50)$ and

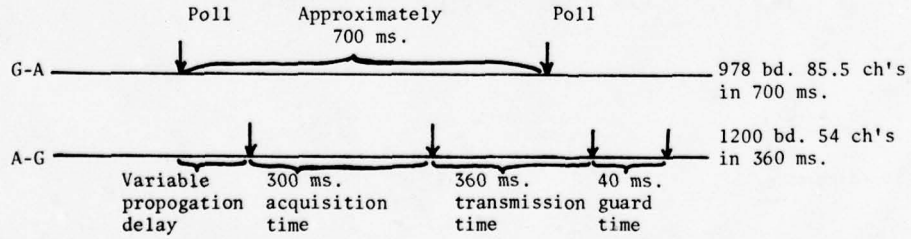
$\alpha = \frac{x}{y} = 0.66$, $\beta = \frac{x}{z} = 0.5$ Hence T = 737 ms.

Now as the values of N_x, N_y, N_z change the value of T is going to change. Let us assume a lower bound of 700 ms. for T and divide this interval into three sections as follows:

1. 300 ms. for data acquisition
2. 360 ms. for transmission (54 ch's at 1.2 kb/s)
3. 40 ms. for guard time

The 9.6 kb/s G-A link serving three starnets operates continuously and provides a received clock to each aircraft on MP (22.5 ms.), byte (0.833 ms.) and bit (0.104 ms.) basis. The relative phase of the clock with respect to the ground clock is a function of the aircraft's position. However, if the aircraft transmits in synchronism with the received clock on byte basis, its transmission will automatically be byte synchronized on the ground. In order to obtain message synchronization on the ground so that the transmissions from different aircraft are received at regular intervals such as 700 ms., the polling time will have to be adjusted to compensate for the propagation delay. This can be done only by estimating the present position of the aircraft based on its previous position. If the estimate is not very accurate it is possible in rare circumstances for two messages to overlap on the same A-G channel. In order to prevent this a guard time of 40 ms. is provided. Similarly the received frequency on the ground is a function of the aircraft's velocity vector, which can also be estimated by the ground computer. The acquisition time may be reduced if this information is supplied to a special device attached to the frequency tracking circuit of the receive modem so that the search can be done over a narrow range. The 300 ms. acquisition time also includes byte synchronization and the resolution of the 2 fold ambiguity of the rate 1/2 encoder. For this reason it may also be possible to improve the acquisition time further by applying to the receive modem a properly modulated signal from a special voltage controlled oscillator during the guard time provided no energy is detected on the input line connected to the antenna.

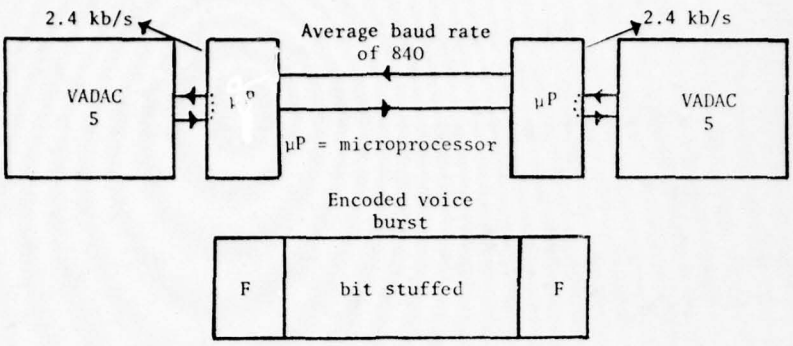
Using some or all of these techniques it may be possible to have a rhythmic operation in each direction based on the master (G) - slave (A) operation. The timing diagram will be as follows:



In order for the ground to receive messages coming from all aircraft at the regular intervals it is necessary to compensate for the propagation delays by either advancing or delaying the poll. This can be done by the communication centre where MP's are generated. Under the control of the inner core protocol start of the A-G transmission can be delayed by any amount by simply issuing a timing instruction to the air-born computer. This delay will be within the tolerance of ±1 bit or ±0.104 ms., i.e. the instruction "delay start of transmission by 144 bits" 15 ms. delay.

4.8 Formatting of Encoded Voice

Vocoders are frame synchronous devices which transmit continuously whether or not anyone is talking. In order to carry out statistical multiplexing of data and voice the active periods of the vocoder are identified by examining the spectrum amplitude information present in each frame. The active periods are then delimited with HDLC flags. The leading and trailing flags indicate the start and the end of a block of encoded voice, respectively. A hangover is provided at the end of each active frame sequence to insure that the speech message is ended. This protocol is implemented by µP's placed in front of the vocoders. These µP's also feed idle frames into the vocoders during the silent periods in order to keep them in synchronism.



It is imperative that the frame pulses in speech are correctly identified. In G-A transmissions 7 byte voice mp's are used so that frame pulse is always in the first bit position of the first byte. The last 2 bits are the trailer as discussed in 4.5. In A-G transmission active frames arrive at random times with respect to the locally generated idle frames. Meshing of these frames is required after proper identification of frame pulses which are not always located in the first bit position of a byte but rather in one of the 4 possible positions. Hence meshing on the ground is more difficult and requires more μP power.

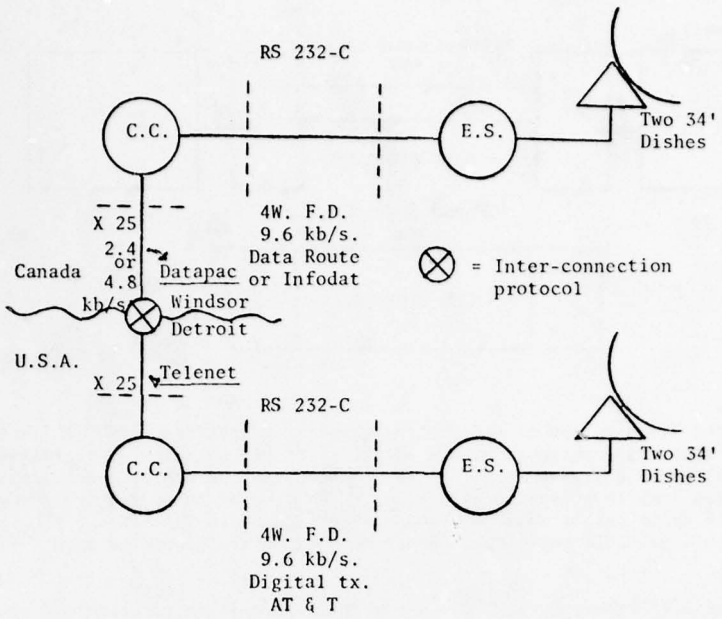
Vocoders like VADAC 5 (FULGHUM, D.P., 1974) are rather insensitive to noise and changes in clock speed. It has been shown experimentally that they operate in channels with error rate as high as 10^{-2} and with clock frequency reduced as much as 240 baud from the normal operating baud rate of 2400.

4.9 Terrestrial Network

In order to retain transparency in protocol, to eliminate packet formation delays and to reduce end to end delay variance leased digital lines are used between the communication centre (CC) and the satellite earth station (ES). In a digital transmission facility such as Data Route (HORTON, D.J., BOWIE, P.G., 1974) the error rate is typically in the range of 10^{-7} - 10^{-8}. The RS 232-C standard interface is used to connect to the digital data sets called "subscriber terminal equipment" or "STE's. They operate with conditioned di-phase at 2.4, 4.8 and 9.6 kb/s. The clock runs at 19.2 kb/s and is supplied either by the "line regenerative repeater", (if the distance is more than 6 miles of 22 GA or the loss is 30 db at 10 KHZ) or by "office loop repeater". The MP's are generated at CC synchronized to the Data Route clock and sent transparently through ES to all aircraft. The return channels which also add up to 9.6 kb/s are buffered for a short time to bring them into synchronization, bit multi-plexed (STDM) and connected into another "STE" through the RS232-C interface. Some of the data, such as the multi-tone ranging information, may be processed by ES computer on the fly and the results may be included in the PP's coming in on the FDMA channels under the control of the inner core protocol. In this way the rhythmic nature of the G-A and A-G transmissions is maintained. If CC and ES are a long way from each other, i.e. Toronto, Canada - Gander, Newfoundland a 9.6 kb/s line (part of a 56 kb/s Data Route full group) having a minimum number of intermediate nodes should be specified. This would minimize the digital transmission delays and the end-to-end message delay variance.

When a MP is received by ES it is transmitted immediately. However as the data is being received by ES it is put through a CRC16 shift register using the same 9.6 kb/s clock. If an error is detected in the MP after its transmission, a short NAK is sent to CC using the 1.2 kb/s emergency channel which is very rarely used. When DD receives the N(S) it immediately re-transmits the MP. In this way the re-transmission time is reduced considerably without jeopardizing the transparency of the system. As was mentioned before the error rate in Data Route is very low, and therefore these types of re-transmissions are expected to be very rare. On the long haul digital transmission links, however, the line problems can result in outages which usually last more than one second. In a case like this perhaps the best thing is for the other CC to take over all the polling functions. This can only be done if each CC transmits to the other its updated track file at the regular intervals over a reasonably high speed line. From this point of view the treatment of the long line outages is not different from the hardware or software failures that may occur in one of the CC's.

Two CC's can be joined together using, for example, Datapac in Canada and Telenet in U.S.A. If the required line capacity is 2.4 or 4.8 kb/s and the transmission time is not critical CC in Canada can be connected to a "network interface machine", "NIM", through an RS232-C interface. When NIM receives data equivalent to a carriage return it will form a packet using X25 (RYBCZYNSKI, A., WESSLER, B., DESPRES, R. and WEDLAKE, J., 1976) link and packet level protocols. Since both Datapac and Telenet are using X25 no interconnect protocol is necessary at the border. For other networks having incompatible protocols a new interconnect standard is currently being prepared. If the transmission time is a critical factor then a 9.6 kb/s leased line will be required between two CC's. In this case X25 protocol will have to be implemented by the CC computer or its front end. This is of course a considerably more expensive approach.



V EXPERIMENTAL RESULTS

As was discussed in section I the down link power of the satellite was assumed to be high enough to support a 9.6 kb/s ground to aircraft channel, i.e. E_b/N_o = 4.5 db. To investigate the behaviour of such a channel under marginal conditions, e.g. when E_b/N_o is as low as 2.8 db, the following experiments were conducted as part of the TI-NET system development. The results of these experiments are discussed in detail elsewhere (ULUG, M.E., GRUBER, J.G., 1977). The highlights of this work, which has a definite bearing on the proposed aeronautical model, will be briefly described.

A 9.6 kb/s link was set up from Carleton University's Wired City Laboratory (COLL, D.C., GEORGE, D.A., STRICKLAND, L.H., GUILD, P.D., PATERSON, S.A., 1975) in Ottawa, Canada over the Hermes (CTS) (EVANS, W.M., DAVIES, N.G., HAWERSAAT, W.H., 1976) satellite to NASA AMES Research Centre in Palo Alto, California, where it was looped back and returned to Carleton as shown in Figure 1. A full duplex audio channel was used for this purpose and a "CODEX 9600" modem was used to transmit data and voice from the TI-NET node over the 9.6 kb/s satellite link and return it to the same node. The audio channel encoder/decoder uses sampling and 8 bit linear quantizing at a video horizontal frequency rate of 15.75 kHz. The resulting digitized (CODEX) modem signal is then multiplexed into a video data stream during the horizontal sync. pulse time. Differential encoding/decoding is used to resolve the phase ambiguity in the QPSK modem's recovered carrier. The rate 1/2, constraint length 7, convoluation encoding with soft decision Viterbi decoding is used as the means of forward error correction. The QPSK modem accepts a composite digital signal of multiplexed audio and compressed video information, and outputs a modulated 70 MHz carrier into IF and RF equipment with 85 MHz bandwidth. Also the Carleton earth station has a 2.4 meter antenna (transmit and receive gain - 48.2 and 46.8 db.) with uplink transmit power limited to about 12 watts. The NASA AMES earth station has a 3 meter antenna (transmit and receive gains - 50.1 and 48.8 db.) with up link transmit power capability of up to 125 watts and higher. The experiments were conducted using 16.124 Mb/s differentially encoded QPSK. Since the measurements were made over a short period of time say 5-20 minutes it is assumed that the results will correspond to a channel with ideal additive Gaussian noise with one-side noise spectral density N_o.

Here

$$\frac{E_b}{N_o} = \frac{P_s/R_b}{P_n/B_n}$$

where P_s = signal power

R_b = bit rate

P_n = noise power

B_n = noise bandwidth

This equation may be written as

$$\frac{E_b}{N_o}\ (\text{db}) = \frac{P_s}{N_o}\ (\text{db}) - 10 \log R_b$$

It can be shown that (VITERBI, A.J., 1976) to achieve a bit error rate of 10^{-5} without coding,

E_b/N_o = 9.6 db is required. On the other hand while using rate 1/2 convolutional code, with constraint length 7, and Viterbi maximum likelihood decoding algorithms with soft decisions (3 bits including 2 quality bits), E_b/N_o need only be 4.5 db. This is a coding gain of 5.1 db. If the symbol energy-to-noise density is considered, E_b/N_o is reduced by 3 db since there is a bandwidth expansion of 2.

The following is an example of measurements made at 12-14 GHz using loop back (500 ms. space travel) from NASA AMES Research Centre via Hermes satellite.

Date:	21/4/77
Carleton Uplink Power:	12.4 watts
NASA AMES Uplink Power:	reduced to 40 watts to generate errors
Weather Conditions in Ottawa:	51% R.H., overcast, no percipiation, winds from SW at 22 km/hour
MP Interval:	20 ms. (24 ch's at 9.6 kb/s)
Length of Header + Data Field:	128 bits
Test Duration:	17.4 minutes
Number of MP's Transmitted:	52354
MP Error Rate:	1.6×10^{-3}
MP Error Free Interval:	Min.: 0.14 seconds, Max.: 70.5 seconds Mean: 12.3 seconds
Number MP's Analysed:	69
Bit Error Free Interval:	Min.: 0 bits, Max.: 9 bits, Mean: 1.46 bits
Distribution of Error Free Intervals (EFI):	37% 0 bit EFI, 25.5% 1 bit EFI, 11.5% 2 bits EFI, ..., 1.5% 5 bits EFI, 0.75% 6 bits EFI.
Number of Bit Errors in the Header and Data Field of an MP (128 bits):	Min.: 0, Max.: 9 bits, Mean: 2.81 bits

From these measurements the mean bit error rate can be calculated as follows:

$$\frac{2.81}{128} \times 1.6 \times 10^{-3} = 3.51 \times 10^{-5}$$

From the performance curves of the decoder this corresponds to E_b/N_o = 4.2 db. with differential coding.

At this level of error rate it was noticed that the damaged MP's were almost unformly distributed over the 17.4 minutes test duration. The clustering of bit errors within an MP is extremely fortuitous. It clearly indicates that if a 26 ch. MP is damaged the probability of having more than one mp being damaged is very small. In fact since mp's take 16-20 bytes out of a 27 byte MP interval it is quite possible that none of the mp's may be in error.

To increase the bit error rate further, NASA AMES uplink power was reduced to 30 watts on the same day. A total of 25750 MP's were transmitted over a period of 8.58 minutes. Under these conditions a MP error rate 3.9×10^{-2} was measured. The mean number of bit errors within a MP was 4.61 bits. From this the mean bit error rate was calculated as follows

$$\frac{4.61}{128} \times 3.9 \times 10^{-2} = 1.4 \times 10^{-3}$$

The corresponding E_b/N_o was obtained from the decoder curves as 2.8 db. In addition to the clustering of bit errors within a MP, this time, clustering of damaged MP's within the test duration of 8.58 minutes was observed. This was somewhat unexpected. However it clearly indicated that when E_b/N_o was reduced to a level as low as 2.8 db the communication system had to have enough excess capacity to perform a series of retransmissions to cope with the situation. The reason for the clustering of damaged MP's lie at least partly on the forward error correction scheme mentioned above, failing to cope with the long error bursts. It is not unusual for the channels subject to fluctuations due to fading and multipath to experience error bursts on the order of thousands of bits in length. In these cases the most effective preventative measure may be to scramble the data plus redundant bits after encoding but prior to transmission, and perform the inverse operation at the receiver prior to decoding (VITERBI, A.J., 1976). This technique has the effect of distributing the burst errors over a long period of time and making them appear to the decoder like random independent events. From the above experiments the following conclusion can be drawn:

i) E_b/N_o = 4.5 db will result in a mean bit error rate of 10^{-5}. This is desirable but not necessary because of the fact that bit errors tend to cluster in a MP, and therefore the probability of more than one mp being damaged is very small. For example when E_b/N_o is 4.2

db, the MP error rate is 1.6×10^{-3}, but the mean number of bit errors in a damaged MP is only 2.8 bits.

ii) When E_b/N_o becomes as low as 2.8 db, the MP error rate increases to 3.9×10^{-2} and the mean number of bit errors in a MP increase to 4.61 bits. Under these conditions damaged MP's seem to cluster. This may be prevented by the scrambling technique. In any case when this sort of fluctuation in E_b/N_o is expected due to fading and multipath, the system should be designed with a fast turn around time and the excess capacity to handle a large number of retransmissions over a short period of time.

It has been demonstrated experimentally that an L band aeronautical satellite channel operating in the range (1540 - 1660 MHZ) may be classified as a Rician fading medium. In such a channel the receiver is presented with a steady undistorted signal and a diffuse scattered signal which has Rayleigh amplitude statistics, plus additive Gaussian receiver noise. In this case the key parameters again are E_b/N_o and direct to indirect power ratio S/I. It has also been shown (WILSON, S.G., SUTTON, R.W., SCHROEDER, E.H., 1976) that to achieve a 10^{-5} bit error probability with uncoded differential PSK with S/I = 10 db, E_b/N_o has to be more than 10db about the free space channel requirements. These results are based on a parameter $\rho = 0.9$ where ρ is related to the fading bandwidth B_m and R_b the bit rate as follows

$$1 - \rho \simeq \pi^2 \cdot B_m^2 \cdot R_b^2$$

In other words if a 10^{-5} probability of error is desired and S/I = 10 db, about 10 db extra E_b/N_o is required above the nonfading requirements, depending on fading bandwidth.

At this point it is possible to relate TI-NET experiments to the Rician channel. First of all because of the special structure of MP's it is desirable but not necessary to have 10^{-5} bit error probability which corresponds to $E_b/N_o = 4.5$ db at 16.124 Mb/s. Secondly forward error correction provides a 5.1 db coding gain. This could be further improved by the use of the scrambling technique mentioned previously to cope with long bursts of errors. Thirdly the proposed communication system with its novel MP structure can easily cope with an E_b/N_o as low as 2.8 db by having 22.5 ms. turn around time and by being able to make 25% of its capacity available for retransmissions without jeopardizing the real time ground to air traffic. Moreover from the clustering of the damaged MP's in these experiments it is clear that the loop back TI-NET channel is far from being an ideal non-fading free space channel. It is believed therefore that the measurements carried out in these experiments to investigated the performance of the proposed communication system under marginal E_b/N_o conditions are applicable to an aeronautical satellite channel provided some caution is exercised in mapping from one medium to the other.

VI SUMMARY AND CONCLUSIONS

This paper describes a novel approach to the design of an all digital aeronautical satellite communication system based on TI-NET principles.

A set of experimental results are also presented which describe the behaviour of system under reduced power conditions. The results clearly indicate the feasibility of the proposed system.

The basic limitation on the bit rate of the ground to aircraft link seems to come from the power limitations of the satellites (E_b/N_o), multipath delay of $\pm$ 2.5 μs, and S/I ratio. These factors are all related to the cost, the number and the logistics of the satellites to be used and whether or not independent surveillance is necessary. However these topics are beyond the scope of this paper.

VII ACKNOWLEDGEMENTS

The financial assistance provided by the National Research Council of Canada has made this work possible and is gratefully acknowledged. The Department of Communications, Carleton University's Wired City Laboratory and NASA AMES Research Centre are acknowledged for the use of their facilities for satellite experiments: in particular the author wishes to thank the personnel of the Wired City and NASA AMES for their assistance throughout the experiments. The author is also indebted to Mr. D. P. Fulghum of E. Systems Inc., Dallas, Texas for the loan of two VADAC V speech processors used in this word. Finally the author wishes to extend his thanks to Mr. J. Gruber who carried out the experiment and Messrs. B. Wilkinson, B. Searle and S. R. Sherman for their invaluable help and advice.

Bibliography

1. Coll, C. D., George, D. A., Strickland, L. H., Guild, P. D., and Paterson, S. A., October 1975, "Multidisciplinary applications of communication systems in teleconferencing and education", IEEE Trans. on Comm., (Special Issue on Social Implications of Telecommunications), Vol.23, No.10, pp.1104-1118.

2. Evans, W. M., Davies, N.G. and Hawersaat, W.H., 1976, "The Communications Technology Satellite (CTS) Program", Communications Satellite Systems: An Overview of the Technology, Ed. by R. G. Gould and Y. F. Lum, IEEE Press, pp.13-18.

3. Fulghum, D. P., October 1974, "An All-Digital Vocoder", EASCON Record, Washington, D.C., (IEEE Pub.

74CH0883-1 AES), pp.50-50G.

4. High Level Data Link Control Procedures, ISO/TC 97/SC 6 (Tokyo-17) October 1974.

5. D. J. Horton and P. G. Bowie, June 1974, "An overview of data route: system and performance" Minneapolis, The Computer Communications Group - Trans-Canada Telephone System, International Conference on Communications.

6. Ornstein, S.M., Heart, F.E., Crawther, W.R., Rising, H.K., Russell, S.B. and Mitchell, A., "The Terminal IMP for the ARPA computer network", Proceedings of AFIPS 1972 Spring Joint Computer Conference, Vol.40, pp.243-254.

7. Rybczynski, A., Wessler, B., Despres, R. and Wedlake, J., "A new communication protocol for accessing data networks - the international packet-mode interface", Proceedings of AFIPS 1976 National Computer Conference, pp.477-482.

8. Ulug, M.E., "Systems design of a transparent intelligent network for data and voice", submitted to IEEE Transactions on Communications.

9. Ulug, M.E., "Unique channel flow control and error correction techniques for a transparent intelligent network", submitted to IEEE Transactions on Communications.

10. Ulug, M.E., Weir, D.F., Morris, L.R., and Gruber, J.G., Toronto, Ontario, August 3-6, 1976, "An experimental transparent intelligent network", Proc. of 3rd ICCC, pp.323-329.

11. Ulug, M.E., Gruber, J.W., September 1977, "Statistical multiplexing of data and encoded voice in a transparent intelligent network", to be published in Conference Record of Fifth Data Communications Symposium; Snowbird, Utah.

12. Ulug., M.E., and Gruber, J.G., December 1977, "An experimental satellite link for a transparent intelligent network", Los Angeles, California, submitted to NTC Conference to be held.

13. Viterbi, A.J., January 1976, "Error control for data communication", Computer Communication Review, Vol.6, No.1, pp.27-37.

14. Ulug, M.E. and Gruber, J.G., December 1977, "An experimental satellite link for a transparent intelligent network", Los Angeles, California, submitted to NTC Conference to be held.

15. Woodford, J.B., September 1973, "Aerosat performance specifications", EASCON Convention Record, pp.139-145.

INDEX OF WORDS

Transparent intelligent networks

Aeronautical communications

Satellite relays

Encoded voice

Statistical Multiplexing

Convolutional encoding

Viterbi soft decision decoding

Rician channel

Fading and multipath

Bit error rate probability

Protocol transparency

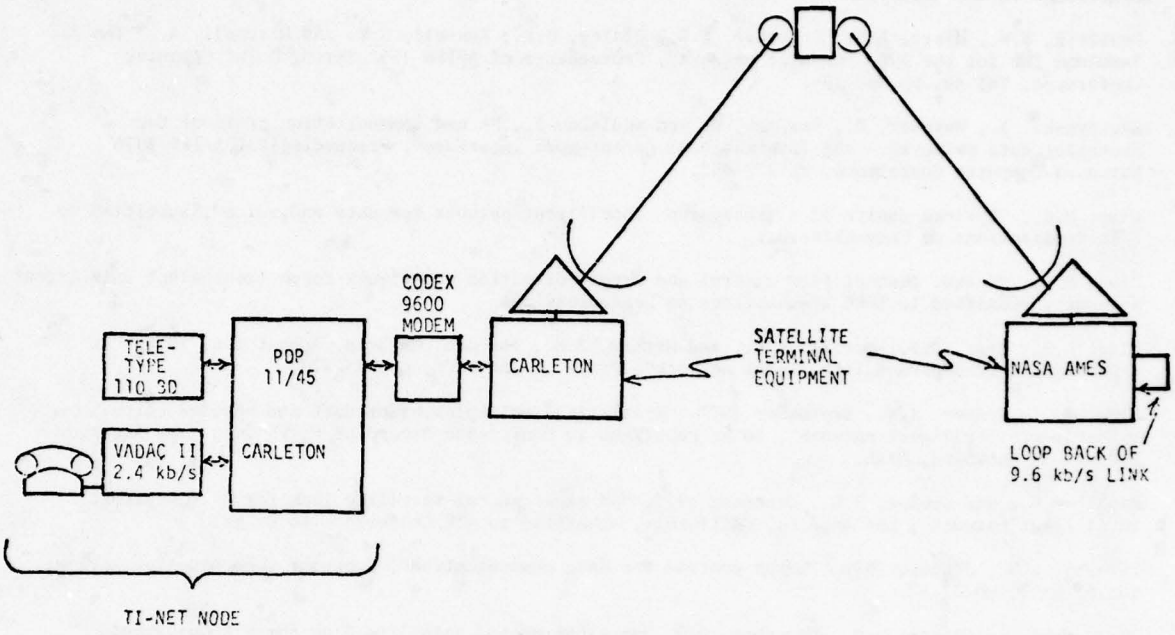


Figure 1 Satellite Loop Back Experiment

CENSAR TDMA
CENTRALIZED SYNCHRONIZATION AND RANGING FOR TIME-DIVISION MULTIPLE ACCESS

Dr. Peter P. Nuspl
Communications Research Centre
Ottawa, Canada

SUMMARY

This paper reports on experiments in synchronization for Time-Division Multiple-Access (TDMA) systems for satellite communications. After a brief introduction to TDMA the report covers the basic principles of the new synchronization ranging concept called CENSAR and explains the technological requirements. Hardware and software features of the implementation are described and the results are discussed. A major result is that this open-loop synchronization method is feasible with guard times of less than 30 ns.

1. INTRODUCTION

The Department of Communications in Canada recognized the desirability of digital communications experiments using the Communications Technology Satellite (CTS), now named HERMES. During the embryonic period in 1971, general guidelines were to investigate new concepts and to develop advanced techniques in high-speed communications technology. The resulting experiment focused on a novel approach to precise synchronization for Time-Division Multiple-Access (TDMA) systems for satellite communications.

This paper covers the basic principles of the new CENtralized Synchronization And Ranging (CENSAR) concept and explains the technological advantages and constraints. Two sections discuss hardware and software features of the implementation using Canadian ground stations and the HERMES satellite. The significant results are presented and their implications for future TDMA systems are discussed.

2. SYNCHRONIZATION FOR TDMA

TDMA technologies for geostationary communications satellite applications have had about a decade of intense development by numerous organizations in several countries. The main reasons, applicable to heavy-route systems, are expectations of higher capacities, greater efficiencies, system flexibility and inter-connectivity in the associated communications systems. Compatibility with the use of spot-beam antennas and satellite switching is also a major factor in creating interest in the techniques. However, tere are only a few TDMA systems in operation, suggesting that further developments and cost reductions are needed before TDMA will be competitive with the more mature FDMA approach. The initial impetus in TMDA technology development was for INTELSAT applications, but recent activities are aimed at regional and domestic systems, sometimes with special-purpose (e.g., data only) applications.

TDMA is the shared use of a satellite repeater by several user stations transmitting bursts which are timed so as to interleave without overlap at the satellite. Since high utilization efficiency is usually a requirement, small guard times, which are deliberate gaps between bursts, are allocated. Since the satellite moves about its nominal position, there is a continuing need for synchronization which is the availability of required timing at all stations in the network. Timing accuracies of better than 100 ns are usually expected in high-capacity networks.

3. THE CENSAR EXPERIMENT

A new synchronization sub-system for TDMA has been given the acronym CENSAR, derived from CENtralized Synchronization And Ranging. The concept was described in a report by de Buda.(de Buda, R., 1972) The objectives of the HERMES CENSAR experiment included demonstration of technical feasibility and assessments of timing accuracies and guard times.

The experimental CENSAR system discussed in this paper consists of a central control station, three ranging stations and a single user station from a possible network of stations. This system and stations were designed for use with a geostationary communications satellite with spot-beam antennas, such as HERMES. The control station sends short bursts of microwave radio signals having wide bandwidth and carrying control data via the satellite, to all stations. Each ranging station sends back a ranging burst at a time specified by the control data. The control station measures the total transmission delays (elapsed time) from the three stations, estimates its own delay to the satellite and corrects the control data accordingly. Becuase each user station calculates its own delay from a knowledge of the station co-ordinates and from the control data, and thus achieves synchronization without any transmissions of its own, CENSAR is in the class of 'open-loop' synchronization systems.

Section 4 contains concise statements of CENSAR principles and discussions of them. Computer algorithms are treated in section 5, whereas the specific implementations for the experiment are discussed in Section 6. After a brief description of the field trials, Section 8 summarizes the results.

4. PRINCIPLES AND FEATURES OF CENSAR

The CENSAR concept is based on four principles:

(a) Measure Effective Delays Through Three Satellite Links.

Delay measurements are made at a central control station, with the co-operation of stations called ranging stations. From the geometry illustrated in Figure 1, it will be apparent that only three measurements are necessary. These are best achieved by transmitting and receiving broadband signals in a

burst mode (synchronization and ranging bursts), yielding full information on actual propagation delays through the equipment, propagation paths and the repeater. Each ranging station responds to control data for its use only, establishing three very precise delay-lock loops. The ranging bursts do not interfere with user stations. The required measurements can be made at a central control station with only moderate requirements in signal processing and computation.

(b) Broadcast Relevant Information to the Network.

The control station uses a very small-capacity synchronization burst to send the necessary control data to all stations.

(c) Recover Frame and Symbol Timing.

Each station receives the synchronization burst, recovers frame timing and demodulates the control data in this burst. It is a significant principle of CENSAR that network symbol timing is recovered from the synchronization data bursts; the network is said to be bit synchronous.

(d) Calculate the Transmission Delay at Each Station.

Through a deliberate choice of unique basis vectors, since each user station is given its precise co-ordinates and receives the control data, each station calculates its own delay to the satellite.

From these principles flow some basic features of CENSAR systems. Centralized control can be used in practical applications, measurements and computations being done at a single control station, in co-operation with only three remote stations.

CENSAR is in the class of 'open-loop' synchronization systems. This has been alternately expressed as passive synchronization, because the user stations achieve complete and precise synchronization without transmissions of their own. They are initially synchronized when all synchronization data is received.

Thirdly, through the use of short bursts, with a minimum number of measurements and control data, high efficiencies are achievable even for large numbers of accesses. The remote regeneration of network symbol timing keeps the preamble requirements very low. For example, it is feasible to have 30 accesses at 97% efficient use of available capacity.

It is also noteworthy that CENSAR is compatible with present-generation satellite technology and would also be compatible with future spot-beam systems and satellite-switched TDMA (wherein fast switches in the satellite redirect bursts to designated beams).

Since most applications of TDMA are very sensitive to costs, the potential economies of CENSAR must be emphasized. With only simple equipment of small size and high technology, but requiring no trained operators, a cost-effective TDMA network can be envisioned.

5. COMPUTER ALGORITHMS

In the development and understanding of the computational procedures, it is very useful to view the problem as geometrical. A geocentric cartesian co-ordinate system is selected and the locations of all stations are found by accurate surveys. Other small effects temporarily put aside, the requirement is to solve for the position of the satellite.

In a system having spot beams, it is usually the case that a station cannot receive its own signals. This implies that two-hop signalling is necessary between pairs of stations. In CENSAR, the control station and a ranging station are foci of an ellipse; the satellite lies on the ellipse defined by the sum distance T_1+T_K, where K = 2,3,4, in Figure 1. The satellite resides on the ellipsoid of rotation obtained by rotation of the ellipse about its major axis. Three such rellipsoids with foci carefully selected to provide independent information are necessary and sufficient; in principle, the satellite is at the intersection of the ellipsoids.

A direct solution is unattractive due to computational requirements. Indeed, the viability of the CENSAR concept is attributed to vector algebraic transformations which yield tractable formulations. (Nuspl, P.P., 1974) The first of these starts with survey data and computes coefficients for use in the other algorithms. Such calculations are done off-line, with high precisions and accuracies; they are required only once for a given network. Another related off-line processing stage consists of calculation of CENSAR co-ordinates for each of the user stations.

Real-time algorithms have been devised to achieve high accuracies and yet be programmed for the modest resources of a mini-computer. At the control station only, the main procedures are the control of the delay-lock loops and the calculation of the delay (T_1+T_1) (which cannot be measured in a spot-beam system). There are numerous other housekeeping, data logging and check functions performed also. At each user station, a simple computation with known coefficients and received data yields the required burst delay for the station.

6. IMPLEMENTATION OF THE EXPERIMENT

Whereas the above has discussed principles and algorithms applicable generally, the following sections describe the technical preparations for the experiment with HERMES and ground stations in Canada. In all phases of the experiment, the central control station was in Ottawa at the 9m antenna. The required three ranging stations were located at London, Rouyn and Quebec City in Phase I and at Brandon, Thompson, (Manitoba) and Thunder Bay (Ontario) in Phase II, the latter illustrated in Figure 2. A convenient method of evaluating the accuracy of the synchronization technique was to measure the burst-timing error of the signals from a fifth station which utilized only the broadcast synchronization data,

thereby simulating one of the many user stations of a CENSAR-timed network. The design of the experiment included calculations in real-time and off-line of the difference between predicted timing and measured timing for the station.

There were serious constraints on the implementation. The HERMES transponder configuration had been selected; the two separate channels and two spot beams required new algorithms and influenced the bandwidth and type of signals. The ground stations had also been selected as to type and function; the transmit power was increased to 20w in part to satisfy the ranging station requirements. Even so, ranging bursts had to be longer than planned and coded to permit possible processing gains at the control station. Some problems and delays were caused by the need to co-locate and to share the stations with other experimenters.

Accurate station surveys are necessary for precise open-loop synchronization and for satellite position determination. Through the co-operation of the geodesy group of the Department of Energy, Mines and Resources, all locations were eventually surveyed to accuracies of about $\pm 5m$ in each co-ordinate.

6.1 Measurements

Measurements of delay time T_1+T_K, K = 2,3,4,5 are obtained from four delay-lock loops. Each loop is identical but is processed in its own quarter of the cycle. The fundamental purpose of each loop is to maintain a precise arrival time of a ranging burst, which is suitably observed and controlled at the control station. The control data in each loop are independent and are the required measurements. A cycle repeated every 640 ms is required for synchronization control; so there are many data points available for filtering and calculation of orbital data.

Details of the delay (distance) measuring equipment (DME) are described in another paper;(Nuspl, P.P., 1975) an appreciation of the precision of the method is obtained from a brief discussion. Conceptually, the range is determined in four parts. A large constant is known from the positions of the stations and the satellite. A frame measurement establishes the delay to multiples of 125 μs; this is accomplished by detection of the trailing edge of ranging bursts. A bit measurement is made by correlation of the ranging burst with a reference signal and yields a precision of 1 bit = 15 ns. When such correlations are present, a discriminant function (table look-up) yields a precision of about 1 ns, in a fine measurement portion for which extensive calibrations are necessary.

By their very nature, these measurements must, in large part, be made by hardware carefully designed and constructed. However, in this implementation many tasks have been delegated to a mini-computer. Via real-time software, the raw measurement is acquired, checked and the composite measurement is put together. By smoothing and prediction the loops are kept in precise lock.

6.2 Computations

Over an extended period, the algorithms were programmed on the computers and thoroughly checked. Whereas the hardware was procured on contract, all software was developed at Communications Research Centre with contract support. Through extensive orbit simulations, many situations could be tested while the software was being developed. The integration, interfacing and testing of the complete synchronization system required much special programming. Separate, stand-alone routines were developed for the user station computer.

6.3 Start-Up

The CENSAR concept does not require orbital information, so a very important procedure is start-up, by which the synchronization network becomes operational. Procedures were developed by which fully automatic start-up is achieved with up to four ranging stations simultaneously. Typically, the four loops become locked in less than 10 seconds and seldom require more than 30 seconds.

6.4 Data-Log and Control Functions

A set of new measurements are made every 640 ms; for post-processing in TDMA and orbital studies, a concise data set is written on magnetic tape. This facility has enabled many repeated processings of interesting data.

The events in this experiment occur too quickly for detailed operator roles, but for experiment control it was desirable to have a control/monitor program. Digital displays and pen recorders, the status of ranging stations and key parameters were all under console control.

7. FIELD TRIALS

The first field trials were conducted prior to launch of HERMES. With a transponder simulator on a boresighted tower 15 km away, RF tests checked out the delay-lock loops and calibration procedures. In this phase, many problems in interfacing were encountered and solved. Confidence in the expected accuracies was achieved.

Phase I took place in the spring and summer of 1976, with stations in Ontario and Quebec. The feasibility of CENSAR and the validity of the algorithms were verified. There were many problems in keeping CENSAR equipment and stations in operation.

Phase II was carried out in the spring of 1977 with stations in the Manitoba and Ontario areas. More extensive data over longer periods were accummulated. Operations were much smoother and numerical results were yet more pleasing.

8.0 RESULTS

The essential results of the synchronization portion of this HERMES experiment are that the CENSAR concept had been shown to be valid and that the selected implementation is feasible. Preliminary results have been reported (Brown, K.E., 1976 and Nuspl, P.P., 1977) and a paper will be presented at the fourth International Conference on Digital Satellite Communications, to be held in Montreal, Canada in October 1978. Specific technological advancements and technical data are discussed in this section.

8.1 Development of the CENSAR Concept

In its original form, this new type of TDMA synchronization scheme applied only to global systems, wherein geometries are spheroidal and stations can receive their own signals. By extension of the concept to spot-beam satellites, the scope of potential applications was considerably widened. The advent of satellite switching is expected to increase interest in centralized synchronization.

8.2 Algorithms

Computational procedures have been developed for spheroidal and ellipsoidal geometries. The satellite is at the intersection of three of these surfaces; since a direct solution in real-time is not viable,(Nuspl, P.P., 1974), an algorithm which solves this problem is a key constituent of CENSAR. Algorithms have also been designed to control the delay-lock loops and the open-loop bursts by user stations. A start-up algorithm is completely automatic and requires no orbit data. All of these algorithms were extensively analyzed, checked in simulations and verified in operations.

8.3 Hardware

Equipments have been designed according to CENSAR principles and constructed using 1975 technology. The new developments include a multi-function synchronization burst generator, timing recovery circuits, burst ranging techniques which have a precision of 1 ns, delay-measuring equipment which is used for four measurements, and interface circuits. Crucial to the performance are the two-hop delay-lock loops which operate independently in a time-share mode with three ranging and one check stations.

8.4 Software

Algorithms have been programmed to operate off-line as much as possible, yielding coefficients and parameters for real-time operation. Real-time software was prepared for a modest mini-computer facility (PDP 11 family). These programs include operations control and display, and also data log routines.

8.5 Integration

A significant aspect of this project has been the integration of technologies and equipments. The delay measurements are a prime example of effective integration of hardware and software methods. Due attention was paid to interfacing of the computer and the measuring equipment and of the IF and RF equipments. Calibration of the complete system required careful procedures to yield precision and accuracy.

8.6 Technical Performance

The performance of the CENSAR system, as implemented for synchronization tests with HERMES, is summarized in Table 1. A user station recovered symbol timing to about 2 ns under operating conditions; the self-noise contribution has been measured to be ± 0.2 ns rms. The control bits broadcast to all stations were sent in an erasure channel, with erasures at one per minute under operating conditions. From measurable error rates in the signalling system, the calculated error rate for these control bits is negligible (1 in 10^{20} approximately). Figure 3 is a typical result of range measurement and Figure 4 illustrates system errors from all sources.

Start-Up Time:	typical - 10s
	worst case - 30s
Symbol Timing Recovery:	accuracy - < 2 ns peak error
	± 0.2 ns rms self-noise
Control Bit:	erasure rate, typical - 1/min
	error rate - < 1 in 10^{20} (calculated)
Synchronization:	loop precision - ± 1 ns
	accuracy, short term - ± 5 ns
	24 hour - ± 20 ns (3σ)
	guard time - < 30 ns (2 bit durations)

TABLE 1: CENSAR Performance Data

8.7 Data Base

All data produced during CENSAR operations has been recorded in concise form on a few magnetic tapes. This data base was used in checking the accuracy of the CENSAR system and in orbital studies. Some of the data is also being used to investigate doppler effects on time transfer.

8.8 TDMA Synchronization Costs

Being experimental and pioneering, the CENSAR experience cannot be expected to yield accurate information on costs for future systems. However, project allocations in dollars and man-years are indicators which can be used for planning and budgetary estimates. For example, IF equipment for five stations was procured for less than $500 000 (1975). DOC man-year allocations ranged from one professional to a high of two professionals plus two technicians before and during operations. Estimated costs for specific-purpose TDMA stations are $25 000 per station (IF equipment for TDMA, in quantity production).

9.0 CONCLUSIONS

The CENSAR TDMA synchronization experiment via HERMES has met its objectives. This new system of centrally-controlled synchronization was developed and demonstrated. The technical performance illustrates guard times of 30 ns and short preambles, which make high efficiencies possible. This proven open-loop system is being followed by other investigations for applications in international and domestic systems.

REFERENCES

1. Brown, K.E., and Nuspl, P.P., 1976, "Early Operational Experience with a New TDMA Synchronization System Through CTS", Communications Research Centre, Technical Note #682, October 1976.

2. de Buda, R., 1972, "Synchronization for Time Division Multiple Access Experiments", Technical Report RQ72EE4, Canadian General Electric Company Limited, Toronto, Canada, June 1972.

3. Nuspl, P.P., and de Buda, 1974, "TDMA Synchronization Algorithms, EASCON '74 Record, Washington, D.C., 1974, pp 656-663.

4. Nuspl, P.P., Davies, N.G., and Olsen, R.L., 1975, "Ranging and Synchronization Accuracies in a Regional TDMA Experiment", INTELSAT/IECE/ITE, Third International Conference on Digital Satellite Communications, Kyoto, Japan, November 1975, pp 292-300.

5. Nuspl, P.P., and Mamen, R., 1978, "CENSAR (CENtralized Synchronization And Ranging) TDMA and OPME (Orbit Perturbation Measurement Experiment)", Proceedings of the HERMES Symposium, Ottawa, Canada, 29 November - 1 December 1977, in publication.

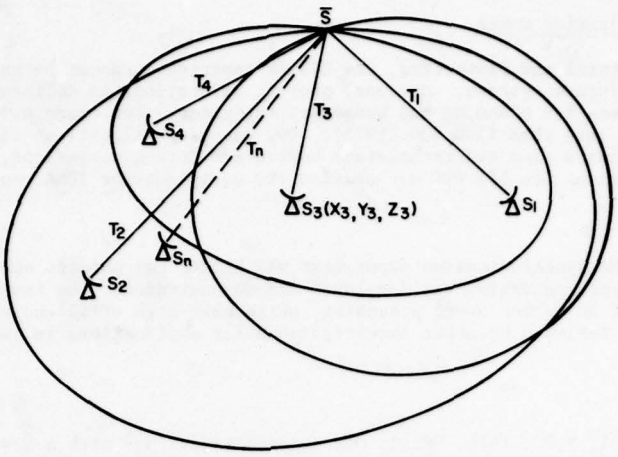


Figure 1: Geometry for Spot Beams.

(x_1, y_1, z_1) and (x_k, y_k, z_k) are foci of ellipsoids of rotation, k = 2,3,4

$\bar{S}$, satellite is on the surface of each ellipsoid.

The three measurements T1+T2, T1+T3, T1+T4 are necessary and a fourth measurement T1+Tn is used to check the system performance.

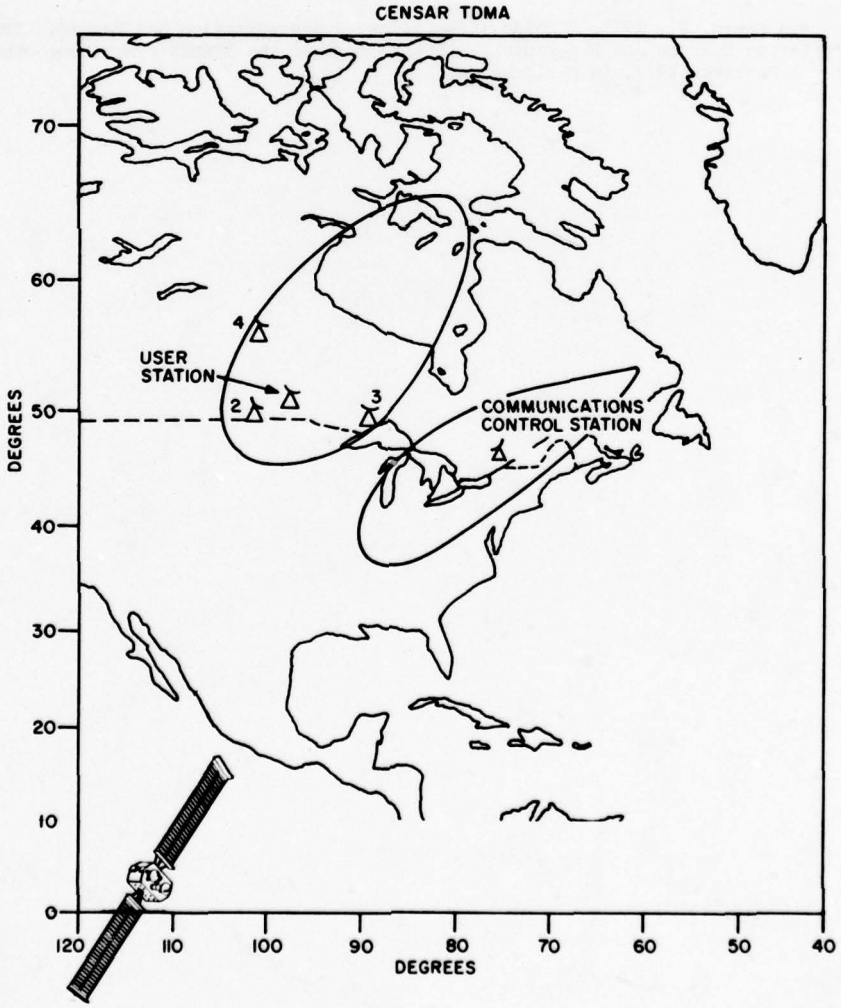


Figure 2: Typical Regional Coverage and TDMA Network.

This configuration for Phase II shows the triangle of ranging stations and a user station. The separated spot-beam coverages by HERMES are illustrated. HERMES is located at 116°W.

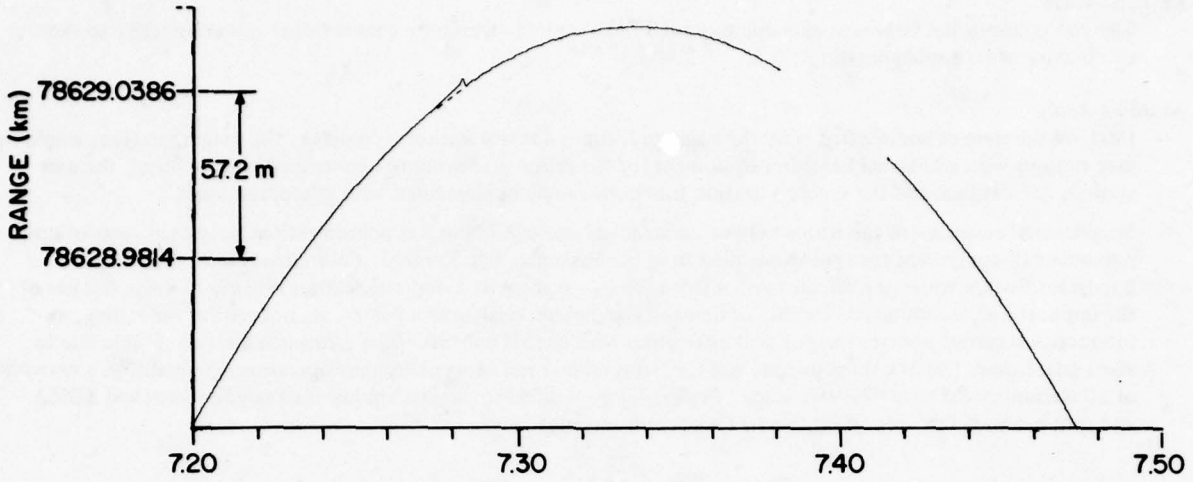


Figure 3: Range SUM T_1+T_2 Versus Time.

In the period 720-750 GMT, 21 August 1976 this range measurement was made during a transition through a maximum range. The resolution is about 0.3m.

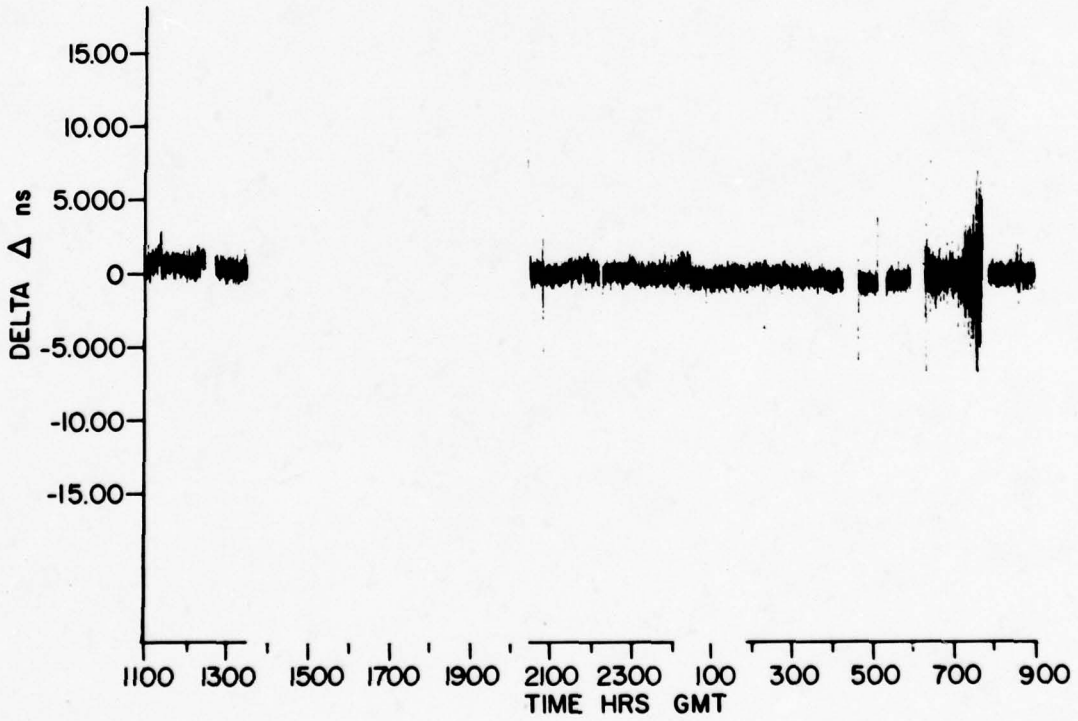


Figure 4: System Error (calculated delay-measured delay).

The delay is calculated from the synchronization control data and the measured delay is from the fourth measurement.

DISCUSSION

I.L.Lebow, US

Can you quantify the improvements due to your TDMA system, over more conventional systems relative to the complexity of its implementation?

Author's Reply

First, on the item of complexity, only the communications control station is complex; the ranging stations would be user stations with additional baseband equipment for the purpose of sending a controlled ranging burst; the user stations are identical and the synchronization functions can be implemented with microprocessors.

So, *potential economy* of the whole network is an advantage of CENSAR synchronization; projected costs of non-redundant IF equipment for synchronization have been estimated at $25,000. *Centralized control* is a very important feature which is difficult to quantify as to its advantages; a domestic carrier is likely to make full use of the implications, including possibilities of demand assignment, peak load adjustment, and centralized billing; an international carrier perhaps has political difficulties with central control. *High efficiencies* are achievable due to short guard times (30 ns), short preambles (4 to 10 symbols) and bit-synchronous operation; to quantify, a network of 30 stations could have 97% efficiency. *Compatibility* with advancing technologies of satellite-switched TDMA and spot-beams is very important from a long-term viewpoint.

A DIGITAL COMMUNICATION SYSTEM AS GATEWAY BETWEEN ADJACENT COMPUTERIZED AIR TRAFFIC CONTROL CENTRES

Dr.-Ing. M. Baum
EUROCONTROL Agency
Maastricht U.A.C.
Postbus 78
Beek (L)
The Netherlands

SUMMARY

This paper describes a practical approach to improve data communication between adjacent computerized air traffic control centres. Different point to point links and local networks can be bridged by a transit facility. The applied concept is "communication procedure conversion" by which systems with incompatible transmission procedures can communicate with each other via a gateway system. Procedure conversion meets the basic requirement which is not to intervene in the functioning of the involved local networks and not to require modifications of locally used procedure envelopes. The transit function is realized by conversion of the procedure specific message type and priority on one hand and of address information on the other hand. A flexible table mechanism is chosen instead of fixed coding. Outstanding problems are an efficient link failure processing and an envisaged attachment of a circuit switching function to the message switching principle.

1. INTRODUCTION

The subject described in this paper concerns ground to ground data communication between computerized air traffic control centres. Some of the national Air Traffic Control (ATC) authorities use already data links via private lines between their ATC computers. Such communication systems exist in the United States between FAA Air Traffic Control Centres. In Europe data links are implemented in France between the Air Traffic Control Centres, using the CAUTRA (Coordinateur Automatique de Traffic Aerien) and in the Federal Republic of Germany between centres equipped with the DÜV (Datenübertragungs- und Verteiler-System). All these systems are local configurations with their own different communication procedures. They are not compatible with each other so that communication and coordination across national borders in Europe is still a problem.

2. COMMUNICATION ENVIRONMENT

2.1 Point to Point Data Links for MADAP

Maastricht U.A.C., a EUROCONTROL air traffic control centre is responsible for the control of the so-called General Air Traffic in the upper airspace over Belgium, Luxembourg and the northern half of the Federal Republic of Germany as well as for the control of the Operational Air Traffic in the airspace over the northern part of Germany through the services of the German Air Force. Its highly automated system MADAP (Maastricht Automatic Data Processing and Display System) has to exchange up to date flight plan information with its adjacent national centres for each aircraft which passes its boundaries (EUROCONTROL, 1977). For this purpose a point to point data link between the Paris centre (CAUTRA) and Maastricht has been installed (figure 1). A dual channel point to point link protected by the DÜV procedure has been tested and is ready for operational use. It is planned to install a link with the DCTS (Digital Communication Terminal System) at the EUROCONTROL centre at Karlsruhe, using the transparent line procedure ADIS/CIDIN. Basic flight plan information as filed flight plans (FPL), departure messages (DEP), delay messages (DLA), etc. are still received via the Aeronautical Fixed Telecommunications Network (AFTN), a telex-type network with low transmission rates of 50 to 100 baud.

Whilst technical means for data exchange are available to some extent, little has been achieved in the definition of messages which are to be exchanged between automated control centres. One of the reasons is that the original concept of exchanging current flight plans (CPL) could not be implemented as expected. Other modes of data exchange between ATC computers were therefore investigated. In 1973, first proposals were made to introduce the Activate (ACT) message as a substitute for the verbal boundary estimate in the usual centre to centre coordination of aircraft movements in the event of hand-over from one unit to the next.

About 15 minutes prior to the time at which the aircraft is estimated to arrive over the agreed coordination point an ACT is to be sent to the next centre which is to activate the corresponding flight plan in the receiving computer system automatically. A further type of data which is expected in inter-centre data exchange are transmissions from data bank to data bank of adjacent ATC computers with interactive controller access.

For an exchange of such and other messages hardware and software have been developed for the handling of transmission procedures for a protected data exchange via point to point links. A special communication system has been installed which serves MADAP in handling different procedures with their special treatment of channel and system failures. This Digital Communication Terminal System (DCTS) is designed as an

autonomous front-end computer complex for the exchange of data between MADAP and external units (BAUM, M., 1975 and CIT-ALCATEL, 1978).

2.2 DCTS Functioning

DCTS is based on a duplicate MITRA 15 minicomputer of 32 K 16 bit words core memory (figure 2). Discs and magnetic tape units are used to extend its storage capacity for functions which allow slower access times. For communication purposes micro-programmed telecommunication processors are available which execute duty-cycle expensive processing, as e.g. detection of control characters and flags, calculation of character parity and of block check characters. All peripherals are duplicate, and both chains are processing in parallel in order to be immediately ready for an automatic reconfiguration in case of pilot chain stops. AFTN message reception via low speed channels requires uninterrupted availability of the system. DCTS is running autonomously, although it is a subsystem of MADAP. In case of MADAP outages it must be able to function without interruption. All AFTN messages are stored for at least one hour on disc from which they can be retrieved on request. On printers all messages which enter the system via low or medium speed channels can be printed in case of link failures. In the following two national communication networks, with which DCTS has established technical links via protected medium speed lines and their special transmission procedures will be sketched.

2.3 DÜV Network

The DÜV network (Datenübertragungs- und Verteiler-System) is a star-type network of the German administration for air traffic services BFS (Bundesanstalt für Flugsicherung) with its centre in Frankfurt (AEG-TELEFUNKEN, 1974). Via this network NOTAM bulletins and flight progress strips are transmitted to the regional control centres Frankfurt, Düsseldorf, Bremen and Munich with a large amount of attached civil and military terminals (figure 3). In a further development phase also the distribution of AFTN messages within Germany is planned in order to save line costs by replacing bundles of low speed lines by several, in total less expensive, medium speed lines. The applied DÜV transmission procedure is a character-oriented (CCITT5) synchronous serial line procedure for medium speed data exchange from 200 to 4800 bps in full duplex mode of transmission. Four message priorities are available in order to insert urgent, in general short, messages into long sequences of less urgent messages. For instance NOTAM bulletins will be interrupted by flight progress messages. A usage of two lines per circuit is foreseen in order to enable line reconfiguration without message rerouting. A line reconfiguration is executed automatically after three transmission trials replied by NAK (Negative Acknowledgement) or after time-out.

The DÜV procedure has been implemented in DCTS and tested between DÜV Frankfurt and Maastricht. It has been proven that the point to point link is technically ready for operational use.

2.4 CAUTRA Network

A further link has been established between DCTS Maastricht and CAUTRA Paris. The basic CAUTRA network has the form of a triangle with the edges Paris, Bordeaux and Aix-en-Provence (figure 4). The herein applied INTERCAUTRA transmission procedure is a synchronous serial line procedure for half duplex mode of transmission in 2400 bps. It is a character oriented procedure using CCITT5 code similar to DÜV. Only one block priority exists. Error control is done both by character parity checks and a longitudinal redundancy check. A complex system of time-out controls has been installed in order to overcome blocking situations.

2.5 Transit Links

The necessity to handle several different procedures as front-end processing system for MADAP involves to transform different procedure envelopes into one common message header. Procedure conversion is only a little step further, from the common message header a new procedure envelope must be formed. This facility can be used for a point to point link between Paris and Karlsruhe via DCTS Maastricht as transit station (figure 5). In this case conversion would be required between the INTERCAUTRA and the CIDIN procedure.

Basic principle was not to require any modification of the involved procedures. A pragmatic solution had to be developed in short time. Nevertheless a simple procedure conversion from INTERCAUTRA to CIDIN and vice versa could be avoided, and a more flexible approach has been chosen.

3. PROCEDURE ENVELOPES

Instead of converting from only one into another procedure, the analysis was extended to the problem to convert from any transmission procedure, available in DCTS for point to point links, to any other available transmission procedure. There are four procedures concerned : DÜV, INTERCAUTRA, WMO and CIDIN (figure 6). The first three of them are character oriented and use CCITT5 code, whilst the fourth one is a transparent procedure with a bit stuffing mechanism identical to that used by HDLC (High Level Data Link Control) adopted by the International Standardization Organization (ISO). Comparative analysis showed that all investigated line procedures have sufficient common properties to enable the implementation of a general conversion algorithm. All procedure envelope information can be grouped into three fields : Data Link Control Field (DLCF), Communication Control Field (CCF) and Communication Data Field (CDF). The DLCF contains block separation elements as control characters STX, ETX, ETB, etc. for character oriented procedures and flags for CIDIN. Further elements of the DLCF are the block type, block numbering and a longitudinal block or frame check information BCC (Block Check Character), LRC (Longitudinal Redundancy Check) or FCS (Frame Check Sequence). The CCF contains message type, priority, and address information of the sending and receiving stations and terminals. Message code and format are regarded as part of the

message type. The CDF contains the message text which is to be copied without modification from procedure to procedure.

The DLCF is to be striped off, before the conversion mechanism is triggered. Procedure conversion deals only with the CCF. After CCF conversion a new DLCF is added for the control of message output.

4. PROCEDURE CONVERSION BY ADDRESS AND TYPE/PRIORITY CONVERSION TABLES

The CCF can be subdivided into an address part and a message type/priority part. The address part contains the numbers of the emitting and receiving stations and the numbers of the corresponding input and output terminals. The message type/priority part can contain several classifications of messages as priority, type, code and format.

A standard transit message header has been defined which can express all relevant communication control information of the affected procedures in a normalized form. Each input procedure envelope is converted into the transit header, from which it is converted into the output procedure envelope. This transit header is used also for messages which enter and leave the system in the same procedure, i.e. transit traffic without effective procedure conversion.

4.1 Input Address Conversion

The receiving and emitting station numbers of the input procedure are transformed into common logical numbers which are procedure independent by the means of a table INRECMOD (figure 7). The index of this table is a combination of the procedure number and the four low order bits of the procedure specific receiving and emitting station numbers ADEST and AEMM for INTERCAUTRA, DÜNRE and DÜNRA for DÜV and AE and AX for CIDIN.

An invalid address means that the emitting or receiving station number is not defined for the corresponding procedure. Such an indication is marked in the INRECMOD table and forces an error print-out of the message.

4.2 Output Procedure Determination

The transit message header is procedure transparent. MADAP which generates a transit message header for its outgoing messages does not specify the procedure in which DCTS will transmit them. The procedure number is determined by the means of different tables which correlate the logical number of the receiving station with an appropriate output circuit and the output circuit with a procedure.

4.3 Output Address Conversion

Based on the procedure number and the transit addresses the receiving and emitting station numbers of the output procedure are determined by the means of a table OURECMOD (figure 7). The index of this table is a combination of the procedure number and the transit address. An invalid address means that a station number is inconsistent with the procedure derived from the output circuit. A print-out of such a misrouted message is triggered.

Subgroup/terminal numbers are not converted. The originator has to specify the destination terminal in the numbering system of the output procedure. The transit function for terminal numbers consists only in copying this information from the input procedure envelope via the transit header into the output procedure envelope.

4.4 Input Type/Priority Conversion

Type/priority conversion is realized in a similar way as address conversion (figure 8). The type/priority information of the input procedure is transformed into a common transit type information by the means of a table INTYPE which is applied to all available transmission procedures. Table index is an extract of the type/priority information which is relevant for the conversion process.

4.5 Output Type/Priority Conversion

The transit type information is transformed into procedure specific type/priority information by the means of a table OUTYPE which is applied to all procedures. The index of this table consists of the procedure number and the transit type. Both at the access of the INTYPE and OUTYPE tables invalid type and priority indications can be detected. Such conversion errors are in general due to undefined ways of message flow which have not been foreseen at table generation or which are not permitted. In both cases an error print-out is forced in order to trigger corrective actions.

5. GATEWAY

In order to discuss some general aspects of application of the described conversion mechanism some terms of data communication networks will have to be introduced (LLOYD, D. and KIRSTEIN, P.T., 1975). An interconnection of local networks A and B might be needed (figure 9). In such a case a basic requirement is not to interfere with the functioning of the local networks by actions of the others. For short-term solutions it cannot be expected that local procedures may be modified for interconnection purposes. In this context the implementation of an interconnection protocol for bridging different existing networks must be excluded, because this would force important modifications to the involved local networks. Procedure conversion is the only practical solution for the transit station which is called a gateway, when the

conversion is applied to networks. The presented solution is not specific for the DCTS but could easily be implemented in computer systems which have to link incompatible communication systems. The conversion table approach has the advantage to be independent of the characteristics of the available computer and the used high level or assembler language.

6. PRESENT TECHNICAL STATUS

Presently procedure conversion is implemented in DCTS for the three transmission procedures DÜV, INTER-CAUTRA and ADIS/CIDIN. It will be used for a limited re-routing in case of link outages of direct data circuits. For example messages which must be exchanged between CAUTRA Paris and DCTS Karlsruhe can be re-routed via DCTS Maastricht (EURO-DATEX GROUP, 1976).

In case of outages of transit links for longer periods (more than 3 minutes) the concerned messages are printed on emergency printers. A problem occurring at such link failures is the requirement to transmit such messages either to their origin or destination. A possible solution to overcome this problem could be to send system messages to the originator with identification of those messages which cannot be forwarded to their destination. In combination with the system messages the transit station would cancel the blocked messages after a certain time automatically.

7. COMBINED MESSAGE AND CIRCUIT SWITCHING

A further technical development, which is presently in work under the aspect of line cost-saving, is the application of circuit switching techniques in DCTS. Concerning the so far developed gateway function DCTS applies the message switching principle : messages to different destinations and for different customers are transmitted via the same lines.

As for ground to ground communication between air traffic control centres bundles of telephone lines are available, it is aimed to share such lines between voice and data communication. The technical means of interfacing the DCTS systems with the circuit switching systems MOX and KOX of Maastricht and Karlsruhe respectively will be developed in the near future. Based on the 200 circuit series of CCITT5 V24, procedures for seizing and release of a line are specified in relation to the ADIS/CIDIN procedure.

Beyond the technical problems several operational and organizational questions must be answered before an application can start. Priorities between voice and data lines had to be agreed, but will certainly undergo changes in the forthcoming development. Schedules of voice and data line charges under the authority of post offices have far-reaching consequences on the cost aspect of the application of these techniques.

8. CONCLUSIONS

Technical means of data communication between ATC computers are urgently required. It is obvious that the described gateway system is neither intended nor can essentially change the general data communication situation of air traffic control centres in Europe. The development of a Common ICAO Data Interchange Network (CIDIN) is an important aim and must be regarded as the only efficient long-term solution for ATS data exchange. The proposed gateway solution applied to DCTS and other systems might improve data communication between air traffic control centres in an interim period, before CIDIN will reach a significant stage of implementation leaving apart any further CIDIN planning. In relation to a future CIDIN network the present DCTS systems will not represent switching centres but will have the qualifications of tributary stations. Nevertheless the techniques of flexible procedure conversion should be applicable for CIDIN switching stations and other tributary stations in order to convert between CIDIN and other local procedures.

9. REFERENCES

AEG-TELEFUNKEN, 1974, "Datenübertragungs- und Verteiler-System - Systemkurzbeschreibung".

BAUM, M., 1975, "DCTS - a system for data communication between air traffic control centres", Paca Press, Conference Proceedings EUROCOMP 75.

CIT-ALCATEL and MAASTRICHT SYSTEMS DIVISION, 1978, "DCTS - software documentation, contract C/26/E/HG/76".

EUROCONTROL MAASTRICHT U.A.C. OPERATIONS DIVISION, 1977, "MADAP - Operations System Manual".

EURO-DATEX GROUP, 1976, "WP15 - Considerations of the projects in the Eurocontrol 5-year plan for automatic data communications in relation to possible developments with the Common ICAO Data Interchange Network (CIDIN)".

LLOYD, D. and KIRSTEIN, P.T., 1975, "Alternative approaches to the interconnection of computer networks", Paca Press, Conference Proceedings EUROCOMP 75.

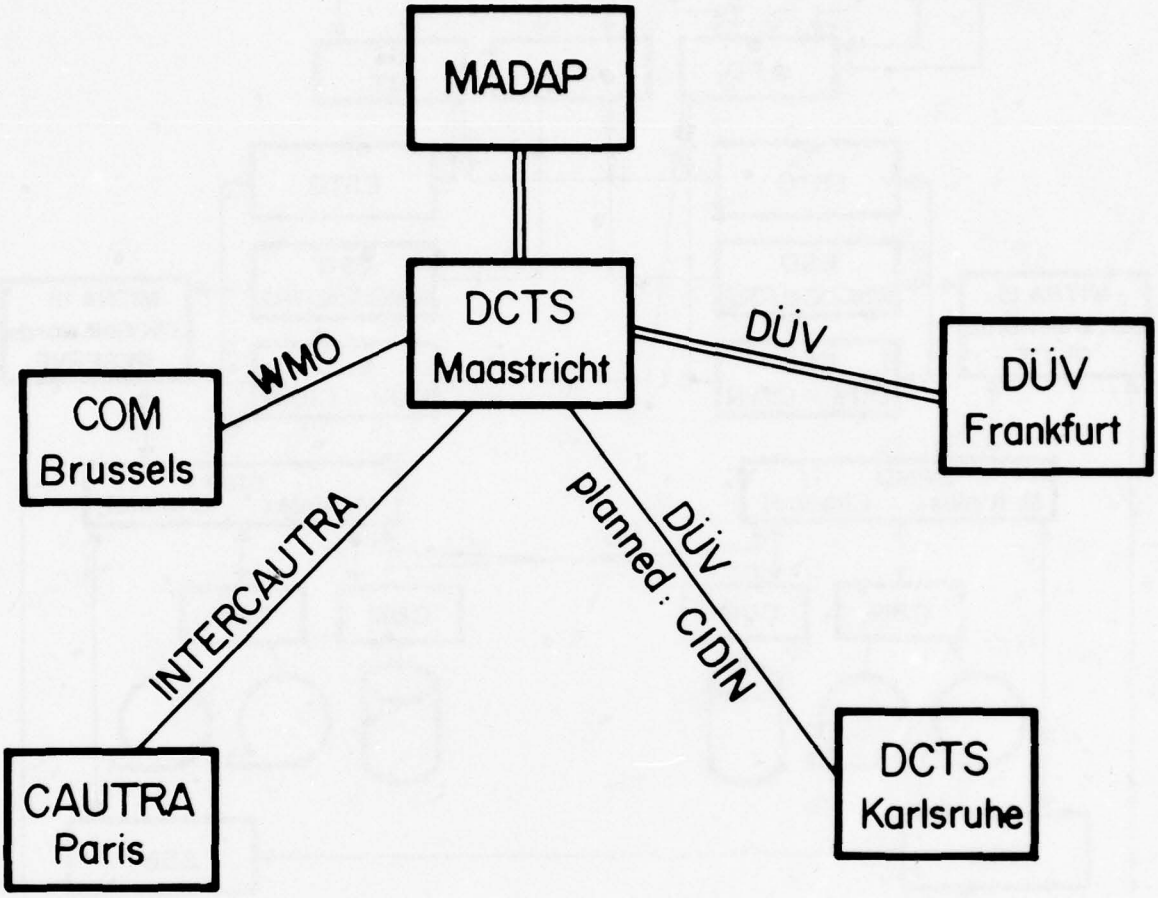


Fig.1 Point to point links for MADAP (Maastricht Automatic Data Processing and Display System)

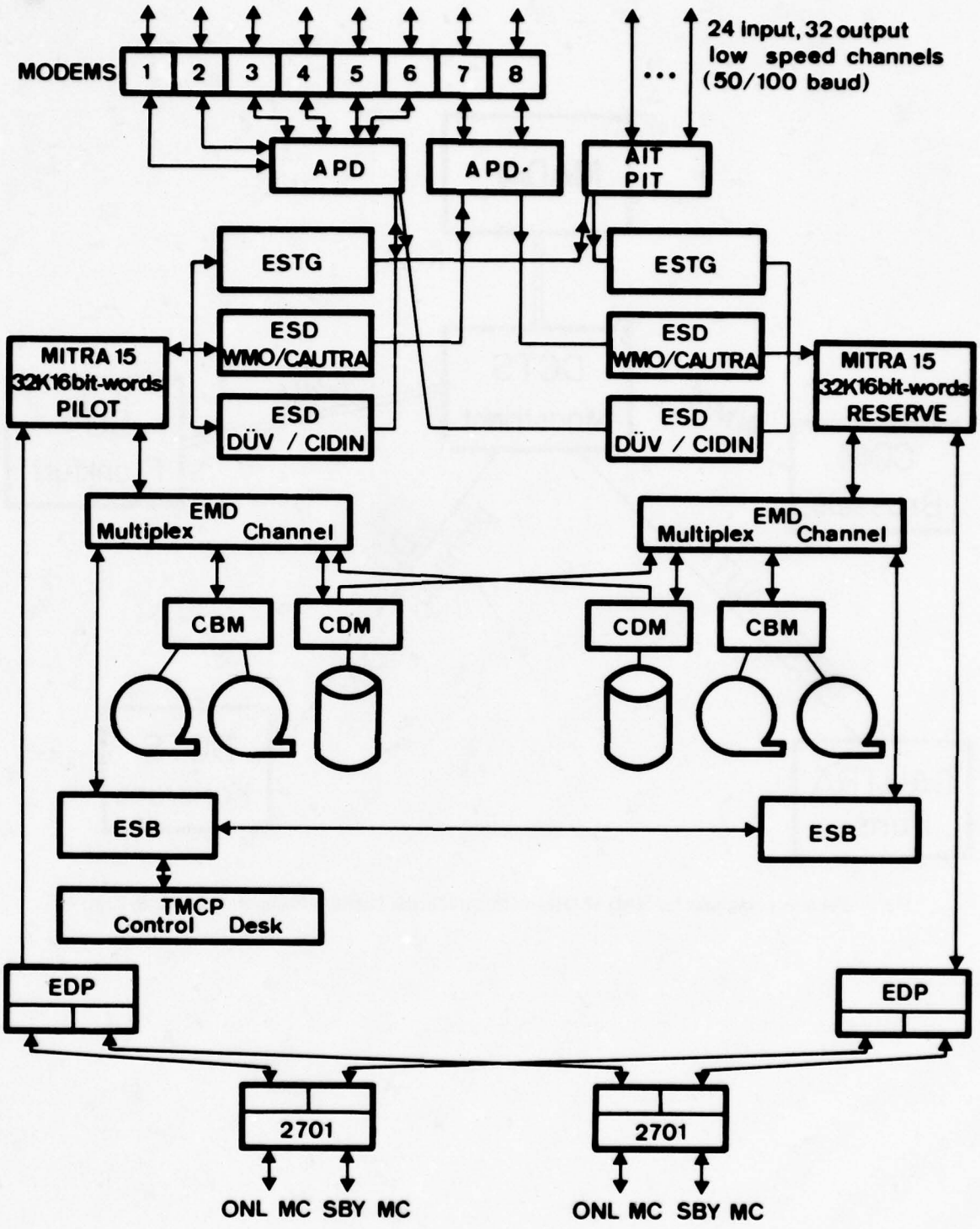
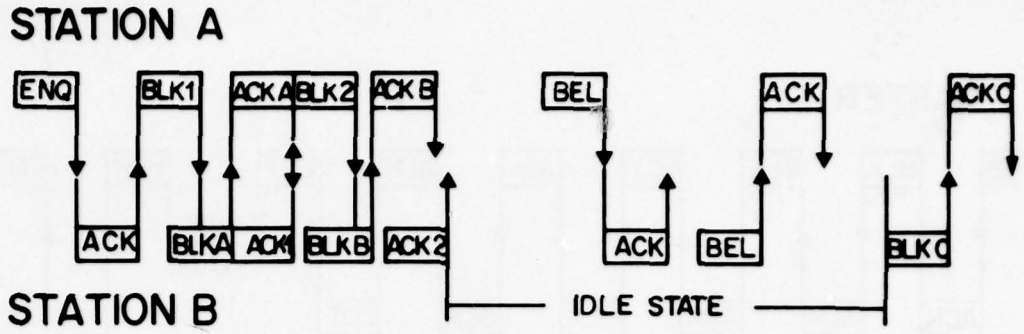


Fig.2 DCTS system configuration



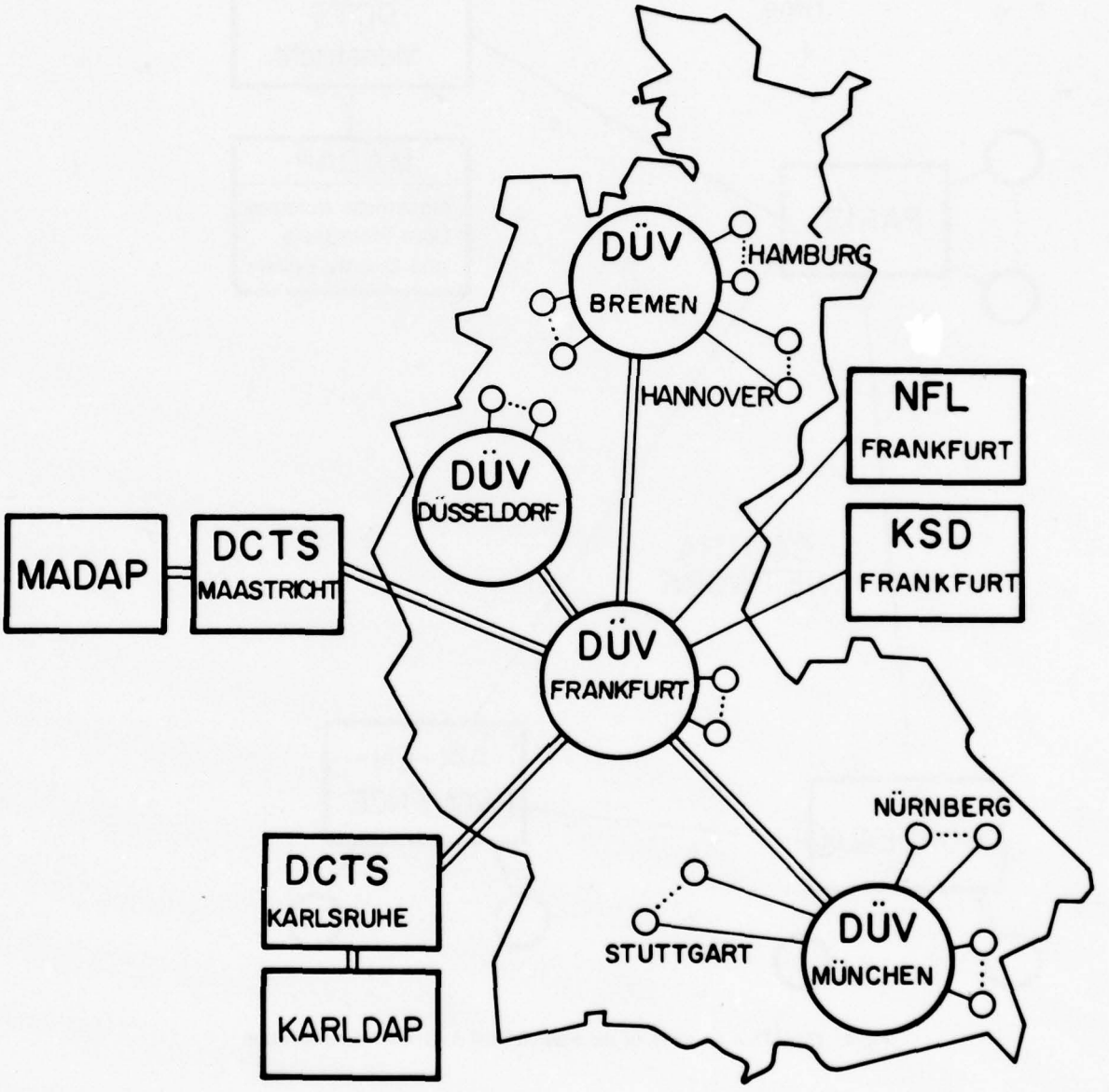


Fig.3 DÜV network

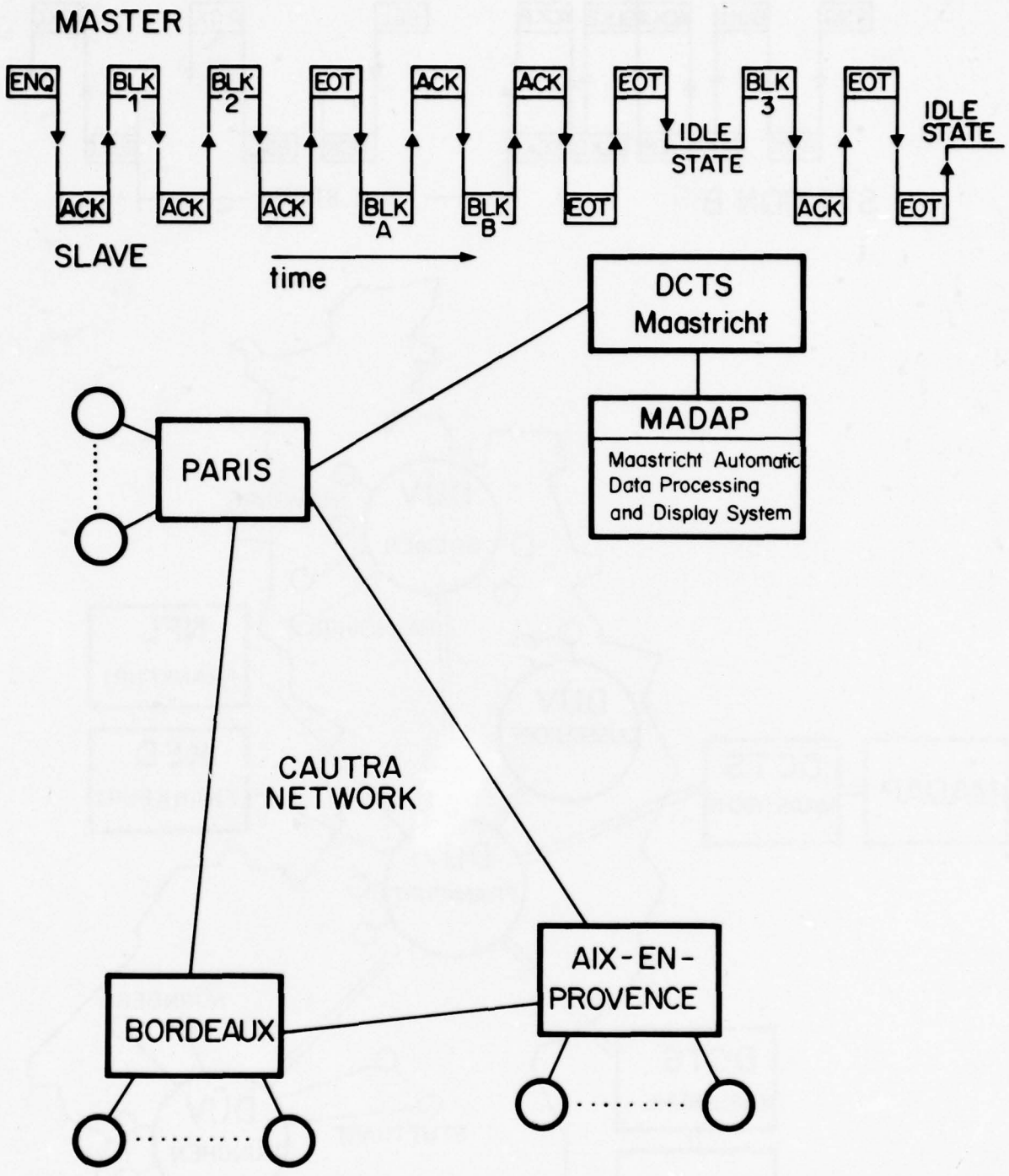


Fig.4 CAUTRA network of the French Civil Aviation Administration

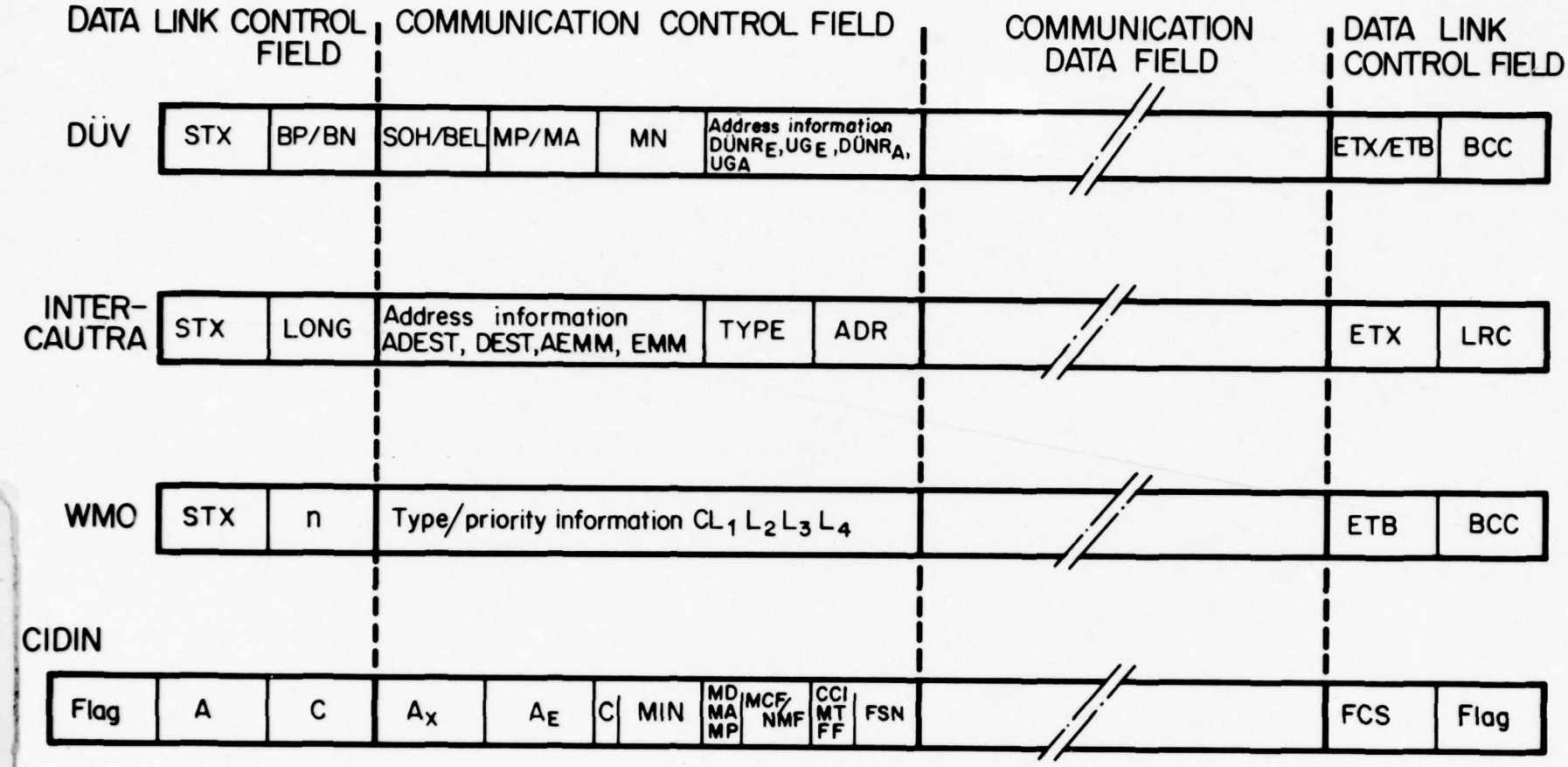


Fig.6 Comparison of procedure envelopes

PRECEDING PAGE BLANK-NOT FILMED

PRECEDING PAGE BLANK-NOT FILMED

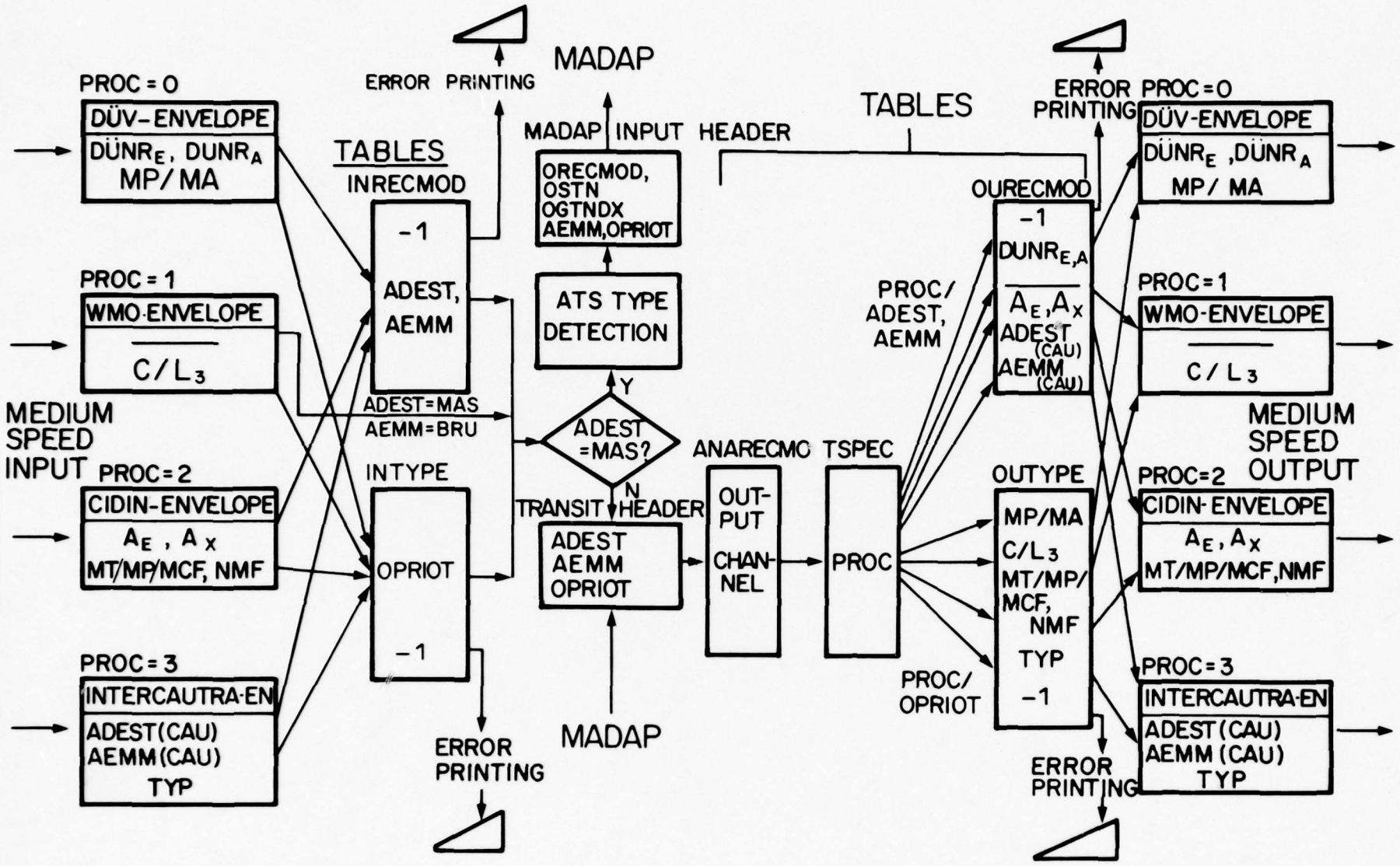


Fig.8 Procedure conversion by transit tables

DISCUSSION

I.L.Lebow, US

Has anything like this work been done elsewhere in the Air Traffic Control Environment?

Author's Reply

Similar work has been done in other areas of data communication for the interconnection of packet-switching networks. In the Air Traffic Control Environment this type of development has, according to my knowledge, not yet been tackled by other bodies to overcome the present heterogeneous data communication scenery.

It will certainly be a very important means in order to establish international links between European Air Traffic Control Centers in the next years. Flexible procedure handlers and converters should be developed and implemented also in the long-term development of a Common ICAO Data Interchange Network for interfacing with local networks.

PRECEDING PAGE BLANK-NOT FILMED

2. THE MARKOV CHAIN MODEL

Let us consider the channel model represented in fig.2.1. The source emits a sequence (b_n) of independent, identically distributed discrete random variables that is fed into the coder. We assume that the coder is a finite-state machine, so that, under our hypotheses on the statistics of (b_n), its output (c_n) is a sequence of random variables that can be modeled as a homogeneous Markov chain (TAKACS, 1960). The modulator operates by associating to each coded symbol c_n a waveform $c_n h(t-nT)$, where $h(t)$ has a finite duration T, T^{-1} being the rate at which coded symbols are fed into the modulator.

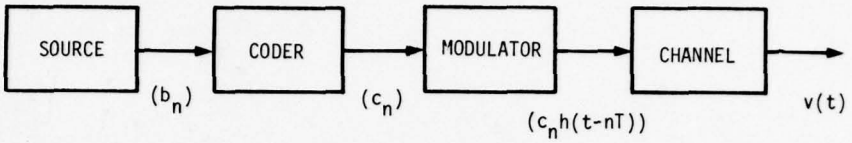


fig.2.1

We also assume that the memory of the channel is finite, i.e., at any given time instant t the channel output depends only on a finite number, say J, of symbols c_n. Thus, we can say that the channel output $v(t)$ is a function of the type

$$v(t) = V(t, c_{\ell_1(t)}, \ldots, c_{\ell_J(t)}) \tag{2.1}$$

where $\ell_1(t), \ldots, \ell_J(t)$ are integers depending on the value of t.

Furthermore, under our assumptions we can say that, if we observe $v(t)$ for T seconds, this waveform will take on only a finite number of possible shapes. In fact, as t ranges into a finite interval of duration T, the integers $\ell_1(t), \ldots, \ell_J(t)$ will take on different values, but still in a finite range.

In general, if M denotes the number of values taken by the random variables c_n, and L is the number of c_n's on which $v(t)$ depends as t runs in an interval of duration T, the number M' of different waveforms observed at the output of the channel is bounded above by

$$M' \leqslant M^L$$

where the inequality accounts for the situation in which the sequence (c_n) is coded, and hence not every L-tuple of symbols may be allowed.

In conclusion, at the output of the channel, and assuming for the moment that there is no noise, we can write the following signal

$$v(t) = \sum_{n=-\infty}^{\infty} q(t-nT; \xi_n) \tag{2.2}$$

where $q(t)$ is a waveform with duration T, and (ξ_n) is a sequence of random variables taking values in the set $\{1,2, \ldots, M'\}$.

Example 1

An important real-world example of nonlinear channels arises from digital satellite communication systems, as mentioned in the Introduction. A useful tool for analyzing a communication situation in which a digital signal is sent over a satellite channel is offered by Volterra series theory (BIGLIERI,1977 ; BENEDETTO,BIGLIERI and DAFFARA,1978).

The model represented in fig.2.2 is usually assumed for digital satellite channels.

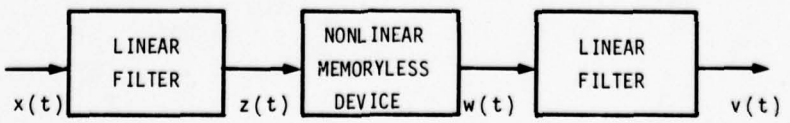


fig.2.2

PRECEDING PAGE BLANK-NOT FILMED

Here, a nonlinear memoryless part, representing the on-board amplifier, is preceded and followed by two bandpass linear systems. The first one represents the cascade of earth station transmitting filter and satellite input filter; the other one represents the cascade of satellite output filter and earth station receiving filter. Assume that $\tilde{h}'(\cdot)$, $\tilde{h}''(\cdot)$ are the equivalent low-pass impulse responses of the linear filters, and that the nonlinear device has an input-output relationship

$$w(t) = \sum_{n=1}^{\infty} \gamma_n \frac{z^n(t)}{n!}$$

Then (BENEDETTO,BIGLIERI and DAFFARA,1978), if the input signal has a complex envelope

$$x(t) = \sum_{n=-\infty}^{\infty} c_n \delta(t-nT)$$

the complex envelope of the first spectral zone of the channel output signal is given by

$$(2.3) \qquad v(t) = \sum_{k=0}^{\infty} L_k \sum_{n_1} \cdots \sum_{n_{2k+1}} c_{n_1} \cdots c_{n_{k+1}} c^*_{n_{k+2}} \cdots c^*_{n_{2k+1}} \cdot$$

$$\cdot \tilde{h}_{2k+1}(t-n_1T, \ \ldots \ , t-n_{2k+1}T)$$

where

$$\tilde{h}_{2k+1}(\tau_1, \ \ldots \ , \tau_{2k+1}) = \frac{\gamma_{2k+1}}{(2k+1)!} \int_{-\infty}^{\infty} \tilde{h}''(\tau) \prod_{r=1}^{k} \tilde{h}'^*(\tau_r - \tau) \prod_{s=k+1}^{2k+1} \tilde{h}'(\tau_s - \tau) \, d\tau$$

and

$$L_k = \binom{2k+1}{k} 2^{-(2k+1)}$$

The assumption that the channel has a finite memory is equivalent to saying that the functions $\tilde{h}_{2k+1}(\ldots)$ are zero when their arguments are outside a certain finite range. This implies that only a finite number of c_n's is involved in the summations (2.3). Thus, eq.(2.3) turns out to take the form (2.1). ■

We are now able to characterize the communication situation described at the beginning of this Section, provided that we are able to determine the statistics of the discrete-time random process (ξ_n).

To do that, let us assume for notational simplicity that, as t runs in the interval $[kT, \ (k+1)T]$, v(t) depends on the following L coded symbols:

$$(2.4) \qquad c_k \ , \ c_{k-1} \ , \ \cdots \ , \ c_{k-L+1}$$

Therefore, ξ_k will be a function of the same random variables. We can also assume, without loss of generality, that ξ_k is a one-to-one function of these random variables.

Consider now the time interval $[(k+1)T, \ (k+2)T]$. Here v(t), and hence ξ_{k+1} , will depend on

$$(2.5) \qquad c_{k+1} \ , \ c_k \ , \ \cdots \ , \ c_{k-L+2}$$

and the following conclusions can be drawn.

Consider $Pr\{\xi_{k+1} | \xi_k \ , \ \xi_{k-1}, \ \ldots\}$: ξ_{k+1} is a one-to-one function of the random variables (2.5), whereas ξ_k , ξ_{k-1}, ... (i.e., the past of ξ_{k+1}) is a one-to-one function of c_k, c_{k-1} , ...But (c_k) is a Markov chain, so the values taken on by c_{k+1} , ... , c_{k-L+2}

will depend only on $c_k, \cdots, c_{k-L+1}$.

This means that the values taken on by ξ_{k+1} depend only on the value of ξ_k, and not on those of $\xi_{k-1}, \xi_{k-2}, \cdots$. In other words, the sequence ξ_k forms another Markov chain.

Let us now compute the probability transition matrix of this chain. Assume that ξ_k takes the value i when $c_k = i_o$, $c_{k-1} = i_{-1}, \cdots, c_{k-L+1} = i_{-L+1}$, and ξ_{k+1} takes value j when $c_{k+1} = j_1, c_k = j_o, \cdots, c_{k-L+2} = j_{-L+2}$. We get

$$\Pr\{\xi_{k+1} = j | \xi_k = i\} = \Pr\{c_{k+1}=j_1,\ldots,c_{k-L+2}=j_{-L+2} | c_k=i_o,\ldots,c_{k-L+1}=i_{-L+1}\}$$

$$= \begin{cases} \Pr\{c_{k+1} = j_1 | \ c_k = i_o\} & \text{if } j_\ell = i_{\ell+1}, \ -L+2<\ell<0 \\ 0 & \text{elsewhere} \end{cases} \tag{2.6}$$

Eq. (2.6) also shows that the Markov chain (ξ_n) is homogeneous.

It must be observed that the structure of the chain (ξ_n) does not depend on the actual behavior of the channel, but only on its memory length L. the channel behavior is reflected by the shapes of the waveforms $q(t;k)$, $k=1,\ldots,M'$.

Example 2

Assume L=3, and a binary, uncoded signal entering the channel. Since the random variables c_n turn out to be independent and equally likely, the process (c_n) is described by the trivial transition matrix

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

The transition probability matrix of (ξ_n) is 8x8. Labeling its rows and columns by the corresponding values of the triplets c_k, c_{k+1}, c_{k+2} we get

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{matrix} 000 \\ 001 \\ 010 \\ 011 \\ 100 \\ 101 \\ 110 \\ 111 \end{matrix} \tag{2.7}$$

$$000 \ \ 001 \ \ 010 \ 011 \ \ 100 \ \ 101 \ \ 110 \ \ 111$$

Example 3

For a more sophisticated -- and perhaps more enlightening -- example, consider binary sorce data being encoded by the convolutional coder depicted in fig. 2.3 (the resulting code has constraint length 2 and rate 1/2).

This coder can be modeled as a finite-state machine with 8 states (each one corresponding to a content of the three-stage binary shift register of the encoder) . To each state it corresponds a pair of binary symbols : for example, if the register content is 110, the output is the codeword 01. Since the memory length of this machine is 3, the transition probability matrix of the sequence of coded binary pairs is the same as given in

(2.7). It must be noticed here that there are only 4 distinct codewords (namely, 00, 01, 10 and 11), but we have to take 8 distinct states into account.

Assume now that the channel has memory L=2. The resulting output can be modeled by a 64-state Markov chain, whose transition probability matrix is computed according to the rule (2.6). ■

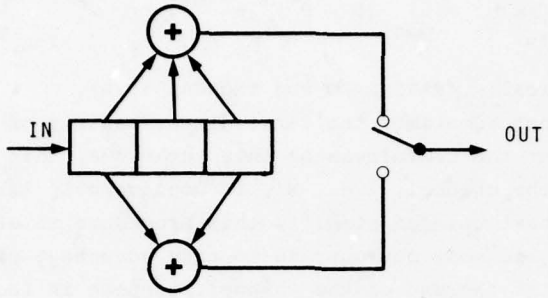


fig.2.3

3. COMPUTATION OF THE POWER SPECTRUM

Let us return to (2.2). From that equation we can see that -- to specify the output of the channel -- we have to give a set of waveforms $q(t;k)$, $k=1,\ldots,M'$, together with the statistics of the Markov chain (ξ_n). From now on we shall make the further assumption that the Markov chain is regular (GANTMACHER,1960), so that its transition probability matrix will suffice for its description.

Consider now the computation of the power spectrum of the signal $v(t)$. Such a computation is often a relevant task in assessing the performance of a digital communication scheme: in fact, in many systems data streams from users are assigned adjacent frequency bands, that interfere with each others in a larger or lesser extent depending on the bandwidth occupancy of the modulated signals. Thus, spectrum occupancy is a measure of adjacent channel interference.

Assume now that the waveforms $q(t;k)$, $k=1,\ldots,M'$, and the transition probability matrix $\underset{\sim}{P}$ of the chain (ξ_n) have been specified. Then(AJMONE MARSAN and BIGLIERI,1977) the power spectrum of $v(t)$ can be written as

$$\mathcal{G}(\omega) = \mathcal{G}_c(\omega) + \mathcal{G}_d(\omega)$$

where $\mathcal{G}_c(\omega)$ is the continuous part of $\mathcal{G}(\omega)$, and $\mathcal{G}_d(\omega)$ is the line spectrum; these two functions are given by

$$\mathcal{G}_c(\omega) = \frac{2}{T} \mathrm{Re}\{\underset{\sim}{Q}^t(\omega)\underset{\sim}{\Pi}\underset{\sim}{\Lambda}(\omega)\underset{\sim}{Q}(\omega)\} - \frac{1}{T}\underset{\sim}{Q}^t(\omega)\underset{\sim}{\Pi}(\underset{\sim}{I} + \underset{\sim}{P}^{\infty})\underset{\sim}{Q}(\omega)$$

$$\mathcal{G}_d(\omega) = \frac{2}{T^2}\underset{\sim}{Q}^t(\omega)\underset{\sim}{\Pi}\underset{\sim}{P}^{\infty}\underset{\sim}{Q}(\omega)\sum_{m=-\infty}^{\infty}\delta(\omega - m\frac{2\pi}{T})$$

where $(\,)^t$ denotes conjugate transpose, $\underset{\sim}{I}$ is a M'xM' identity matrix, $\underset{\sim}{Q}(\omega)$ is the column M'-vector whose k-th entry is the Fourier transform of $q(t;k)$, $\underset{\sim}{\Pi}$ is a diagonal matrix with entries

$$(\underset{\sim}{\Pi})_{ij} = \begin{cases} \Pr\{\xi_n = i\} & i = j \\ 0 & i \neq j \end{cases}$$

$\underset{\sim}{P}^{\infty}$ is the limiting matrix

$$\underset{\sim}{P}^{\infty} = \lim_{n\to\infty} \underset{\sim}{P}^{n}$$

and $\underset{\sim}{\Lambda}(\omega)$ is a matrix series

$$\underset{\sim}{\Lambda}(\omega) = \sum_{m=0}^{\infty} (\underset{\sim}{P} - \underset{\sim}{P}^{\infty})^{m} e^{-jm\omega T}$$

which can be evaluated numerically (CARIOLARO and TRONCA, 1974).

It must be observed that sometimes the large dimensionality of the transition probability matrix $\underset{\sim}{P}$ can hinder the usefulness of this technique. This occurs when the memory of the coder and/or of the channel, i.e., M', is considerably large.

In any case, several shortcuts can simplify this procedure in order to improve its practicalness. The simplest, and most obvious, is to take advantage of any linear system that can possibly be present at the end of the channel. Suppose in fact that the nonlinear channel can be split into the cascade of a nonlinear part and a linear system with transfer function $H(\omega)$. This is the case of many practical communication systems, where the demodulator is preceded by the receiver filter (see also Example 1). Under these conditions, we can compute the power spectrum before this linear system, and the multiply it by $|H(\omega)|^2$.

Other shortcuts take advantage of the special structure of the matrix $\underset{\sim}{P}$ for the computation of $\underset{\sim}{\Lambda}(\omega)$ and $\underset{\sim}{P}^{\infty}$, but we shall not pursue this topic any further here. The interested reader is referred to the forthcoming paper (AJMONE MARSAN, BIGLIERI and ELIA, 1978).

4. OPTIMUM (MAXIMUM LIKELIHOOD SEQUENCE) RECEIVER

We shall now show how the Markov chain model derived in Section 2 allows the derivation of the structure of the maximum likelihood sequence receiver for a situation in which the channel consists of a known, finite-memory part followed by a noisy memoryless part. In particular, we assume that the received signal is the sum of a signal like (2.2) plus white Gaussian noise. This situation has been considered in a general, abstract setting by (OMURA,1971) and for the specific case of a bandpass nonlinear channel by (MESIYA et al., 1977).

Since the channel has a finite memory, the signal received at time t can be written as

$$z(t) = \sum_{n=n_1(t)}^{n_2(t)} q(t-nT;\xi_n) + \nu(t) \tag{4.1}$$

where $n_1(t)$ and $n_2(t)$, $n_2(t) \geq n_1(t)$, are integers depending upon the actual value of t, and $\nu(t)$ is white Gaussian noise.

Assume that z(t) is observed over the time interval $0<t<KT$. Denoting by N_1 and N_2 the following integers

$$N_1 = \min_{0\leq t\leq KT} n_1(t)$$

$$N_2 = \max_{0\leq t\leq KT} n_2(t) \tag{4.2}$$

we see that the observation will depend on the values taken by the random variables $\xi_{N_1}, \ldots, \xi_{N_2}$. This sequence of random variables can take on one of

$$I = (M')^{N_2-N_1+1} \tag{4.3}$$

possible states, and to each state it corresponds a received waveform like

$$v_{\underset{\sim}{m}}(t) = \sum_{n=N_1}^{N_2} q(t-nT;m_n) \qquad 0<t<KT \tag{4.4}$$

where $\underset{\sim}{m} = (m_{N_1} , \dots , m_{N_2})$ is an integer sequence denoting a possible state taken by the sequence of random variables $\xi_{N_1} , \dots , \xi_{N_2}$.

Compute now the log-likelihood ratio for $\underset{\sim}{m}$. We get

$$\Lambda_{\underset{\sim}{m}} = \frac{2}{N_o} \int_0^{KT} v_{\underset{\sim}{m}}(t) z(t) dt - \frac{1}{N_o} \int_0^{KT} v^2_{\underset{\sim}{m}}(t)\, dt \ . \tag{4.5}$$

Using (4.4), it follows that

$$\Lambda_{\underset{\sim}{m}} = \frac{2}{N_o} \sum_{n=N_1}^{N_2} \int_0^{KT} q(t-nT;m_n) z(t) dt - \frac{1}{N_o} \sum_{\ell=N_1}^{N_2} \sum_{n=N_1}^{N_2} \int_0^{KT} q(t-\ell T;m_\ell) q(t-nT;m_n) dt \tag{4.6}$$

Recall now that, under our hypotheses, $q(t;.)$ has a finite duration T. Assuming that K is large enough so that we can disregard end effects, we have

$$\int_0^{KT} q(t-nT;m_n) z(t) dt = \int_{nT}^{(n+1)T} q(t-nT;m_n) z(t) dt \tag{4.7}$$

Similarly, we can observe that

$$\int_0^{KT} q(t-\ell T;m_\ell) q(t-nT;m_n) dt = \begin{cases} 0 & \ell \neq n \\ \int_0^T q^2(t;m_n) dt & \ell = n \end{cases} \tag{4.8}$$

Thus, defining

$$\alpha_n(m_n) = \int_{nT}^{(n+1)T} q(t-nT;m_n) z(t) dt \tag{4.9}$$

and

$$\mathbf{E}(m_n) = \int_0^T q^2(t;m_n) dt \tag{4.10}$$

we finally get

$$\begin{aligned} \Lambda_{\underset{\sim}{m}} &= \frac{2}{N_o} \sum_{n=N_1}^{N_2} \alpha_n(m_n) - \frac{1}{N_o} \sum_{n=N_1}^{N_2} \mathbf{E}(m_n) \\ &= \frac{1}{N_o} \sum_{n=N_1}^{N_2} \{ 2\alpha_n(m_n) - \mathbf{E}(m_n) \} \ . \end{aligned} \tag{4.11}$$

We can observe that:

(i) $\alpha_n(m_n)$ can be obtained by sampling at time (n+1)T the output of a filter, matched to $q(t;m_n)$, to the input z(t), $nT<t<(n+1)T$.

(ii) $\mathbf{E}(m_n)$ is the energy of the waveform $q(t;m_n)$.

The maximum likelihood decoding rule requires $\Lambda_{\underset{\sim}{m}}$ to be maximized over the set of possible sequences $\underset{\sim}{m}$. With obvious notations, $\Lambda_{\underset{\sim}{m}}$ can be rewritten in the form

$$\Lambda_{\underset{\sim}{m}} = \sum_{n=N_1}^{N_2} \lambda_n(m_n) \tag{4.12}$$

from which it is seen that $\Lambda_{\underset{\sim}{m}}$ is a function of the vector $\underset{\sim}{m}$ through a sum of functions of its components.

This last observation, together with the Markov chain structure of the process (ξ_n), forms the basis for the application of Viterbi algorithm to teh solution of this demodulation problem (actually, the maximum likelihood sequence receiver simultaneously performs the function of demodulation and decoding). In fact, the assumption that the process (ξ_n) is a Markov chain means in particular that the set of values that ξ_n is allowed to take on depends upon the future values ξ_{n+1} , ξ_{n+2} , ... only through ξ_{n+1}. Since we have denoted these values by m_n, this is equivalent to saying that the allowable values for m_n depend on the values of the other components m_{n+1} , m_{n+2} , ... of $\underset{\sim}{m}$ only through m_{n+1}. This statement is crucial in order to allow the Viterbi algorithm to be applied for the solution of this maximization problem. To see why this is true, consider for notational simplicity the case $N_1=1$, $N_2=N$. Then the demodulation problem is equivalent to finding

$$\mu = \max_{m_1,\ldots,m_N} \sum_{n=1}^{N} \lambda_n(m_n)$$

i.e., maximizing a function of N arguments made up of the sum of N functions, each of them dependent on only one of the arguments. Denote by

$$m_i \rightarrow m_{i+1}$$

the set of values that m_i is allowed to take under the constraint that the following component of $\underset{\sim}{m}$ takes value m_{i+1} . Then the maximization (Viterbi) algorithm is

$$\mu_2(m_2) = \max_{m_1 \rightarrow m_2} \lambda_1(m_1)$$

$$\mu_3(m_3) = \max_{m_2 \rightarrow m_3} \{\lambda_2(m_2) + \mu_2(m_2)\}$$

$$\vdots$$

$$\mu_N(m_N) = \max_{m_{N-1}\ m_N} \{\lambda_{N-1}(m_{N-1}) + \mu_{N-1}(m_{N-1})\}$$

$$\mu = \max_{m_N} \mu_N(m_N).$$

Notice that sometimes further simplifications can occur. For example, if $\mu_i(m_i)$ does not depend on m_i, i.e., all the values of m_i give rise to the same value for $\mu_i(m_i)$, the following iterations can be simplified by taking advantage of the fact that μ_i is now a constant.

The performance of such an optimum receiver can also be evaluated: upper bounds to the bit error probability can be computed(MESIYA et al.,1977 and VITERBI and OMURA, 1978) depending on the set of distances

$$d^2(\underset{\sim}{m},\underset{\sim}{m}') = \int_0^{KT} |v_{\underset{\sim}{m}}(t) - v_{\underset{\sim}{m}'}(t)|^2 dt \ .$$

REFERENCES

M.AJMONE MARSAN and E.BIGLIERI, 1977, "Power spectra of complex PSK for satellite communications", Alta Frequenza, Vol.XLVI,n.6,pp.263-270

M.AJMONE MARSAN,E.BIGLIERI and M.ELIA, 1978, "Power spectra of digital signals after non linearities with memory", to be published

S.BENEDETTO,E.BIGLIERI and R.DAFFARA, 1978, "Modeling and performance evaluation of nonlinear satellite links - A Volterra series approach", to be published

E.BIGLIERI, 1977, "Digital transmission over nonlinear channels with memory - A Volterra series analysis", NATO Advanced Study Institute on Communication Systems & Random Process Theory, Darlington,U.K.

G.L.CARIOLARO and P.TRONCA, 1974, "Spectra of block coded digital signals", IEEE Trans. on Commun., Vol.COM-22, n.10, pp.1555-1564

F.R.GANTMACHER, 1960, Matrix Theory,vol. II. Chelsea, New York

M.F.MESIYA,P.J.McLANE and L.L.CAMPBELL, 1977, "Maximum likelihood sequence estimation of binary sequences transmitted over bandlimited nonlinear channels", IEEE Trans. on Commun., Vol.COM-25, n.1, pp.12-22

J.K.OMURA, 1971, "Optimal receiver design for convolutional codes and channels with memory via control theoretical concepts", Inform. Sciences, Vol.3, pp. 243-266

L.TAKACS, 1960, Stochastic Processes. J.Wiley & Sons, New York

A.J.VITERBI and J.K.OMURA, 1978, Digital Communication and Coding. McGraw-Hill,New York

STATE OF THE ART OF ERROR CONTROL TECHNIQUES

Jack Keil Wolf
Department of Electrical and Computer Engineering
University of Massachusetts
Amherst, Massachusetts 01003
USA

SUMMARY

A survey of error control techniques for achieving reliable transmission over noisy communication channels is presented. Both binary and nonbinary codes are considered. Block codes and tree codes are described along with their decoding algorithms. The parameters of the most frequently utilized codes are given. Finally, the performance of such codes are considered for an additive Gaussian noise channel with and without Rayleigh fading.

1. INTRODUCTION

Assume that you as an author of a paper at this symposium have received the following telegram from the Program Chairman: PRLGRAM FOB ASARD AVP SYRPOSIAM CHLNGED YSURLPAPRR NQW SCHEDULAD FOR JONE 7 AT 13,45. After some effort at error detection and correction, you could probably correctly interpret the non-numerical portion of the text since words in the English language are redundant. That is, not every combination of the 27 symbols (26 letters plus the space symbol) forms an acceptable message. For example the sequence of letters "PRLGTAM" is not an English word and thus errors have been detected in this sequence of symbols. Since it differs from the word "PROGRAM" in only one letter and differs from other words in more than one letter it is "closest" to the word "program". Thus in decoding this word to the word "PROGRAM" we have accomplished error correction.

The errors in the numerical portion of the telegram (the date and the time) present a different problem. In general, numbers do not possess the redundancy of non-numerical text. Thus the "7" could have been in error and we could not detect or correct this error from the natural redundancy of the message.

In designing a system for the reliable transmission of data over a noisy communication channel, one cannot rely on the natural redundancy of the message to detect and correct errors since the system must work for all types of messages (even those without natural redundancy such as certain types of computer data). Thus, we must introduce an artificial redundancy into the messages in order to effect error control. This artificial redundance, called coding for error control, is the subject of this paper.

Codes for error control come in two distinct flavors: block codes and tree (commonly called convolutional) codes. The next two sections are concerned with the definitions and important characteristics of these two classes of codes.

2. BLOCK CODES (WOLF. J. K., 1973)

A block code of length n and size M is a collection of M distinct vectors called codewords, each vector having n components belonging to some finite alphabet $X = \{0,1,2,\ldots,q-1\}$. The rate of the code, R, is defined as

$$R = \frac{\log_q M}{n}.$$

Since the codewords are distinct, $1 \leq M \leq q^n$ and $0 \leq R \leq 1$. For binary codes, $q = 2$, while for nonbinary codes $q > 2$. Usually q is chosen equal to a prime or a power of prime.

The Hamming weight of a codeword is equal to the number of nonzero components in that vector. The minimum weight of a code is the positive integer equal to the smallest nonzero Hamming weight of a codeword in the code.

We assume henceforth that the elements of X form a finite field GF(q) (so that q is equal to a prime or a power of a prime). The code is linear if the codewords are all the solutions to a set of r homogeneous linear equations, called generalized parity-check equations. The coefficients of these equations are elements from X. Let $k = n - r$. If the equations are linearly independent, $M = q^k$, $R = k/n$, and the code is termed an (n,k) code. A code which is not linear is said to be nonlinear.

The Hamming distance between two n-vectors is equal to the number of components in which these vectors differ. For a linear code, the number of codewords of Hamming distance i, $i = 0,1,2,\ldots,n$, from any given codeword is equal to the number of codewords of weight i. The minimum Hamming distance between a pair of distinct codewords in a code, d_{min}, (or the minimum weight of a linear code) yields important information regarding capability of the code to a correct and detect random errors. A code can correct all patterns of t or fewer random errors and in addition detect all patterns having no more than d errors (where $d \geq t$) provided that

$$d + t + 1 \leq d_{min}.$$

If the code is used for error correction only then $d = t$ and the code can correct all patterns of t or fewer random errors provided that

$$2t + 1 \leq d_{min}.$$

2.1. Example of a Binary Block Code

To illustrate the ideas introduced in the previous section we consider the following simple example of a binary ($q = 2$), (7,3) code. Such a code has $M = 2^3 = 8$ code words, each of block length 7. If $\underline{x} = (x_1,$

$x_2, x_3, x_4, x_5, x_6, x_7$) represents a code word in the code, and if these symbols satisfy the following set of linear (parity check) equations (+ means modulo 2 sum)

$$x_1 + x_2 + x_3 = x_4$$
$$x_1 + x_3 = x_5$$
$$x_2 + x_3 = x_6$$
$$x_1 + x_2 = x_7,$$

then the 8 code words are:

0 0 0 0 0 0 0
1 0 0 1 1 0 1
0 1 0 1 0 1 1
0 0 1 1 1 1 0
1 1 0 0 1 1 0
0 1 1 0 1 0 1
1 0 1 0 0 1 1
1 1 1 1 0 0 0 .

The minimum distance of the code is 4 which is the minimum weight of any nonzero code word. (In this special case all nonzero code words have the same weight but this is not usually the case.) Thus the code can correct a single error while detecting but not correcting a double error.

Note that in this case, the first 3 digits in any code word can be considered as the message digits while the last 4 digits which are calculated from the first 3 are the redundant digits or parity digits.

2.2. Some Important Classes of Block Codes (PETERSON, W. W., E. J. Weldon, Jr., 1972)

The following is a brief summary of the characteristics of some important classes of block codes:

2.2.1. Binary Hamming Codes (q = 2)

Let m be any positive integer ≥ 2. Then for each m there is a linear code with parameters

$$n = \text{block length} = 2^m - 1,$$
$$k = \text{message digits} = 2^m - 1 - m,$$
$$n - k = \text{check digits} = m.$$

These codes all have minimum distance equal to 3 and thus can correct any single error in the block of length n digits.

2.2.2. Bose-Chaudhuri-Hocquenhem (BCH) Codes

These are linear codes with coefficients from any field GF(q). Let m be any positive integer ≥ 1, let c be any integer which divides $q^m - 1$ and let t be any positive integer. Then the code has parameters:

$$n = \text{block length} = (q^m - 1)/c,$$
$$n = k = \text{check symbols} \leq \begin{cases} 2mt & q \neq 2 \text{ (nonbinary codes)} \\ mt & q = 2 \text{ (binary codes)}, \end{cases}$$
$$d_{min} = \text{minimum distance} \geq 2t + 1.$$

2.2.3. Reed-Solomon (R S) Codes

These are a special case of nonbinary BCH codes formed by choosing m = c = 1. These codes have parameters:

$$n = \text{block length} = q - 1 \quad ,$$
$$n - k = \text{check symbols} = 2t = d_{min} - 1.$$

2.2.4. Simplex Codes

These are a special case of binary BCH codes formed by choosing c = 1. These codes have parameters:

$$n = \text{block length} = 2^m - 1,$$
$$k = \text{message digits} = m \quad ,$$

d_{min} = minimum distance = 2^{m-1}.

2.2.5 Golay Code

This is a special binary code that has a very high error correction capability for the amount of redundancy utilized. It is also a special case of the BCH codes. It has parameters:

n = block length = 23 ,

k = message digits = 12,

d_{min} = minimum diatance = 7.

The code thus can correct 3 random errors in a block of 23 digits. The code is often used as a (24,12) code by adding an extra parity digit which is a parity check over all digits in the block. The resultant (24,12) code then has minimum distance equal to 8.

2.2.6. Majority Logic Decodable Codes

These are a class of codes that because of the special form of their parity check equations lead to a particularly simple decoding algorithm. (See the next section for a further discussion.)

2.3. Decoding of Block Codes (WOLF, J. K., 1973)

If the received word were always an exact replica of the transmitted word when a codeword is transmitted over a communications channel, there would be no need for coding. Rather, a noisy communications channel distorts the transmitted codewords in a stochastic manner. A channel with input n-vectors from $(X)^n$ (the space of sequences of n symbols from the input alphabet X) and output n-vectors from $(Y)^n$ (where Y is the output alphabet) can be described by a conditional probability distribution $P_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x})$ for all $\underline{x} \in (X)^n$ and $\underline{y} \in (Y)^n$. Here $\underline{X}$ and $\underline{Y}$ are random n-vectors representing the input and output n-vectors for the channel, and $\underline{x}$ and $\underline{y}$ are the specific values which can be assumed by these vectors.

A decoder is a device that instruments a decoding rule for choosing among the transmitted code words on the basis of the received vector $\underline{y}$. A possible option, termed error detection, is to choose no codeword at all if the received sequence is not a code word. This option is often utilized when the codeword can be retransmitted or reread from memory. A particular decoding rule which always decodes to a codeword is the one that chooses the codeword having the highest conditional probability of being transmitted, given the received vector $\underline{y}$. If all codewords have equal probability a priori, then this rule, called a maximum-likelihood decoding rule, chooses the codeword $\underline{c}_i$ for which $P_{\underline{Y}|\underline{X}}(\underline{y}|\underline{c}_i)$ is the largest. A brute-force application of this rule requires M calculations of the conditional probability distribution. For a binary code of block length n = 100 and rate $R = \frac{1}{2}$, this works out to $2^{50} \approx 10^{15}$ calculations—a hopeless task even with a large computer. It is the algebraic structure of the codes that allows us to escape form this dilemma.

Most decoding rules for algebraic block codes do not realize a maximum-likelihood decoding rule. Rather, they decode to the most likely codeword only if the noise on the channel is not too large. Otherwise they utilize the option of not decoding. Such a rule is called a bounded-distance decoding rule.

The Berlekamp algorithm for decoding BCH codes (PETERSON, W. W. and WELDON, E. J., 1972) is a bounded-distance decoding rule that requires that X = Y and that will decode correctly if and only if the Hamming distance between the received vector and the transmitted codeword does not exceed $(d_{min} - 1)/2$.

A class of codes that are not as powerful as BCH codes but that allow a simpler decoding algorithm are the majority-logic decodable codes. The generalized parity-check equations of these codes are based upon the combinatorial configurations of finite geometries. In the simplest case, decoding for these codes is performed on a symbol-by-symbol basis. For each symbol, several generalized parity-check equations are checked, each equation predicting that the symbol be a particular element of GF(q). The field element receiving the most votes is taken to be the correct value for that symbol. It has been shown that any decoding rule for any code can be realized, in principle, by properly weighting the votes of generalized parity-check equations. (RUDOLPH, L. D., ROBBINS, W. E., 1972)

3. TREE CODES, TRELLIS CODES AND CONVOLUTIONAL CODES (WOLF, J. K., 1973)

Consider a tree as shown in Figure 1. The small circles are nodes, and the lines emanating from each node are branches. We assume that every node has Q branches emanating from it. Associated with each branch is a sequence of n_o symbols from the alphabet $\{0,1,2,\ldots,q = 1\}$. A tree code is the set of (possibly infinite) sequences obtained by concatenating the symbols on the branches of each unique path through the tree. Note that although there are an infinite number of codewords in our code, the first n_o symbols for every codeword can assume only Q different realizations. Note further that if we truncated the tree by allowing each path to contain only L branches, we would have a block code of block length $n = n_o L$ with $M = L^Q$ codewords. (Here the codewords may not all be distinct.) The rate of the tree code is defined as $R_o = (1/n_o) \log_q Q$.

We now introduce some structure in the tree. We assume that the tree is generated by a K-state machine with states $S_0, S_1, \ldots, S_{K-1}$. The machine has inputs from the set $\{0,1,\ldots,Q - 1\}$ and outputs from $(X)^{n_o}$.
We assume the machine always starts in state S_0. The machine is thought to reside in a state until an input is imposed. As a result of this input, the machine produces an output n_o-vector and assumes a next state. This change of states and production of outputs is described by a state-transition table that lists,

for every state and every input, the next state and the corresponding output.

To obtain a tree code from a K-state machine, associate a state with each node. The input then determines which of the Q branches to take from that node (state) to the next node (state). The n_o symbols on each branch are the outputs of the machine.

An example of a state-transition table for a four-state machine, with Q = 2, X = {0,1}, and n_o = 2, and its corresponding tree code is shown in Fig. 2(a) and (b). An input sequence and the corresponding codeword are given in Fig. 2(c). Note that there are only four states, so that several of the nodes in the tree can be collapsed into a single node. Upon collapsing these nodes, the tree forms a trellis, as shown in Fig. 2(d). Thus we say a finite-state machine generates a trellis code.

Consider a trellis code where X is the finite field GF(q) and $Q = q^{k_o}$, $1 \leq k_o \leq n_o$. Then each input can be considered a k_o-vector with components from X = GF(q). Let the n_o components of the outputs be a fixed linear function of the k_o components of the present input vector and the vk_o components of the v immediately preceding vectors. "Linear" here means a weighted sum of the components with respect to addition and multiplication as defined in GF(q). The number of states of the machine need never exceed q^{vk_o}. The resulting trellis code is said to be a convolutional code of constraint length v (or $k_o v$). The rate of the code is $R_o = (1/n_o) \log Q = k_o/n_o$.

A convolutional code is called systematic if k_o of the output symbols are equal to the current input k_o-vector. Otherwise the code is nonsystematic. Nonsystematic convolutional codes are superior to systematic convolutional codes for maximum-likelihood decoding on a random-error channel. This surprising result is related to the fact that every block code is equivalent to a systematic block code, but not every convolutional code is equivalent to a systematic convolutional code.

Two distance measures have been suggested for convolutional codes. The first, d_{min}, is the minimum nonzero Hamming distance between the first $(v + 1)n_o$ symbols of distinct codewords. The second, d_{free}, is the minimum nonzero Hamming distance between distinct infinite-length codewords. The free distance d_{free} seems to be more closely related to the performance of the code for the more powerful decoding algorithms.

Given a systematic convolutional code of minimum distance d_{min}, the first k_o message digits can be decoded correctly if t or fewer errors occurred in the first $(v + 1)$ n_o transmitted digits provided that

$$2t + 1 \leq d_{min}.$$

The relationship between d_{free} and the error correction capability of the code is more obtuse.

3.1. An Example of A Convolutional Code

The state-transition table and trellis of a convolutional code with parameters q = 2, k_o = 1, n_o = 2, v = 2 are given in Fig. 3(a) and (b). A realization of this finite-state machine in terms of a two-stage shift register is given in Fig. 3(c). The code has $d_{min} = d_{free} = 4$.

3.2. Some Convolutional Codes

Very little is known about constructing tree or trellis codes that are not convolutional codes. Thus in this section we restrict our attention to convolutional codes. Indeed, even for convolutional codes, there is a scarcity of techniques for constructing good codes.

3.2.1. Single Error Correcting Binary Codes (q = 2)

Let v be any positive integer. Then the code has parameters:

n_o = symbols per branch = 2^v,

k_o = message symbols per branch = $n_o - 1$,

Q = branches per node = $2^{k_o} = 2^{n_o - 1}$,

d_{min} = minimum distance = 3.

3.2.2. Double Error Correcting Binary Codes (q = 2)

This code is based upon a binary BCH code of minimum distance 6. For any positive integer m, it has parameters

n_o = symbols per branch = $2^m - 1$,

k_o = message symbols per branch = $2^m - 2 - 2m$,

v = constraint length = 1,

d_{min} = minimum distance = 6.

3.2.3. Self-Orthogonal Binary Codes (q = 2)

The construction of these codes is based upon difference triangles. They have parameters:

n_o = symbols per branch = any integer,

d_{min} = minimum distance = any integer,

k_o = message symbols per branch = $n_o - 1$

v = constraint length $\geq (n_o - 1)(d_{min} - 1)(d_{min} - 2)/2$.

3.2.4. Computer Generated Codes

Most good convolutional codes have been found by computer search rather than by algebraic construction procedures.

3.3 *Decoding of Tree, Trellis and Convolutional Codes*

Sequential decoding is an efficient method for finding the most probable codeword in a tree code, given the received sequence y, without searching the entire tree. In sequential decoding, the received alphabet Y need not be equal to X. One begins at the first node and tentatively chooses the branch whose code symbols are most likely to have produced that portion of the received sequence. A measure of the difference between the tentatively chosen code symbols and the corresponding received sequence is retained. One proceeds by tentatively choosing the most likely branch from each successive node until the rate of growth of the difference measure indicates that the path being followed is incorrect. One then backtracks by going back to a previous node and taking a less likely branch. Backtracking and trying alternate paths continues until a path is found on which the rate of growth of the difference measure is satisfactory. Of course, the critical factors in this approach are the choice of the proper difference measure and a procedure to decide whether the rate of growth of this measure is or is not satisfactory.

An interesting modification of this algorithm is the stack algorithm. Here the decoder stores the difference measure on several paths and extends that path which appears most likely to be correct. When that path temporarily loses favor because of the rate of growth of its difference measure, the next most likely path is extended. All paths investigated are stored in the decoder until the storage capacity of the decoder is exceeded. Then the least likely paths are dropped from consideration.

Viterbi's maximum-likelihood decoder for convolutional codes makes use of the fact that there is a trellis structure for convolutional codes (VITERBI, A. J., 1967). In fact, it applies to any trellis code, not just convolutional codes. The essence of the procedure is to keep only one path to any node in the trellis; of course, the path to keep is the most likely one. The discarded paths to any node can never lead to the most likely codeword. If the trellis is generated from a K-state machine, only K paths ever need be retained by the decoder.

Algebraic decoding algorithms exist for certain convolutional codes. Some codes are majority logic decodable in that several parity checks are calculated for each message digit and a majority vote on the correctness of the digit is taken. In other cases a form of syndrome decoding is employed.

4. PERFORMANCE

Of prime interest to communications engineers is the increase in performance furnished by coding systems as compared to uncoded systems. We will take the probability of error in our binary message stream (either the bit error probability or the probability of error in a block of k message digits) as our measure of performance.

The efficacy of coding depends heavily on the particular communications channel. We will consider here two different channels. In the first, the only channel perturbation on the transmitted signal is additive white Gaussian noise. In the second, we will assume that the transmitted signal experiences Rayleigh fading and also is corrupted by additive Gaussian white noise. We consider both hard and soft decision receivers.

4.1. Additive Gaussian White Noise Channel (Hard Decisions)

This channel model which is a good approximation to transmission from deep space has been well studied in the literature. We will take as our baseline system an uncoded binary, phase-shift keyed system employing coherent detection. For a bit error probability of 10^{-7}, an 11 db signal to noise ratio is required while for a bit error probability of 10^{-5} the required ratio is about 9.6 db. (By signal-to-noise ratio we mean the ratio of the received energy per bit to noise power density.)

When we consider coded systems, we will assume that the information rate (in bits per second) for all systems fixed. Thus, the pulse duration of the uncoded and coded systems differ. The required ratio of received energy per information bit to noise power density as measured in db for a block error rate of 10^{-7} for various block codes is given in Table I. For each code we assume hard decisions at the receiver and bounded distance decoding where the decoder corrects all error patterns containing t or fewer errors.

Table I

m	k	R'	t	P_{ew}	(S/N_0R) db	Comments
23	12	.522	3	1×10^{-7}	9.3	Golay code
21	12	.571	2	1×10^{-7}	10.0	BCH
31	16	.517	3	1×10^{-7}	9.3	BCH
45	29	.644	2	1×10^{-7}	9.3	BCH
31	21	.678	2	1×10^{-7}	10.3	BCH
63	36	.571	5	1×10^{-7}	8.0	BCH
63	39	.619	4	1×10^{-7}	8.5	BCH
63	45	.714	3	1×10^{-7}	9.0	BCH
73	45	.616	4	1×10^{-7}	8.5	BCH
127	92	.724	5	1×10^{-7}	8.0	BCH
127	71	.559	9	1×10^{-7}	7.0	BCH
255	179	.702	10	1×10^{-7}	7.0	BCH
255	115	.451	21	1×10^{-7}	6.5	BCH
1	1	1.000	0	1×10^{-7}	11.0	Uncoded

We note that codes of moderate complexity save approximately 2 to 3 db in required signal-to-noise ratio over uncoded systems while the very complex (255,115) 21 error correcting code achieves a saving of 4.5 db. It is to be noted that we are comparing codes with different block lengths and that we have fixed the block error probability and not the bit error probability. However, essentially the same result is obtained when we fix the bit error probability. At a block error probability of 10^{-5} approximately 1 db less signal-to-noise ratio is required.

For the same channel model, convolutional codes outperform block codes of the same rates. A rate 1/2 convolutional code of long constraint length employing sequential decoding or the stack algorithm requires a ratio of energy per bit to noise power density of approximately 4.5 db in order to achieve a bit error probability of 10^{-7}. This is a saving of 5.5 db over the uncoded system but requires a very complex decoder.

Short constraint length convolutional codes employing Viterbi decoding also outperform block codes for this channel. A binary rate 1/2 convolutional code of moderate constraint length requires a ratio of energy per bit to noise power density of about 7 db in order to achieve a bit error probability of 10^{-5}. Shorter constraint length codes require somewhat higher signal-to-noise ratio but savings of more than 3 db are obtained for relatively simple codes (and decoding algorithms) (HELLER, J. A., JACOBS, I. M., 1971).

4.2 Additive Gaussian White Noise Channel (Soft Decisions)

For an additive Gaussian white noise channel, the maximum likelihood receiver for uncoded bipolar signalling consists of a matched filter followed by a threshold decision device. In the previous section it was assumed that such a detector was used for the coded case prior to the decoding circuitry. Thus the decoder was presented with a sequence of 0's and 1's at its input.

It is well known that in the coded case the analog signal at the output of the matched filter prior to the thresholding contains more information than the "hard decisions" emanating from the threshold device. In fact no information is lost by the matched filtering and these "soft decisions" at the output of the matched filter contain all the information required to make a maximum likelihood decision in the coded case.

As a rule of thumb, one can say that for any given code, one achieves an additional savings of approximately 2 db by using the soft decisions at the decoder input rather than the hard decisions. In principle, this 2 db savings can be obtained for both block and convolutional codes. In practice, however, soft decision decoding is much easier to use for convolutional codes than for block codes.

At a bit error probability of 10^{-5}, the following table (HELLER, J. A., JACOBS, I. M., 1971) gives the performance of some convolutional codes of constraint length 7 using the Viterbi algorithm and soft decision decoding.

Type	Rate	(Energy per bit/noise power density) db
Convolutional	3/4	4.4 db
Convolutional	1/2	5.5 db
Convolutional	1/3	4.0 db

Longer constraint length codes achieve even better performance but for very long constraint lengths the Viterbi algorithm is impractical and one must use sequential decoding or the stack algorithm.

Soft decision decoding of block codes theoretically show comparable performance to convolutional codes but the complexity of decoding often makes such a scheme impractical. One can always build an optimum maximum likelihood decoder with a decoding complexity proportional to the number of codes. For a binary (n,k) code, this means the decoding complexity is proportional to 2^k. Recently (WOLF, J. K., 1978) an optimum algorithm was presented which has a complexity proportional to $2^{(n-k)}$ for such codes. Various sub-optimum algorithms show promise.

The performance of such codes improve with blocklength. The following table gives the required ratio of energy per bit to noise density for orthogonal codes using soft-decision maximum likelihood decoding in order to achieve a bit error probability of 10^{-7}.

(n,k)	2^k	(required signal-to-noise ratio) db
(8,3)	8	10 db
(16,4)	16	9 db
(32,5)	32	8.3 db
(64,6)	64	7.6 db
(1024,10)	1024	5.9 db
$(2^{15},15)$	2^{15}	4.9 db
$(2^{\infty},\infty)$	2^{∞}	-1.6 db (limiting case)

The extended Golay_5(24,12) code requires about 5.5 db signal to noise ratio in order to achieve a bit error probability of 10^{-5}.

4.3. Rayleigh Fading Channel

We first consider a block coding scheme (PIEPER, J. F., PROAKIS, J. G., REED, R. R., WOLF, J. K.) where the data bits are represented by n bits using an (n,k) block code. These n bits are assigned to n frequency slots so that if the bit is a 1 that frequency is transmitted while if it is a 0 the frequency is not transmitted. It is assumed that the frequencies fade independently in accordance with Rayleigh statistics and that all frequency channels are corrupted by independent Gaussian noise of flat spectrum.

To decode, the squared magnitudes of the responses of the matched filters corresponding to the n frequency cells are first formed. We call these "decision variables". Then, for each code word, these decision variables corresponding to 1's in the code word are summed. If all the code words have the same Hamming weight, that is, the same number of 1's, then the maximum likelihood decoder decodes to that code word having the largest sum of decision variables. Such a scheme is only practical for moderate values of k (say $k < 10$).

A similar transmission scheme can be considered for convolutional codes. For example for a rate 1/2 code, every message digit corresponds to two channel symbols and thus two frequencies.

Curves of performance for both block and convolutional codes are shown in Figure 4. It is seen that the savings in signal-to-noise ratio achieved by coding is much greater here than in the non-fading channel model.

5. SUMMARY

The purpose of this paper was to present an overview of various coding techniques available for error control over noisy communications channels. The parameters of several common codes were given. The performance of these codes for signalling over two common communications channels was then presented.

ACKNOWLEDGEMENT

This research was supported by the United States Air Force, Office of Scientific Research under Grant AFOSR-74-2601. Portions of this paper were taken from the paper "A Survey of Coding Theory: 1967-1972," IEEE Trans. on Information Theory, Vol. IT-19, pp. 381-389, July 1973.

REFERENCES

HELLER, J. A. and I. M. JACOBS, 1971, "Viterbi Decoding for Satellite and Space Communications," IEEE Transactions on Communications Technology, Vol. COM-19, pp. 835-848.

PETERSON, W. W. and E. J. WELDON, Jr., 1972, Error Correcting Codes, Second Edition, M.I.T. Press, Cambridge, MA.

PIEPER, J. F., J. G. PROAKIS, R. R. REED and J. K. WOLF, "Design of Efficient Coding and Modulation for a Rayleigh Fading Channel," to be published in the IEEE Transactions on Information Theory.

RUDOLPH, L. D. and W. E. ROBBINS, 1972, "One-Step Weighted-Majority Decoding," IEEE Transactions on Information Theory, Vol. IT-18, pp. 446-448.

VITERBI, A. J., 1967, "Error Bounds for Convolutional Codes and An Asymptotically Optimum Decoding Algorithm," IEEE Transactions on Information Theory, Vol. IT-13, pp. 260-269.

WOLF, J. K., 1973, "A Survey of Coding Theory: 1967-1972," IEEE Transactions on Information Theory, Vol. IT-19, pp. 381-389.

WOLF, J. K., 1978, "Efficient Maximum Likelihood Decoding of Linear Block Codes Using a Trellis," IEEE Transactions on Information Theory, Vol. IT-24.

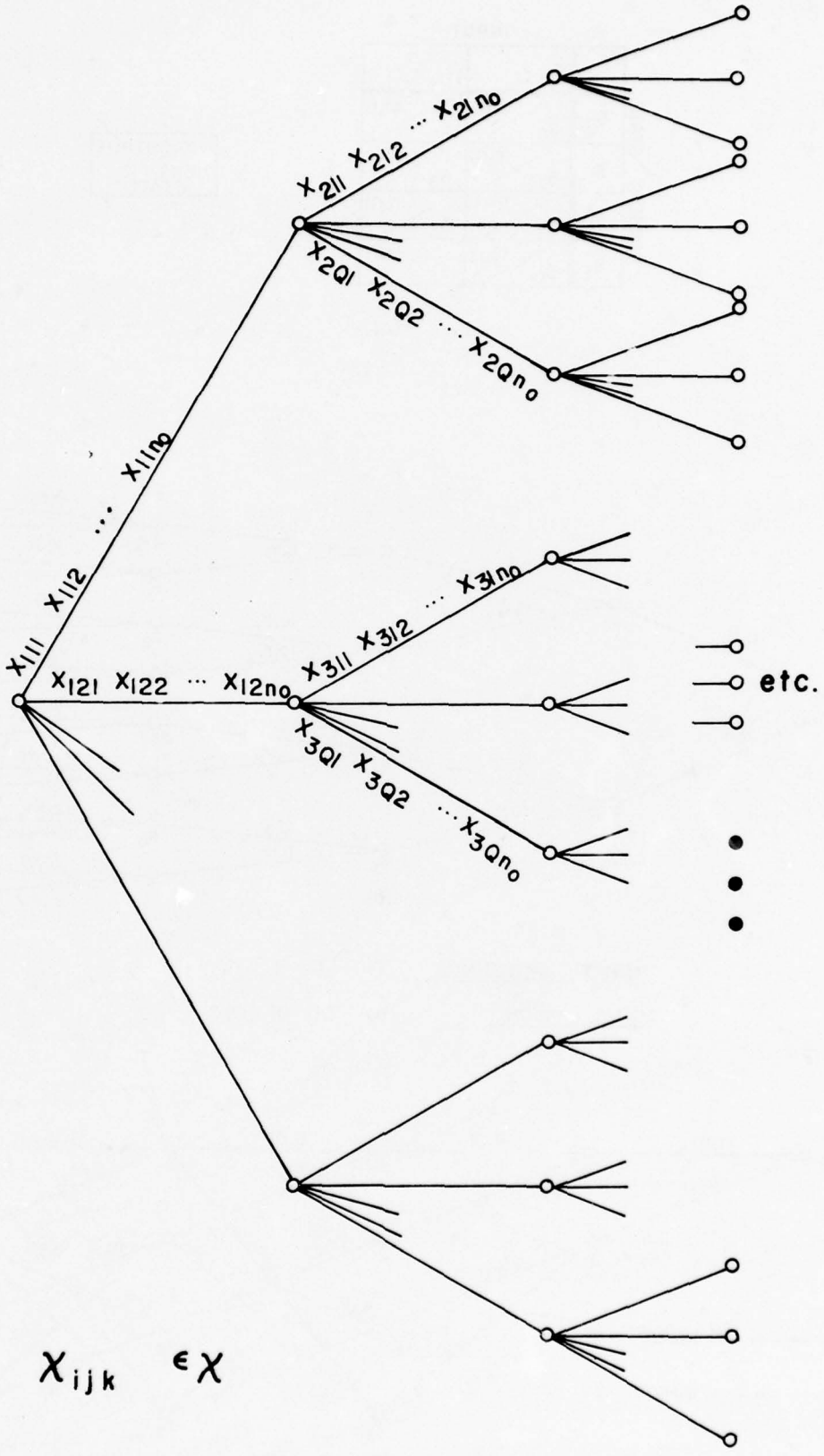
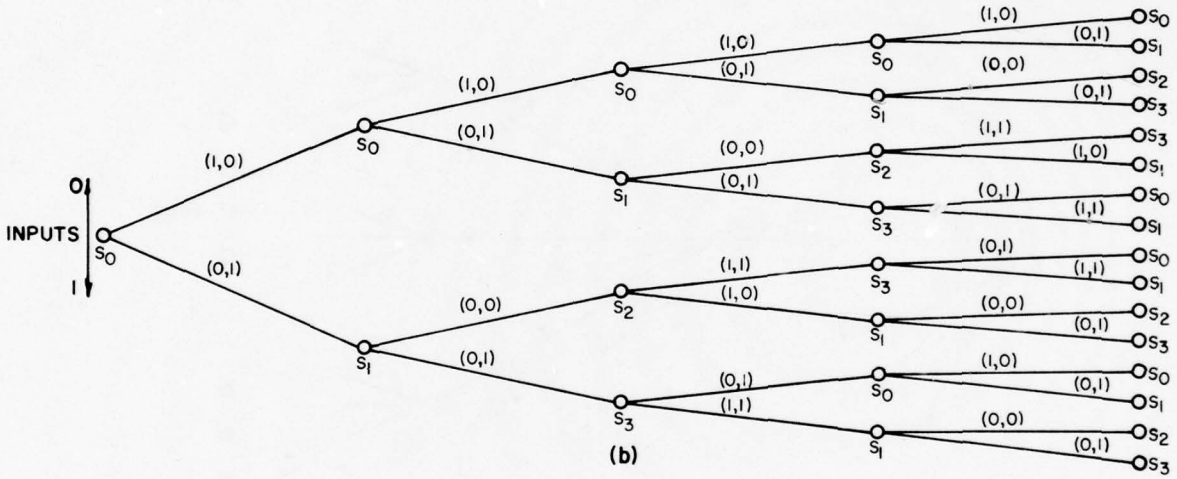


Figure 1 Tree code

PRESENT STATE \ INPUT	0	1
S_0	S_0 / (1,0)	S_1 / (0,1)
S_1	S_2 / (0,0)	S_3 / (0,1)
S_2	S_3 / (1,1)	S_1 / (1,0)
S_3	S_0 / (0,1)	S_1 / (1,1)

(a)

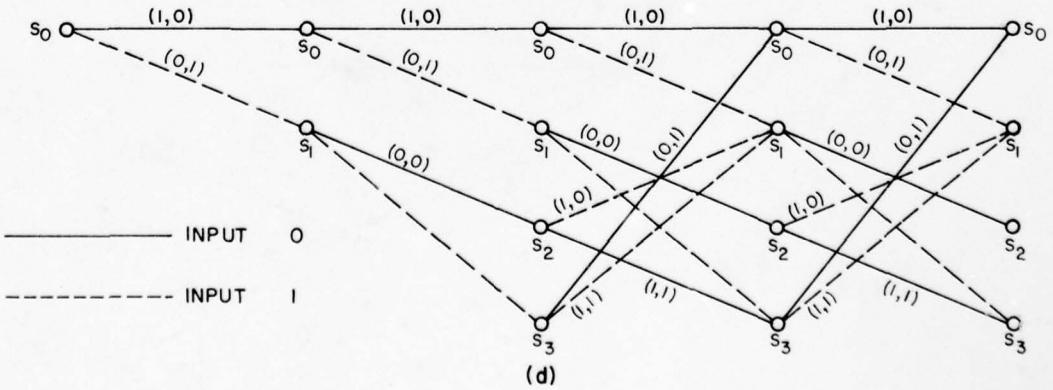
OUTPUT / NEXT STATE



(b)

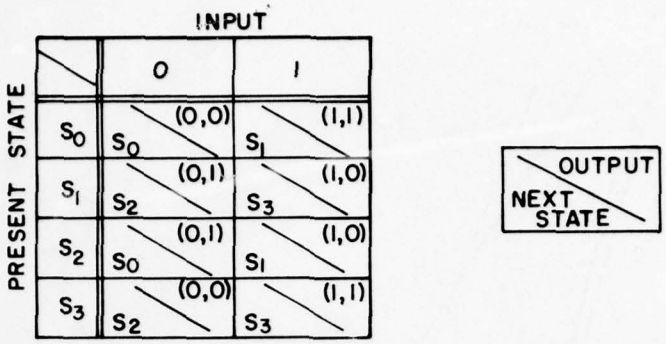
INPUT SEQUENCE	0	1	1	0	...
CODE WORD	10	01	01	01	...

(c)

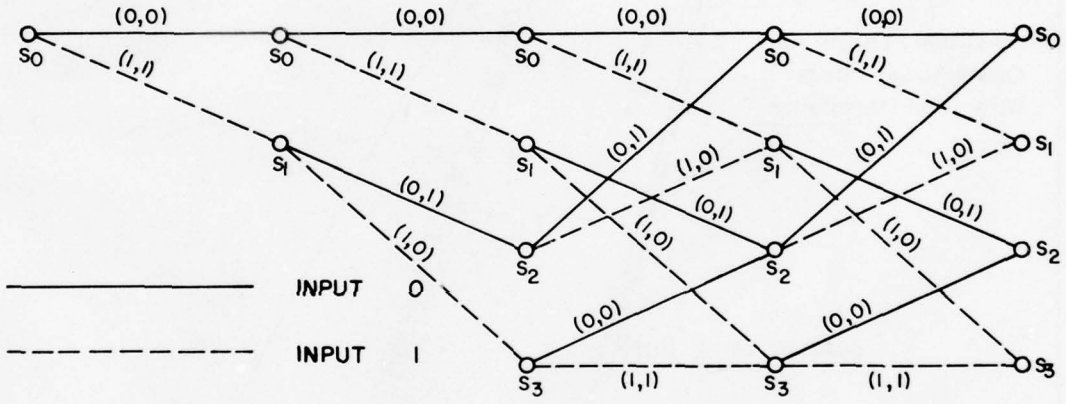


(d)

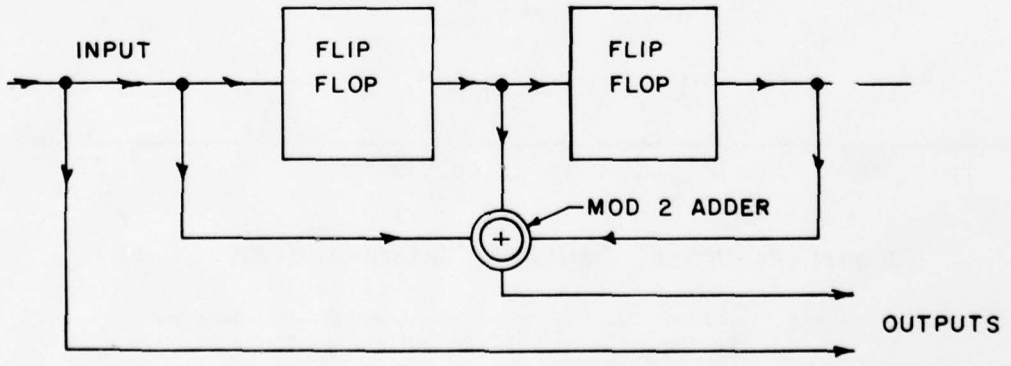
Figure 2 (a) State transition table for tree code
(b) Tree for code
(c) Input sequence and code word
(d) Trellis for code



(a)



(b)



(c)

Figure 3 (a) State transition table for convolutional code
(b) Trellis for code
(c) Block diagram for encoder

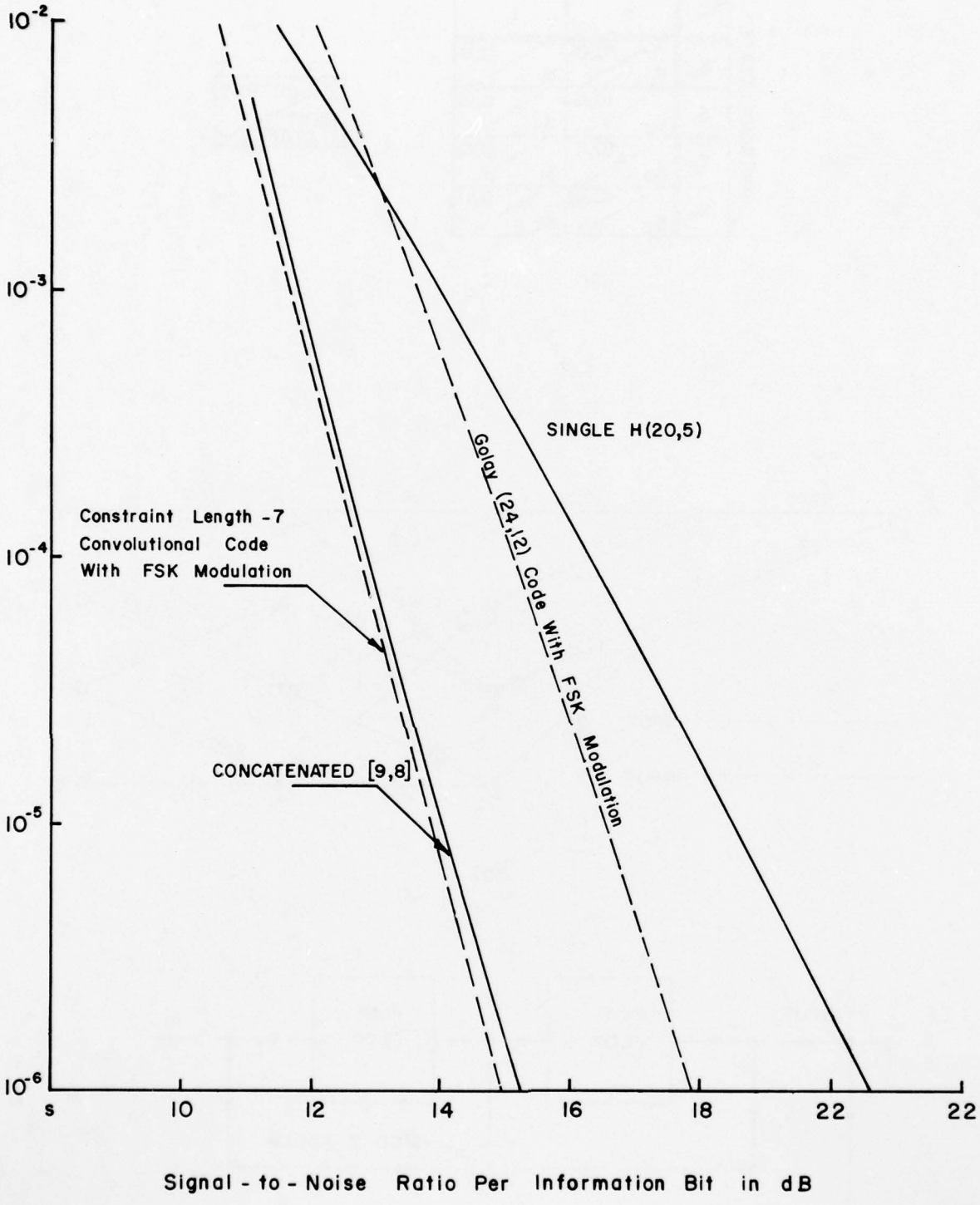


Figure 4 Probability of error versus signal-to-noise ratio per information bit (in db) for certain codes in a Rayleigh fading environment

FORWARD ERROR-CORRECTION FOR THE AERONAUTICAL SATELLITE COMMUNICATIONS CHANNEL

Alan Sewards
Department of Communications
Communications Research Centre
P.O. Box 11490, Station 'H'
Ottawa, Ontario, Canada, K2H 8S2

Leo Beaudet and Hassan Ahmed
Miller Communications Systems Limited
39 Leacock Way
Kanata, Ontario, Canada, K2K 1T1

SUMMARY

The paper identifies the problems caused in an L-band aeronautical satellite communications channel by ocean-surface multipath and shows that data messages transmitted over the channel exhibit unacceptably high error-rates under typical conditions. Channel characteristics and techniques for reducing the error-rate are discussed, leading to the choice of a diffuse threshold-decodable convolutional forward error-correcting code. Two implementations of coder/decoders for this code are described, one using standard IC logic, and the other a Z-80A microprocessor. Results of tests of the IC coder/decoder with noise, data error bursts, and a system simulator are quoted, which show that the codec performed as expected and is capable of reducing the error-rate to 10^{-5} or better.

1. INTRODUCTION

In 1974 a Memorandum of Understanding setting up the AEROSAT program was signed between ESRO, U.S.A. and Canada. This program envisaged the launching of satellites in geostationary orbit to provide communications between aircraft on Atlantic routes and ground stations in North America and Europe. Communications between aircraft and satellite would be in the 1550-1650 MHz L-band, with the satellite/ground backhaul at 5000-5250 MHz. In the same year, the ATS-6 satellite was launched by NASA, and a co-ordinated program of L-band experiments involving ESRO (later ESA), FAA, U.S. Coastguard and Canada commenced (NASA 1973). These experiments had as their objective the measurement of the properties of the transmission link between aircraft and satellites, including propagation and multipath effects, and measurement of the performance of candidate voice and data (digital) modems (channel units) over typical links. Tests were also conducted on several designs of aircraft antenna to determine their characteristics including their ability to discriminate against multipath.

The results of this experimental program were very revealing. It was demonstrated that the error performance on a typical aeronautical satellite communications channel at these frequencies was grossly different from the classical free-space channel often assumed for satellite communications, due to the interference arising from multipath reflections from the surface of the earth, notably the ocean surface. Since multipath interference arrives from a different direction, it can, to some extent, be discriminated against by judicious design of the aircraft antenna, but even in this case the error probability for digital transmissions does not fall exponentially with increasing signal-to-noise ratio. The channel power required under such conditions to achieve an error probability of better than 10^{-5} at information rates of 1200 and 2400 bps may be prohibitive.

The performance of the aeronautical channel can be improved by applying one or more of the following techniques: automatic repeat request (ARQ) followed by retransmission of all or parts of the message, channel diversity, more complex receivers optimized to work in the presence of multipath, and forward error correction coding. The latter offers, perhaps, the most attractive solution in terms of hardware, bandwidth, cost and system design.

Many of the well-known coding techniques are not suitable for the aeronautical channel because of the fact that the effect of multipath is to produce fading which results in bursts of errors. A classical random error-correcting code designed for a memory-less channel must be very powerful to allow correction where a large percentage of consecutive bits are in error. Burst error-correcting codes may be more efficient as they can be designed to trap and correct isolated error bursts. However, the classical burst model is probably not appropriate either, since it is not highly probable that the necessary guard space on either side of the burst will be free from scattered errors.

The channel model appropriate to the aeronautical satellite channel and the choice of a suitable error-correcting code is discussed in the paper. It is concluded that a diffuse convolutional code offers the best compromise between performance and complexity. Two implementations of such a code are described, using random logic and a microprocessor, together with results of tests including tests using an AEROSAT channel simulator. In particular, using such a code, the desired error probability of 10^{-5} at an information bit rate of 1200 bps with a channel quality of 43 dBHz can be met with a typical aircraft antenna signal-to-interference ratio of 10-13 dB.

2. CHANNEL MODEL

The L-band communications channel between aircraft and satellites can be characterized as a channel with the theoretical free-space loss plus effects due to ionospheric propagation and earth-surface multipath. The former is generally covered by adding a margin to compensate for periods of signal attenuation: in the case of the L-band aeronautical channel, a margin of 2 dB will only be exceeded 0.1% of the time. Earth-surface multipath is, however, more difficult to describe and it is only since the results of the joint L-band international tests using the ATS-6 satellite were published that an adequate understanding of its effects has been obtained (Schroeder, E.H., 1976, Chinnick, J.H., 1977, Brown, D.L., 1976).

Earth-surface multipath represents the sum of the direct signal between aircraft and satellite and the signal which is reflected from the surface of the earth. Because aircraft antennas at L-band are by their nature (if not their design) relatively wide beam, for satellite elevation angles of 15° or less a significant reflected signal is received. Due to the better and more consistent reflection characteristics of the sea surface as compared with the land surface, this problem is worse over the ocean and for the remainder of this paper we will concentrate on the question of ocean-surface multipath. The effect of the reflection is to present an interfering signal to the receiver whose magnitude can range from zero up to about 6-8 dB less than the direct signal, and which is modified by the reflection properties of the reflecting surface. This surface produces specular and diffuse reflections and as a result, the reflected signal is delayed by the path length difference (5-25 microseconds), spread in time and frequency (1 microsecond and 10-100 Hz), and doppler shifted due to differential movement by up to about 20 Hz. The combined direct and reflected signals have been found to affect receivers in different ways; however, in general, it can be stated that voice modulation schemes were little affected but digital data PSK transmissions were seriously perturbed. The general shape of the error curves resulting from ocean-surface multipath is shown in Figure 1, where it can be seen that the normal bit-error-rate (b.e.r.) curve is flattened out at high signal-to-interference (S/I) ratios to the point where the b.e.r. is almost independent of channel signal-to-noise ratio (C/N_O or E_b/N_O). It was found that, in many typical aeronautical satellite communications conditions, a b.e.r. of better than 10^{-3} was rarely achieved. Examples of measured results for CPSK obtained by Schroeder, E.H., 1976 are given in Figure 2.

The principal cause of channel errors is destructive interference between the direct and reflected signals, i.e., fades. Due to the nature of the reflected signals, as noted above, these fades do not last long and measurement statistics indicate that the probability of a burst of more than seven consecutive bit errors at a data rate of 1200 bps is very low, as shown in Figure 3 from Schroeder, E.H., 1976. However, unlike some other fading channels, there is not a high probability of an error-free space each side of the burst.

It will be evident from the above discussion that multipath effects on digital PSK signals cannot be simply compensated for by increasing channel C/N_O. As an example, to improve the b.e.r. obtained for a S/I ratio of 8 dB with a doppler spread of 100 Hz on the multipath signal would require an increase in channel C/N_O of about 10 dB. Clearly this is extremely expensive. Other means of obtaining the desired b.e.r. must therefore be sought, and the use of forward error-correction is one such way.

Based on the channel model described above, the code must improve a channel error rate typically in the 10^{-2} to 10^{-3} region to an output data b.e.r. of not worse than 10^{-5}. It must operate in the presence of random errors as well as burst errors, where the burst length is assumed to be no longer than seven bits at 1200 bps. Other parameters assumed are summarized below:

S/I ratio	10-13 dB
C/N_O	43/44 dBHz
Multipath specular reflection delay	5-25 microsec
Multipath spread	1 microsec
Multipath bandwidth	10-100 Hz
Differential doppler	20 Hz

3. CHOICE OF ERROR-CORRECTING CODE

Two basic approaches exist to the problem of combatting noise in a channel where both random and burst disturbances occur. The classical method is to use a code that is good for both random and burst errors, such as diffuse threshold-decodable convolutional codes; interleaved random-error-correcting block codes (with or without channel measurement decoding); character-error-correcting block codes, with or without interleaving; cyclic product block codes; cyclic block codes; and iterated burst and random error-correcting block codes. The other approach is to use an adaptive scheme in which separate decoding algorithms are employed for the same code, depending on whether burst or random errors are detected. Implementations of this scheme have been proposed for block codes and orthogonizable convolutional codes. Its overall decoded error probability is lower-bounded by the probability that the decoder fails to pick the correct decoding algorithm.

A number of coding schemes were considered for the aeronautical satellite application and the pros and cons are discussed in more detail elsewhere (Lyons, R., 1978). The conclusion was that the most suitable code for this application was a rate one-half diffuse threshold-decodable convolutional code. The theoretical basis of the code is described by Wilson, S.G., 1976 and Kohlenberg, A.K., 1968, and it has been shown to provide about 8 dB of coding gain at a 10^{-5} b.e.r. for 1200 bps information transmission over a simulated DPSK AEROSAT channel with an S/I of 10 dB and a channel fading bandwidth of 120 Hz (Wilson, S.G., 1976).

The coder/decoder, to be discussed in more detail below, is shown in block diagram form in Figure 4. The code rate is one-half, and it is systematic, i.e., the information appears explicitly in the symbol stream. Generation is similar to that for any convolutional code, except that the encoding constraint length is made long to diffuse the information over a span of output bits. At least β bits separate each tap in the shift register used to provide the parity or check bits. The greater β is, the more diffuse the code, and the greater its burst-correcting power. In the decoder, parity bits are obtained from the received information bits by the same process used in the encoder, and compared with the received parity bits to produce 'syndrome' bits. A syndrome bit of '1' indicates that the parity bits differ and that an information or a parity bit error has occurred. Syndrome bits are stored in a register of length $3\beta+1$, with majority logic used to correct information bits. For this code, the decision to invert the information bit is made if three of the four syndrome bits examined are '1'. In addition, once the decision to change an information bit is made, the correction is fed to the syndrome register so that the effects of that information bit originally in error are removed from the syndromes.

It has been shown (Kohlenberg, A.K., 1968) that the decoder will correct all bursts up to length 2β (β information bits + β check bits) provided that there is an error-free guard space between bursts of length $6\beta+2$ bits. The decoder will also correct any single or two-bit error pattern in the eleven bits used in the decoding process. The principal source of errors lies in three-out-of-eleven patterns, of which, however, about half are correctable. Of the 165 possible patterns only 85 produce an error, so the 'first error' probability is about $85\ p^3$ (where p = error probability for a binary symmetric channel). The error probability assuming a white Gaussian noise channel with random errors is about $166\ p^3$ for small p (Kohlenberg, A.K., 1968). The closeness of these two values results from the negligible error propagation observed with these diffuse codes. One factor of importance here is that no more than two incorrect syndromes resulting from past decoding errors are used simultaneously. Thus, a decoding error at worst can act like two single errors in the decoder, which will not make another mistake unless a separate third error is present. In other words, the decoder will not continue to make errors if the channel has ceased to produce them.

4. IMPLEMENTATION OF CODER/DECODER

The error-correcting code described above has been implemented in two forms, one using hardwired integrated circuit logic, and the second using a Z-80 microprocessor. Besides encoding the data, the encoder produces a clock at the output bit rate. For message control, the encoder and decoder require a start and end-of-message signal which indicates the presence of valid data at the input. In the case of the encoder, this signal reflects the encoding delay and the presence of the overhead bits. The hardwired version will be described followed by the microprocessor version.

The encoder block diagram is shown in Figure 5. The parity generator incorporates a shift register of length $3\beta+1$, which is initialized to zero. Information bits are clocked into this register at 1200 bps and the bits at four taps are modulo-2 added to produce a parity bit. A parity bit is produced for each information bit clocked into the register and the two bit streams are then multiplexed to produce the 2400 bps output stream. Parity bits are transmitted until the last information bit has been clocked through the parity generator shift register. Zeroes are inserted into the information bit slots during the time it takes for the final bit to clock through the register. This produces an overhead of $3\beta+1$ bits for each message.

The decoder is shown in Figure 6. Information bits are passed through a parity generator identical to that in the encoder. The resulting parity bits are added modulo-2 with the corresponding received parity bits to produce the syndromes. These latter bits are clocked into a shift register and the syndromes at five taps examined by a majority logic circuit to decide whether or not an error has occurred. If so, the erroneous information bit is inverted together with the syndromes which identified it.

One possibility that has been envisaged is that the receiver may lose one bit in which case the information and parity bits demultiplexed by the decoder will be interchanged. If this occurs, a large number of errors will ordinarily be detected. To cope with this situation, a bit synchronizer circuit has been added which counts the errors in the information bit stream, as well as those in the parity stream on the assumption that they are actually information bits. Provided that the error rate of both streams does not exceed a threshold (4 errors in 50 bits), the synchronizer locks the decoder on the stream producing the fewest errors. As the synchronizer requires 50 bits to make a decision, the data is delayed by 50 bits (at the output rate). Resynchronization is achieved by switching between the input data and the input data delayed by one bit.

Because the codec was designed for experimental purposes the value of β was made variable, selectable by on-board switches. β values of 4, 8, 12 and 16 were provided. An additional feature implemented in the codec was the option of selecting a two bit delay between parity and information bits before multiplexing in the encoder, and a corresponding delay in the decoder (shown dotted in Figure 4). If PSK with differential encoding is used to resolve the phase ambiguity at the receiver, with DECPSK demodulation two consecutive bit errors are produced for each isolated channel error. To avoid degradation of codec performance in this situation, the parity bit may be separated by two bits from its corresponding information bit by selecting this delay.

The codec was implemented using CMOS logic operating at 5V with TTL input/output buffering. It was laid out on two 230x100 mm PC boards, one containing the encoder and interface buffers, and the second the decoder. A total of 53 integrated circuits (ICs) were used, 16 in the encoder and 37 in the decoder. 17 of the decoder ICs were required to implement the resynchronization scheme. Power consumption was less than 1 watt at 5V.

5. MICROPROCESSOR IMPLEMENTATION

The codec was also implemented in software using a microprocessor integrated circuit to evaluate the advantages of using a microprocessor and specifically to determine: (1) how much hardware can be eliminated, (2) the amount of software required, (3) the timing constraints and maximum data rates possible, and (4) the interface requirements.

Among the 8 bit microprocessors the Zilog Z-80A microprocessor was chosen because it presently has the most powerful instruction set and the fastest machine cycle time. The Z-80A is similar in architecture to the Intel 8080 but has a much larger instruction set, more addressing modes, and more than twice as many internal registers. The clock frequency of the Z-80A is 4 MHz which gives a machine cycle time of 250 nsec.

The objective was to have one processor handle both the encoding and decoding functions at an input information rate of 1.2 kb/s. The initial approach was to make the interface hardware as simple as possible and to operate on a bit-by-bit basis. Using this approach an interrupt flag is raised by the interface when a valid bit arrives. The microprocessor transfers the bit to a circular buffer in RAM created by software, performs the necessary modulo 2 operations with bits previously entered, and outputs the results. Since the amount of memory is not critical each bit is stored in a separate byte to avoid shifting and masking operations. Thus the encoder requires a buffer of $3\beta+1$ bytes or 49 when β is 16. In addition a pointer is required for each tap used to generate the parity bit, and these pointers must be updated with each new bit. The decoder requires buffers for the input data, the syndromes, and the delay required in the bit synchronizing scheme. The software also has to respond to the input start and end of message (S/E MESS) signal and generate an appropriately timed S/E MESS signal for the output.

The timing contraints for an information rate of 1.2 kb/s are as follows. The critical timing is governed by the interval between the arrival of two bits at the decoder, 0.416 msec, during which an encoder and decoder bit must be processed. At 4 MHz one machine cycle requires 250 nsec and a typical instruction requires 7 to 10 machine cycles. For example a register add requires 4 machine cycles, a read from or write to memory requires 7, and a simple jump instruction requires 10. Thus the encoder and decoder programs together must have less than 150 to 200 instructions. With the overhead required to set up the buffers and pointers the number of instructions was found to surpass this limit by a considerable amount so this approach was discarded.

A more promising approach at the expense of greater interface complexity is a byte-oriented scheme in which the interface accummulates 8 bits of the message before raising an interrupt flag. The microprocessor reads the 8 bits and stores them as a byte in RAM. Furthermore the algorithms for parity bit generation, error correction, and bit synchronization are carried out 8 bits at a time. The time available between successive interrupts of the decoder is now 3.33 msec which allows for 800 to 1200 instructions.

However the complexity of the interface is increased because a counter is required to determine when the 8th bit has arrived, the processor must be told whether the message has ended somewhere within the byte, and the data must be converted from serial to parallel format. Double buffering is required at the interface output to ensure that it will always be ready to accept the most recently processed byte and that there will always be a smooth flow of data.

When 8 bits have accummulated at the interface, they are read by the microprocessor and stored in RAM according to the structure shown in Figure 7. Notice that of the 8 bits that require processing, all of the $3\beta+2$, $2\beta+2$, and $\beta+2$ bits occupy a single byte. Furthermore, the subscript '1' bits can also be made to occupy the same byte through one masking operation.

Consequently, one can form the parity 'word' simply as the exclusive-or of the bytes mentioned above. This requires only four EXOR operations or 4 μsec.

The data structures depicted in Figure 7 have one important feature. Since eight bits are effectively processed in parallel, and since the byte size is related to the chosen values of β in a simple manner a certain amount of masking is required to set up the subscript '1' bits. As a result, for $\beta = 8$ a storage location is required to hold i_1^1. This storage is however not required for $\beta>8$ since i_1^1 will be the first bit in the final byte of the data structure.

Bench mark programs were written for the various values of β to ascertain the timing requirements. Elapsed time per eight bits was split into two parts:-

(a) the actual time for processing to form a parity byte;

(b) the additional time for data management.

The quoted execution times in the following table are for a 0.25 μsec cycle time. All values are given per eight bits for an information bit rate of 1.2 KHz.

β	MANAGEMENT PROGRAM STATES	PROCESSING PROGRAM STATES	TOTAL STATES	EXECUTION TIME (μs)
4	498	273	771	193
8	560	54	614	154
12	684	261	945	238
16	808	82	890	224

TABLE 1: Encoder Processing Requirements for 8 Information Bits.

These are the worst possible times which includes code to initialize the encoder at the end of a message in preparation for the subsequent message. For long messages (compared to eight bits), these states are only executed once such that the average execution time is reduced as shown:

β	STATES SAVED	TIME SAVED (μs)	AVERAGE EXECUTION TIME (μs)
4	158	40	153
8	258	65	89
12	358	90	148
16	458	115	108

TABLE 2: Encoder Processing Time Excluding Initialization Procedures.

Table 2 should be viewed with caution since the worst case timing must be utilized at least once per message.

Finally, it is interesting to note that less time is required for $\beta = 8$ than for $\beta = 4$ and similarly for $\beta = 16$ as opposed to $\beta = 12$. This results from the masking overhead required to form the $3\beta+2$, $2\beta+2$, etc. bytes when β is not an integral multiple of the processing word size of 8 bits.

In the decoder, structures similar to that of the encoder are employed in the data, syndrome and output registers. Again for $\beta<8$, there are two bits at the end of the structure in analogy to the single bit of the encoder. In addition, there is a single bit preceding the data register to save the most recent parity bit because a parity word can only be computed for up to and including the most recent information bit. A similar storage precedes the syndrome register. Its purpose is to cause the $S_{3\beta+1}$, $S_{2\beta+1}$, $S_{\beta+1}$ and S_1 syndromes to occupy single bytes rather than being resident in two bytes, thus reducing the masking overhead.

The software implementation of the decoder follows the hardware structure of Figure 6 very closely with the exception that the output register utilized in the bit synchronization scheme is 12 bytes or 96 bits long as opposed to 100 bits in the hardware version. Thus four information bit errors in 96 bits is the threshold for activation of the bit synchronization scheme provided less than four parity bit errors have occurred.

Benchmark programs were again written to ascertain the worst case timing based on a 0.25 μsec clock cycle. Results are given in Table 3.

β	EXECUTION TIMES (ms)		
	MANAGEMENT PROGRAM	PROCESSING PROGRAM	TOTAL EXECUTION
4	1.078	.39	1.468
16	1.078	.63	1.708

TABLE 3: Decoder Processing Requirements for 8 Data Bits.

To support the encoding and decoding processes which are independent and asynchronous, the interface to the processor was designed with the following characteristics.

Encoder and Decoder Inputs:

(a) For each input the interface receives eight bits and causes an interrupt on the 8th.

(b) At the same time it generates a control byte which contains a '1' in each position in which the data byte contains a valid data bit. The control byte is used by the processor to determine where the last valid data bit of a message occurs.

Encoder Output:

(a) The interface accepts a data byte, a parity byte, and a control byte. It multiplexes the data and parity on a bit by bit basis for serial transmission, and uses the control byte to set and reset the S/E MESS line.

(b) It interrupts the processor when transmission of the current byte is complete. The processor then outputs another byte for transmission.

Decoder Output:

(a) A byte consisting of multiplexed information and parity bits and a control byte are supplied by the processor to the interface. The interface demultiplexes the information bits and uses the control byte to set and reset the S/E MESS line.

(b) It raises an interrupt when it requires more data.

The software for the codec requires about 1.5 K bytes of storage and the complete circuit comprises less than 20 integrated circuits.

6. TEST RESULTS ON FEC CODEC

Three tests have so far been made to measure the performance of the hardwired logic version of the codec. These were performed in the laboratory and consisted of:

(a) tests of codec with PSK channel in presence of additive white Gaussian noise;

(b) tests of codec with simulated error bursts;

(c) tests of codec with a complete PSK channel using AEROSAT channel simulator at Transportation Systems Centre.

Tests over a real satellite link using a C130 aircraft and the ATS-6 satellite are also planned.

Tests (a) and (c) were conducted using the codec with a DECPSK channel unit manufactured by SED Systems Limited, Saskatoon, Canada. In test (a) thermal noise was added to the 70 MHz IF link between the modulator and demodulator, and the E_b/N_o was varied, at 2400 bps with and without the FEC codec in the circuit. A Hewlett Packard Model 1645A Data Error Analyzer was used as the data source. Results are shown in Figure 8, from which it can be seen that the channel unit, without coding, has an implementation loss of about 1 dB compared with the theoretical curve for DECPSK at 2400 bps. The effect of the codec should be to improve the DECPSK performance by 166 p[3]. The coding gain for the theoretical DECPSK is also shown in Figure 8. It can be seen that the codec produces the anticipated gain, amounting to some 3.6 dB at 10^{-5} b.e.r..

The tests with simulated error bursts were performed without noise using a burst error generator operating on data. The generator produces a pseudo-random sequence of length $2^{36}-1$ at a clock rate of 307.2 KHz. Whenever a selected bit pattern in the sequence occurs, a data bit is inverted, thus producing an isolated error. By varying the length of the selected pattern, the b.e.r. produced can be varied. Error bursts are generated by detecting a second pattern, 3 bits long, whenever the isolated error pattern is detected. In effect this produces an error burst instead of an isolated error one time in eight. The error burst length is under switch control and can be varied from 1 to 64 bits, all of which will be in error. Tests were run using this generator with several bursts lengths and a fixed value of $\beta = 8$ for the codec. Results are shown in Figure 9. It can be seen that useful coding gains are obtained for burst lengths up to 16, but with little or no gain for bursts of 20 bits. Tests were also made on the codec for several values of β from 8 to 16 with burst lengths of 16 and 20 bits. Results are shown in Figure 10, and confirm the ability of the codec to correct bursts up to 2β in length. It should be noted that bursts with every bit in error are not likely in actual channel conditions, and so the performance of the codec in a real environment should be better than measured in this test.

The tests using the AEROSAT channel simulator were performed at the Transportation Systems Centre, Cambridge, Massachusetts. The simulator was designed to represent as closely as possible the characteristics of the L-band aeronautical satellite communications cha-nel and allows the various parameters of interest to be varied over ranges of values typically encountered. A full description of the simulator is given by Duncombe, C.B., 1975. The tests were conducted using the SED Limited channel unit noted above. Unfortunately, during transportation the SED channel unit suffered some internal damage which was manifested in an increase of implementation loss to more than 2 dB at 10^{-5} b.e.r., and as a result, the absolute performance of the codec and channel unit is somewhat poorer than anticipated. However, measurements inciated that the coding gains expected were obtained. The channel and decoder b.e.r. were measured for various combinations of C/N_0, simulator multipath bandwidth, and codec β. Values of C/N_0 between 42 and 52 dBHz, multipath bandwidths of 100, 50 and 10 Hz, and β values of 4, 8, 12 and 16 were used. Results are presented in Figures 11. It is interesting to note from Figure 11 that the coding gains obtained are much closer to the results for an AWGN channel than the results from the burst tests where every bit was in error. As the S/I decreases or the channel C/N_0 increases the coding gain tends towards the burst test values.

7. CONCLUSIONS

It has been shown that ocean surface multipath on an L-band aeronautical satellite communications channel will prevent the transmission of 1200 bps digital PSK signals with the required b.e.r. of 10^{-5} without a prohibitive increase in satellite power or avionics performance. Increases of channel quality of the order of 10 dB are required to compensate for multipath where signal to interference ratios lie in the typical region of 10-13 dB. These effects of multipath may be reduced by the use of forward error correction, using a rate 1/2 diffuse convolutional code. Such a code is simple to generate and decode, and will give the required b.e.r. at a C/N_0 of about 43 dBHz. The codec has been implemented in hardwired logic form and also in microprocessor form where the coding and decoding functions are performed in software. Tests on simulated Aeronautical satellite channels have demonstrated that the predicted coding gain was obtained. Use of the technique involves a message delay of approximately 100 information bits plus the transmission of an overhead of 25 information bits for each message. While for some applications these requirements might cause difficulties, for the types of messages expected on the AEROSAT system, this forward error-correction technique is proposed as a cost effective implementation.

8. REFERENCES

Chinnick, J.H., 1977, and Burtt, D., "Canadian Aeronautical Satellite Tests Using the ATS-6 Satellite, 1974-75", Communications Research Centre, Ottawa, Ontario, Report Number 1308, 1977.

Duncombe, C.B., 1975, and Salwen, H., "Performance Evaluation of Data Modems for the Aeronautical Satellite Channel", Transportation Systems Centre, Cambridge, Massachusetts, Report FAA-RD-75-150, September 1975.

Kohlenberg, A.K., 1968, and Forney, G.D., "Convolutional Coding for Channels with Memory", IEEE Trans. Information Theory, September 1968, pp 618-626.

Lyons, R., 1978, and Beaudet, L., "Application of Forward Error-Correction Over Aeronautical Satellite Data Links", Communications Research Centre, Ottawa, Ontario, Report 1314, 1978.

NASA, 1973, "Integrated Test Plan for ATS-F L-Band Experiment", NASA Document TP-750-73-1.

Schroeder, E.H., 1976, and Thompson, A.D., Sutton, R.W., Wilson, S.G., and Kuo, C.J., "Air Traffic Control Experimentation and Evaluation with the NASA ATS-6 Satellite - Volume III: Summary of U.S. Aeronautical Technology Test Plan", Boeing Commercial Airplane Company. Report FAA-RD-75-173, Volume III, September 1976.

Wilson, S.G., 1976, and Sutton, R.W., Schroeder, E.H., "Differential Phase Shift Keying Performance on L-Band Aeronautical Satellite Channels: Test Results and a Coding Evaluation", IEEE Trans. Communications, March 1976, pp 374-380.

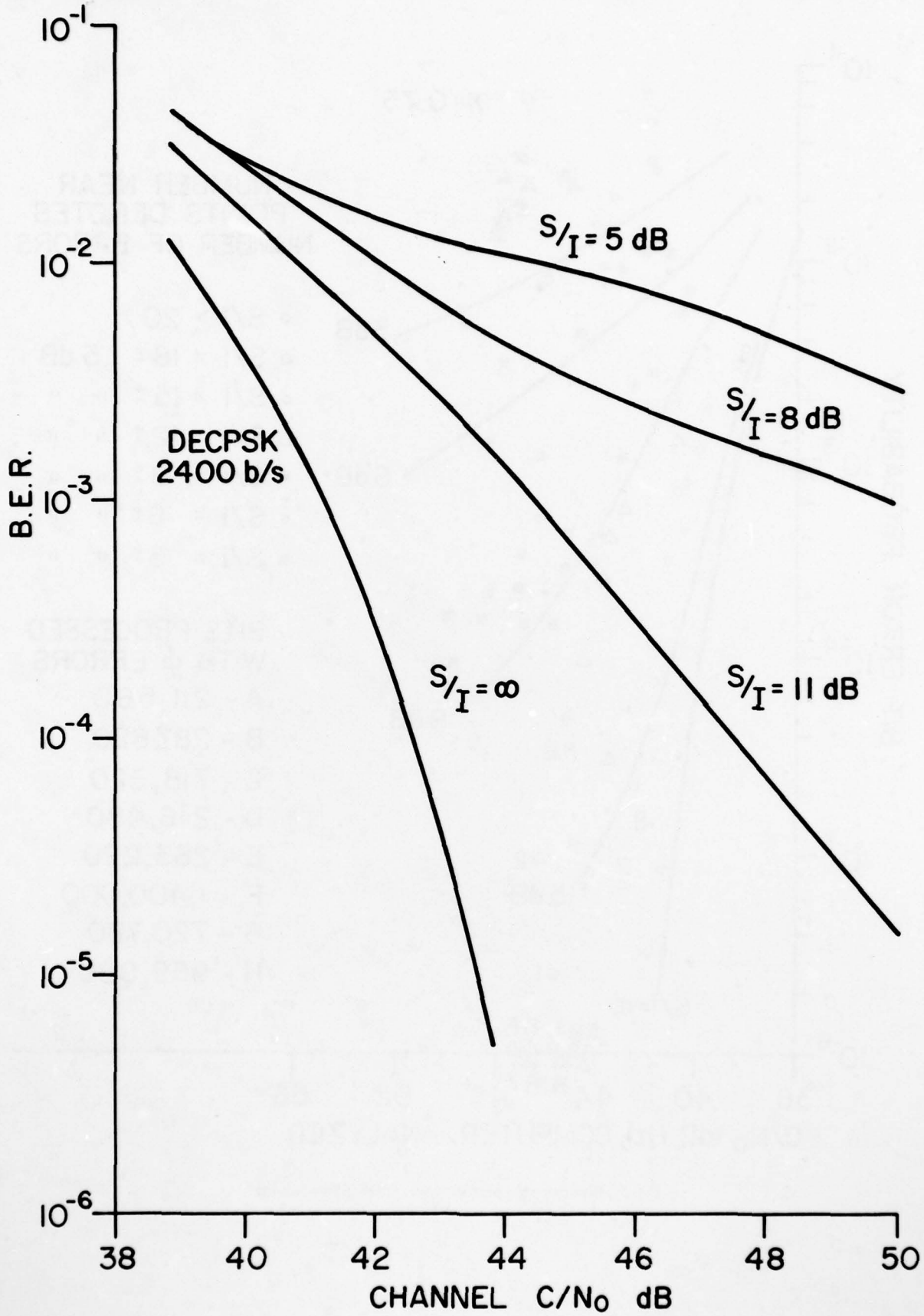


Fig.1 Effect of multipath on aeronautical satellite communications data transmission (measured results)

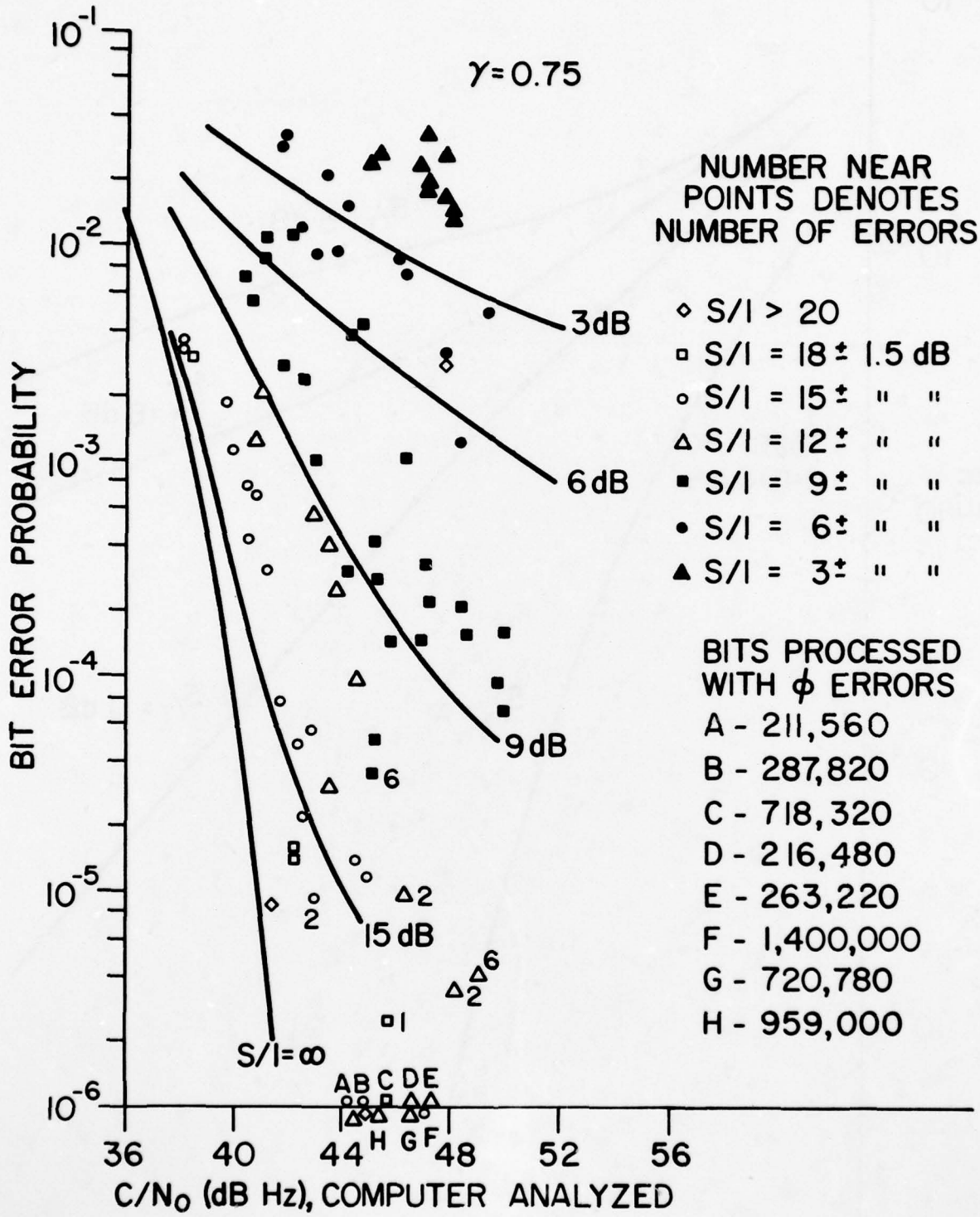


Fig.2 Bit-error-rate performance, CPSK demodulator

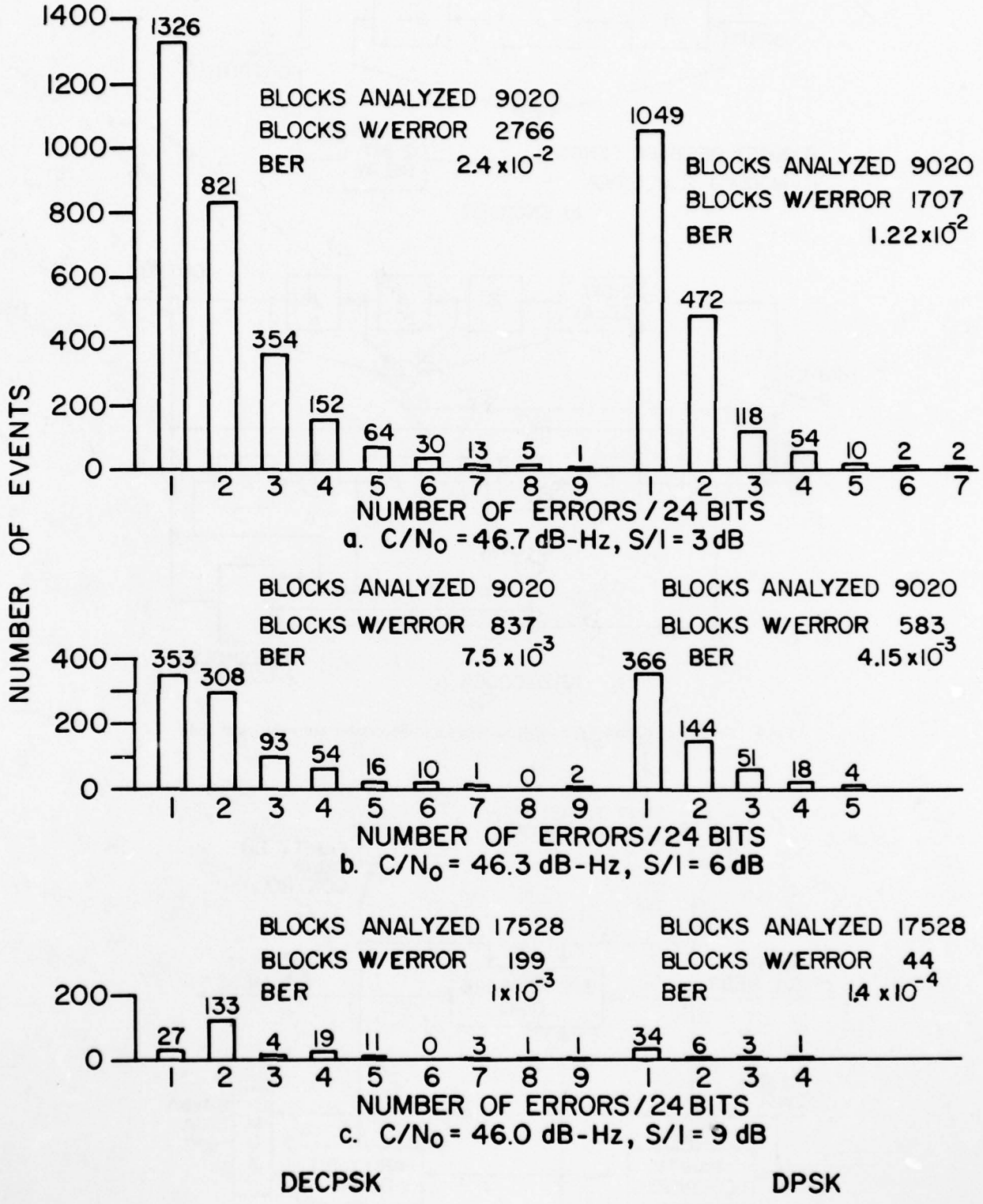


Fig.3 Block error histograms

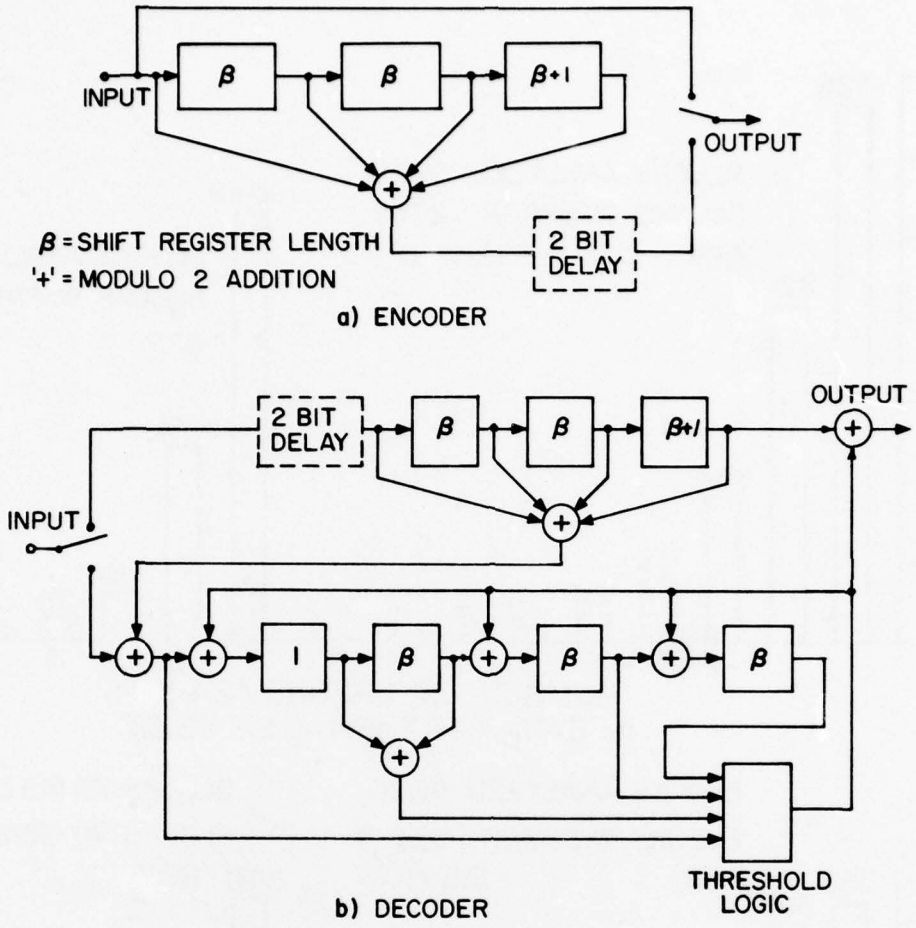


Fig.4 Coder and decoder for a diffuse threshold-decodable convolutional code

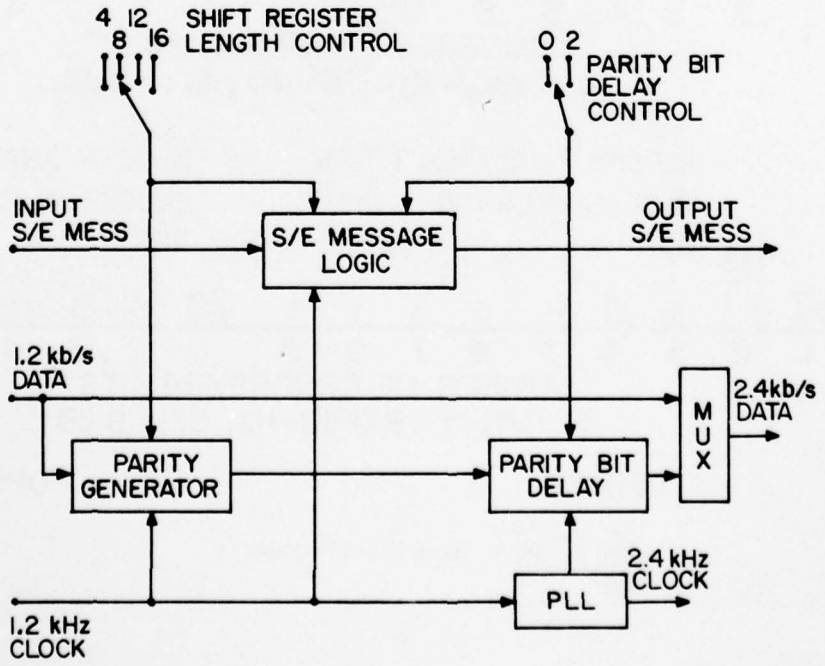


Fig.5 Encoder block diagram

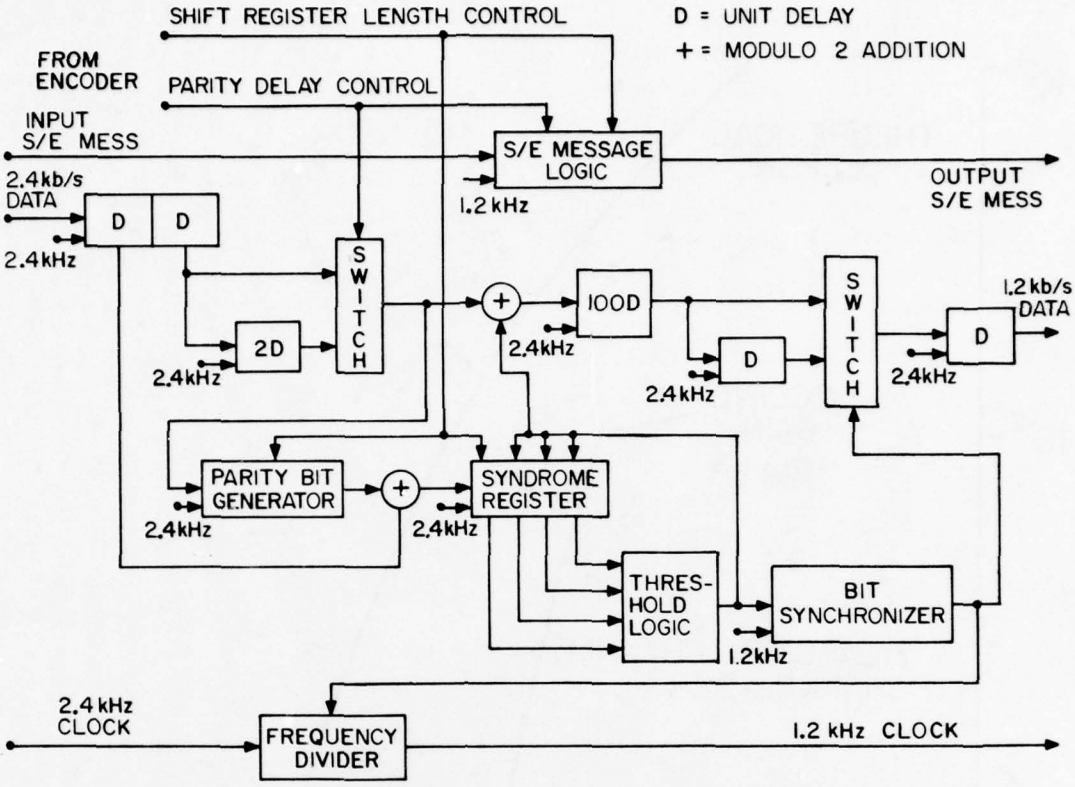


Fig.6 Decoder block diagram

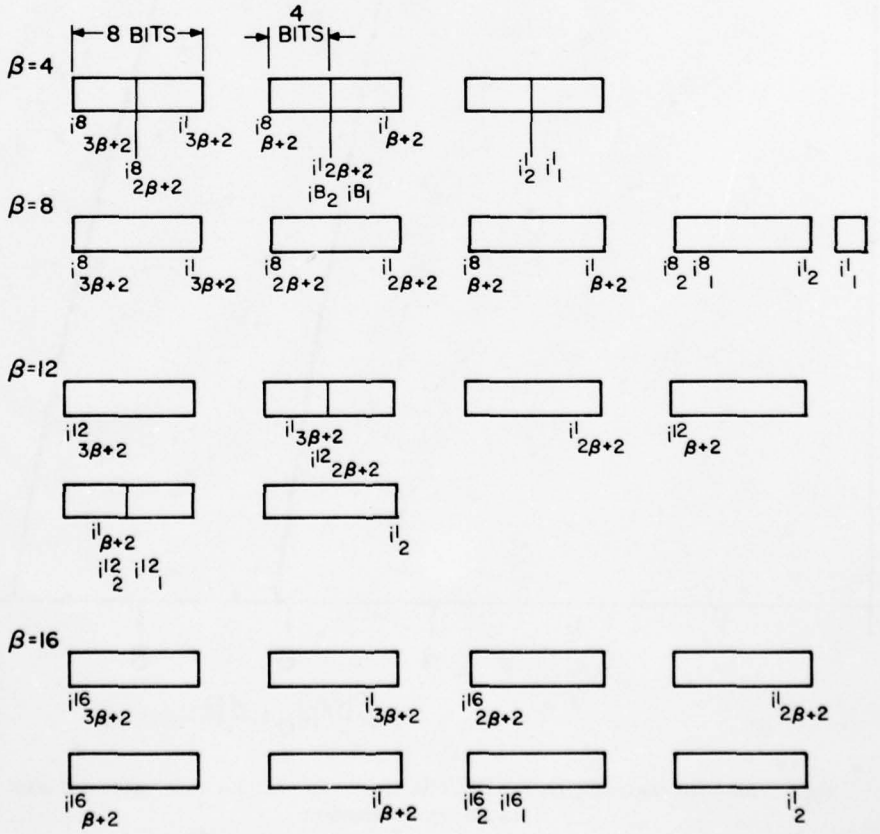


Fig.7 Encoder data structures

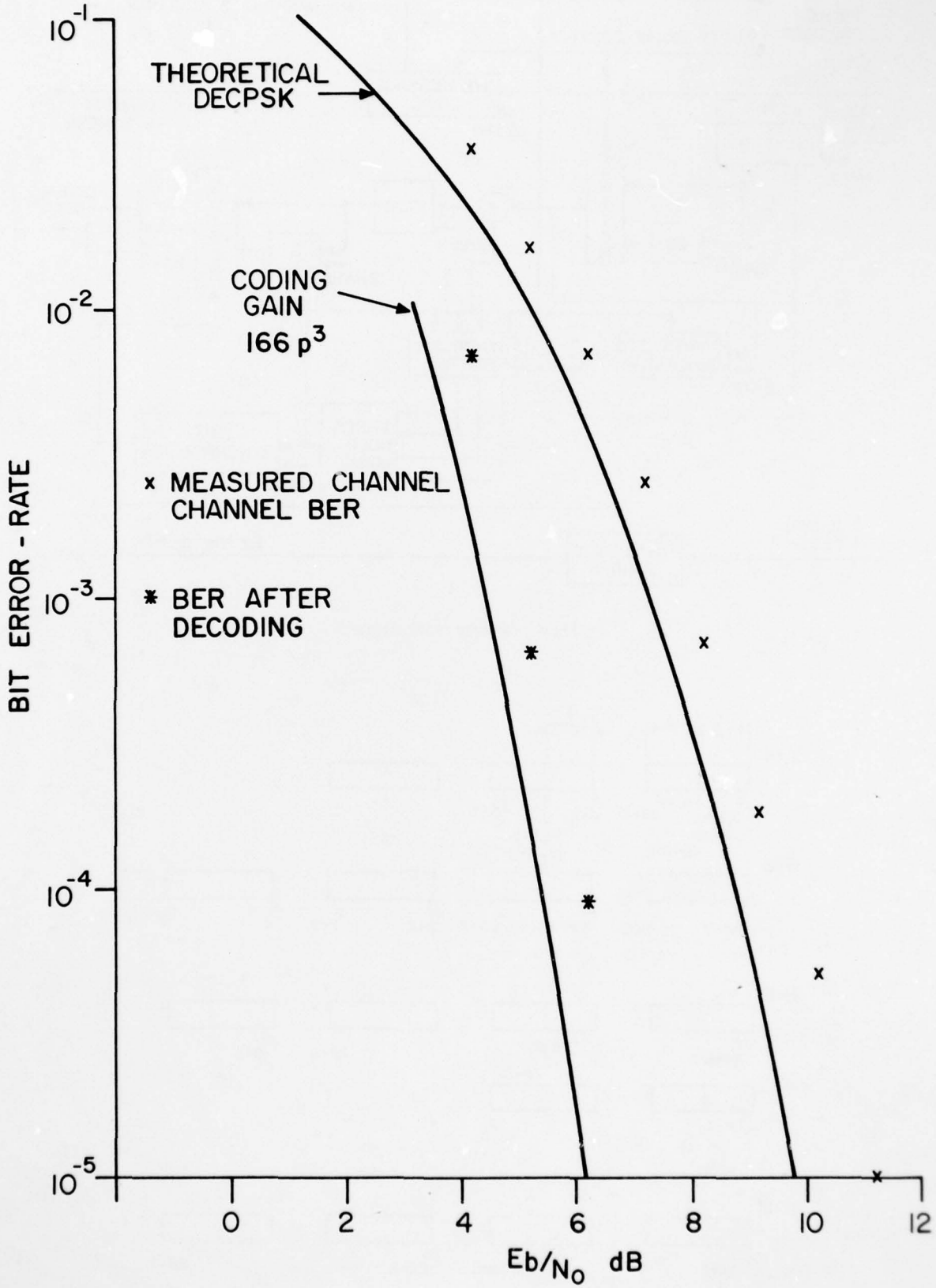
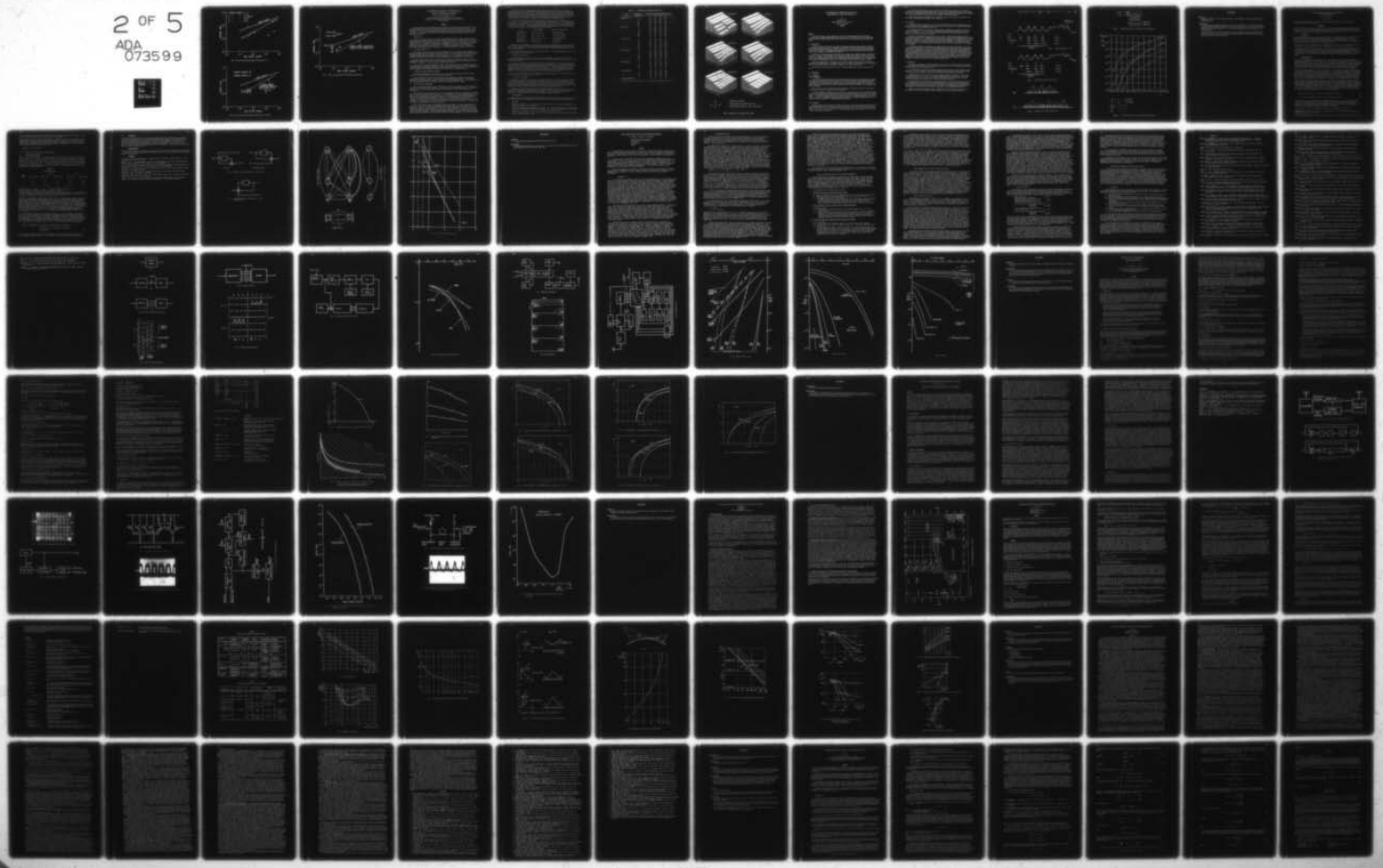


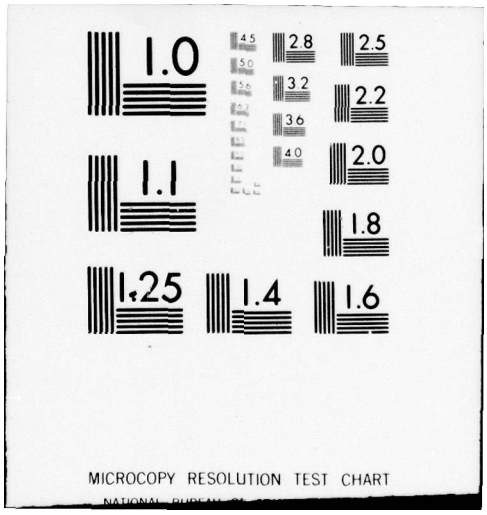
Fig.8 Bit-error-rate vs E_b/N_o for DECPSK transmission at 2.4 kb/s before and after forward error-correction

AD-A073 599 ADVISORY GROUP FOR AEROSPACE RESEARCH AND DEVELOPMENT--ETC F/G 17/2
DIGITAL COMMUNICATIONS IN AVIONICS.(U)
JUN 79 H LUEG

UNCLASSIFIED AGARD-CP-239 NL

2 OF 5
ADA 073599




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
MICROCOPY RESOLUTION TEST CHART

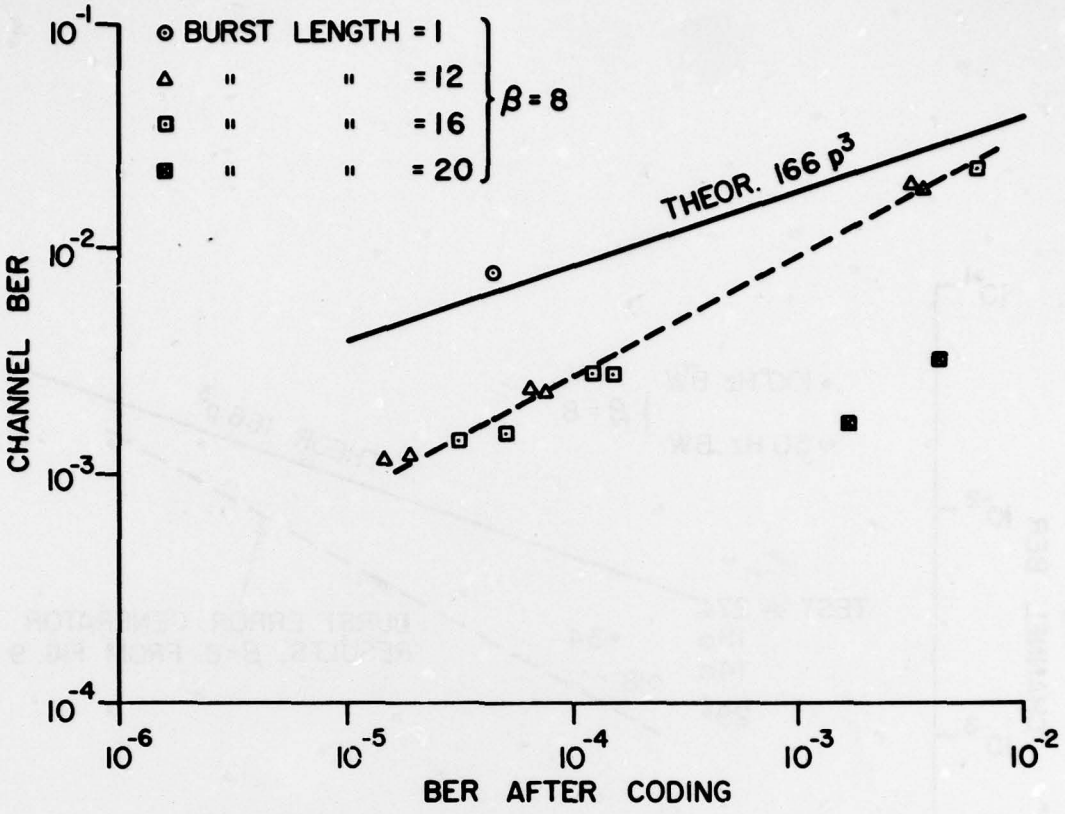


Fig.9 FEC Codec performance using burst error generator: $\beta = 8$

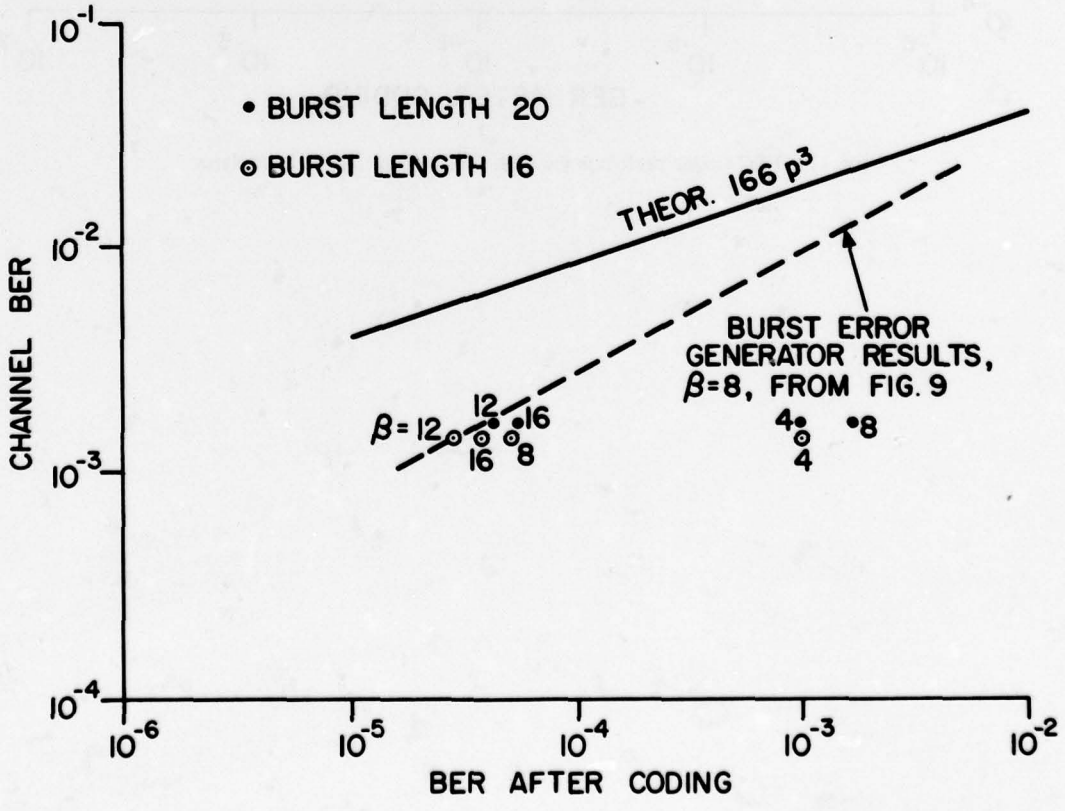


Fig.10 FEC Codec performance using burst error generator: variable β

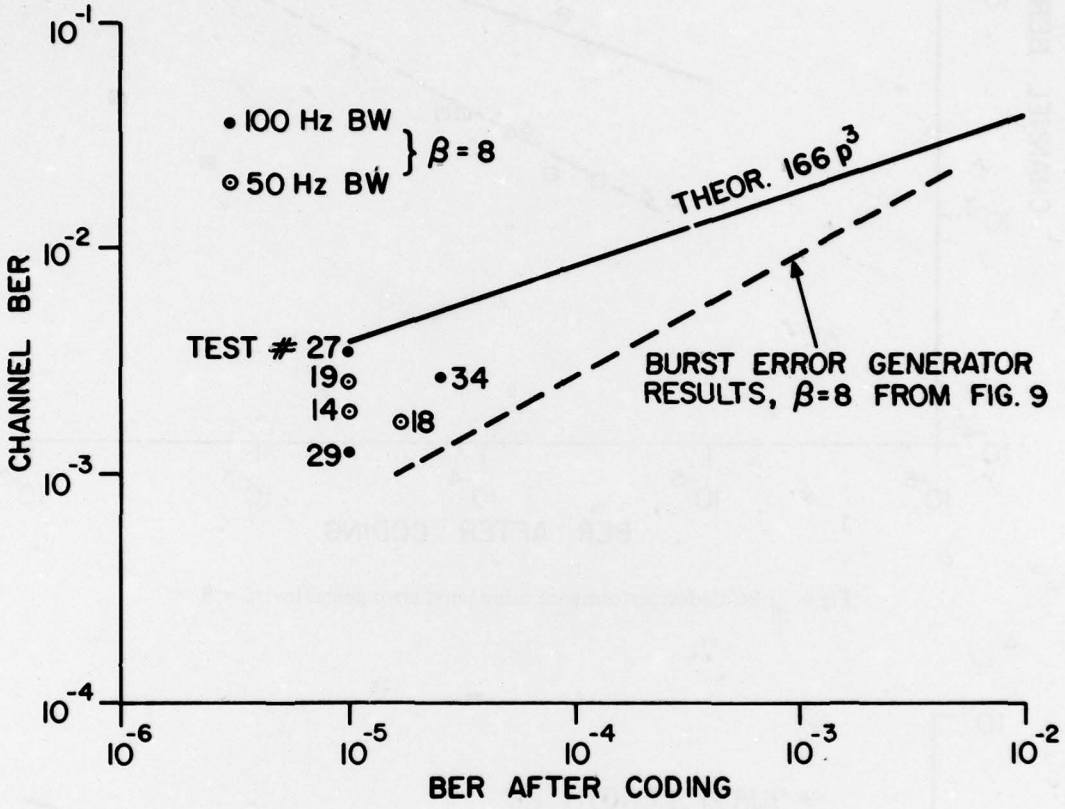


Fig.11 FEC Codec performance with TSC aerosat channel simulator

AN EXPERIMENTAL EVALUATION OF INTERLEAVED BLOCK CODING IN AERONAUTICAL HF CHANNELS

Brian Hillam and Geoffrey F. Gott

Department of Electrical Engineering and Electronics
University of Manchester Institute of Science and Technology
Manchester, UK.

SUMMARY

This paper presents results obtained from the application of forward error-correcting block codes to a 75 baud frequency shift keyed data transmission system, operating in aeronautical hf channels. Six different binary cyclic codes were tested, using detailed error structures obtained from point-to-point link tests. In an attempt to randomise error bursts, which arise because of fading and interference on the channel, interleaving of the codeword bits was employed, using several different degrees of interleaving for each code.

1. INTRODUCTION

The aim of this work has been to investigate the improvement offered by forward error-correcting block codes, when applied to slow rate fsk in aeromobile hf channels, and to examine the effect of several different degrees of interleaving for each code. Classifying the codes as (n,k,t), where n is the total number of bits in the codeword, k is the number of information bits, and t is the maximum of errors per codeword which can be corrected, the codes used were Hamming (7,4,1), BCH (15,7,2), Golay (23,12,3), BCH (31,16,3), BCH (63,30,6) and BCH (127,64,10). The results obtained add to those of previous experimental investigations into block coding at hf, undertaken by Brayer (1), Pierce (2), and others.

2. THE POINT-TO-POINT LINK

An fsk signal, keyed at 75 bauds, with 850 Hz frequency shift, was transmitted from Farnborough, Hampshire, England, and received at Wick, Caithness, Scotland - a south-north path, with a range of 800 km. The frequency chosen was the aeromobile allocation closest to the owf, and varied in the range 3-11 MHz, frequency changes being made according to published predictions. A bicone aerial was used at the transmitter, a monopole at the receiver, and the transmitted power was generally less than 50 watts.

The signal format started with a 44 bit sequence of reversals, to establish bit synchronisation, followed by a 33 bit frame code, which in turn was followed by two 1023 bit pseudo random sequences, which represented the message. This basic format was repeated continuously.

A total of about 40 hours of transmission was made over a one week period, including day and night operation, typical transmissions being about 2 hours. The received signals were recorded on magnetic tape, at audio frequency, for subsequent detection and coding evaluation in the laboratory. So far, about 16 hours of signals have been analysed in detail, and the results are presented in this paper.

3. DETECTION AND ERROR STRUCTURE EVALUATION

A conventional noncoherent fsk detector was used, which incorporated variable threshold decision circuitry to provide protection under selective fading conditions.

To study the effects of various error correcting codes, information was needed on the detailed distribution of errors with time. This was obtained by interfacing the fsk detector with a digital computer, and writing the precise details of the error structure onto magnetic tape. This error structure data was subsequently analysed by software to determine the improvement due to coding.

4. BLOCK CODES AND INTERLEAVING

With block coding, the data bits to be transmitted are divided into groups, and to each group are added parity check bits to form a codeword. The parity bits have a known relationship to the information bits, and it is by checking these relationships at the receiver, that errors may be detected and corrected (3).

Block codes are most effective when the errors occur randomly, but because of interference and fading, errors on hf data links tend to occur in bursts. Interleaving may be used to gain added protection in these conditions, especially when using simple codes. The interleaving and de-interleaving processes can most readily be visualised by considering the formation of two-dimensional matrices from the codeword vectors. To interleave, codeword bits are written to the rows of the matrix, and are read for transmission from the columns. Successive bits from each code word are thus separated by bits from other codewords. At the receiver, which uses an identical matrix, writing column by column, and reading row by row performs the complementary de-interleaving process.

If the total number of bits to be transmitted exceeds the capacity of a single interleaving matrix, the process of writing to rows and reading from columns is simply repeated until the input data is exhausted. However, if we impose the condition that the interleaving should not introduce a dead time, during which no data is to be transmitted, it can be seen that a restriction is placed on the number of rows, which defines the degree of interleaving, such that the available data fits exactly into an integer number of interleaving arrays. Thus for a short message comprised of say, 292 7-bit codewords, the only interleaving values permitted would be those which divide 292 without remainder, i.e. 1,2,4,73,146 and 292.

In processing the error structure data obtained from the point-to-point tests, the effect of interleaving was examined by de-interleaving the received data stream as if the required interleaving had been carried out at the transmitter. This was possible, since the data used in evaluating code performances was the error pattern introduced by the channel, which could be considered to have arisen from the transmission of all-zero codewords. During the analysis, the error structure data was held in the main memory of the computer in a one-dimensional array, and de-interleaving was carried out by indexing along this, rather than by writing to and reading from a two-dimensional array.

In order of increasing difficulty of implementation, the codes applied were Hamming (7,4,1), BCH (15,7,2), Golay (23,12,3), BCH (31,16,3), BCH (63,30,6) and BCH (127,64,10). For the application of these codes to the experimental error structure data, each message must consist of an integer number of codewords, and for each code, a message length was chosen to be as close as possible to 2046 bits, subject to this constraint. The message lengths resulting, and the degrees of interleaving permitted are tabulated below, where no interleaving is represented by degree 1.

Code	Message Length	Degrees of interleaving
Hamming (7,4,1)	$2044 = 7 \times 2^2 \times 73$	1,2,4,73,146,292
BCH (15,7,2)	$2040 = 15 \times 2^3 \times 17$	1,2,4,8,17,34,68,136
Golay (23,12,3)	$2024 = 23 \times 2^3 \times 11$	1,2,4,8,11,22,44,88
BCH (31,16,3)	$2046 = 31 \times 2 \times 3 \times 11$	1,2,3,6,11,22,33,66
BCH (63,30,6)	$2016 = 63 \times 2^5$	1,2,4,8,16,32
BCH (127,64,10)	$2032 = 127 \times 2^4$	1,2,4,8,16

To present a valid comparison of results obtained when applying different codes, the input error structure data (which was for 2046 bit messages) was truncated at the message length required for each code. Hence, regardless of the code used, all messages started immediately after the reception of the corresponding frame code.

All the results presented are for codes decoded by software. The decoding algorithms are not given, and are described in detail elsewhere (3).

5. PRESENTATION OF RESULTS

The results are given in Table 1, for all the codes investigated. They are for 16 hours of transmissions (corresponding to about 4 million bits), and comprise approximately equal contributions from day and night operation.

If coding was not used, the transmitted information would immediately follow the frame code of the signal format, and would therefore be represented by the first fraction, k/n, of the truncated message. Since all the codes used were approximately half redundant, the error rate of the first 1024 bits of each message has been used to define the uncoded error rate in Table 1. Strictly speaking, this 1024 bit information sequence is correct only for the BCH (127,64,10) code. All the codes used have slightly different values of k/n, and consequently the uncoded information sequence would vary from 952 bits for the BCH (15,7,2) code, to 1056 bits for the Golay (23,12,3) and BCH (31,16,3) codes.

Fig. 1 shows the results of Table 1 presented on three-dimensional graphs. The improvement as the degree of interleaving increases may be clearly seen.

6. CONCLUSIONS AND COMMENTS

The decision on which code to use must be a compromise between the error rate improvement obtained, and the decoder complexity. From the results presented, the Golay (23,12,3) code is preferred, since it gave almost as many error free messages as the BCH (127,64,10) code (at their maximum interleaving levels for continuous transmission), and has the advantage of being simpler to instrument. Indeed, the decoding of the Golay code has been conveniently achieved in real time for data rates in excess of 75 bits/second, using a standard microprocessor, and the error trapping decoding method (3).

However, coding alone is probably insufficient to reliably achieve low error rates at hf. In particular, the effectiveness of coding may be considerably enhanced if severe interference from other hf users can be avoided, by making adaptive frequency changes. This technique has been discussed previously (4), and serves to reduce the error rate before the decoding operation.

7. ACKNOWLEDGEMENTS

Thanks are due to the Procurement Executive, Ministry of Defence, who supported this project, and the Department of Electrical Engineering and Electronics, UMIST, for providing the laboratory facilities.

8. REFERENCES

1. BRAYER, K., CARDINALE, O., 'Evaluation of error correction block encoding for high-speed hf data', IEEE Transactions, Vol.COM-15, June 1967.

2. PIERCE, A.W., BARROW, B.B., GOLDBERG, B. and TUCKER, T.R., 'Effective application of forward-acting error-control coding to multichannel hf data modems', IEEE Transactions, Vol.COM-18, No.4, Aug. 1970.

3. PETERSON, W.W. and WELDON, E.J., 'Error correcting codes", MIT Press, 1972.

4. GOTT, G.F. and HILLAM, B., 'Improvements to fsk data transmission', AGARD meeting on 'Radio systems and the ionosphere', Athens, 1975.

Table 1. A comparison of interleaved block codes.

Code	Degree of Interleaving	Percentage of messages with error rate ≤ 0	0.5%	1%	5%
Hamming (7,4,1)	1	19	41	55	82
	2	21	43	57	83
	4	27	48	61	83
	73	44	66	73	85
	146	51	69	76	86
	292	53	71	77	86
BCH (15,7,2)	1	29	46	61	83
	2	37	53	64	84
	4	48	59	67	84
	8	53	64	70	84
	17	59	67	73	84
	34	64	71	76	85
	68	67	75	79	86
	136	70	79	81	87
Golay (23,12,3)	1	35	51	62	83
	2	44	57	66	83
	4	54	63	69	84
	8	60	68	73	84
	11	64	70	74	84
	22	68	74	77	85
	44	72	78	80	86
	88	76	81	82	88
BCH (31,16,3)	1	34	46	58	82
	2	41	54	63	83
	3	44	61	67	84
	6	53	63	69	83
	11	60	68	72	83
	22	66	72	76	85
	33	69	74	77	85
	66	73	78	80	86
BCH (63,30,6)	1	51	59	67	84
	2	58	62	69	83
	4	65	67	72	84
	8	70	73	77	84
	16	76	78	80	85
	32	79	81	83	87
BCH (127,64,10)	1	58	61	66	83
	2	64	66	70	83
	4	70	72	75	83
	8	76	77	79	84
	16	80	81	82	86
Uncoded Performance		12	37	52	81

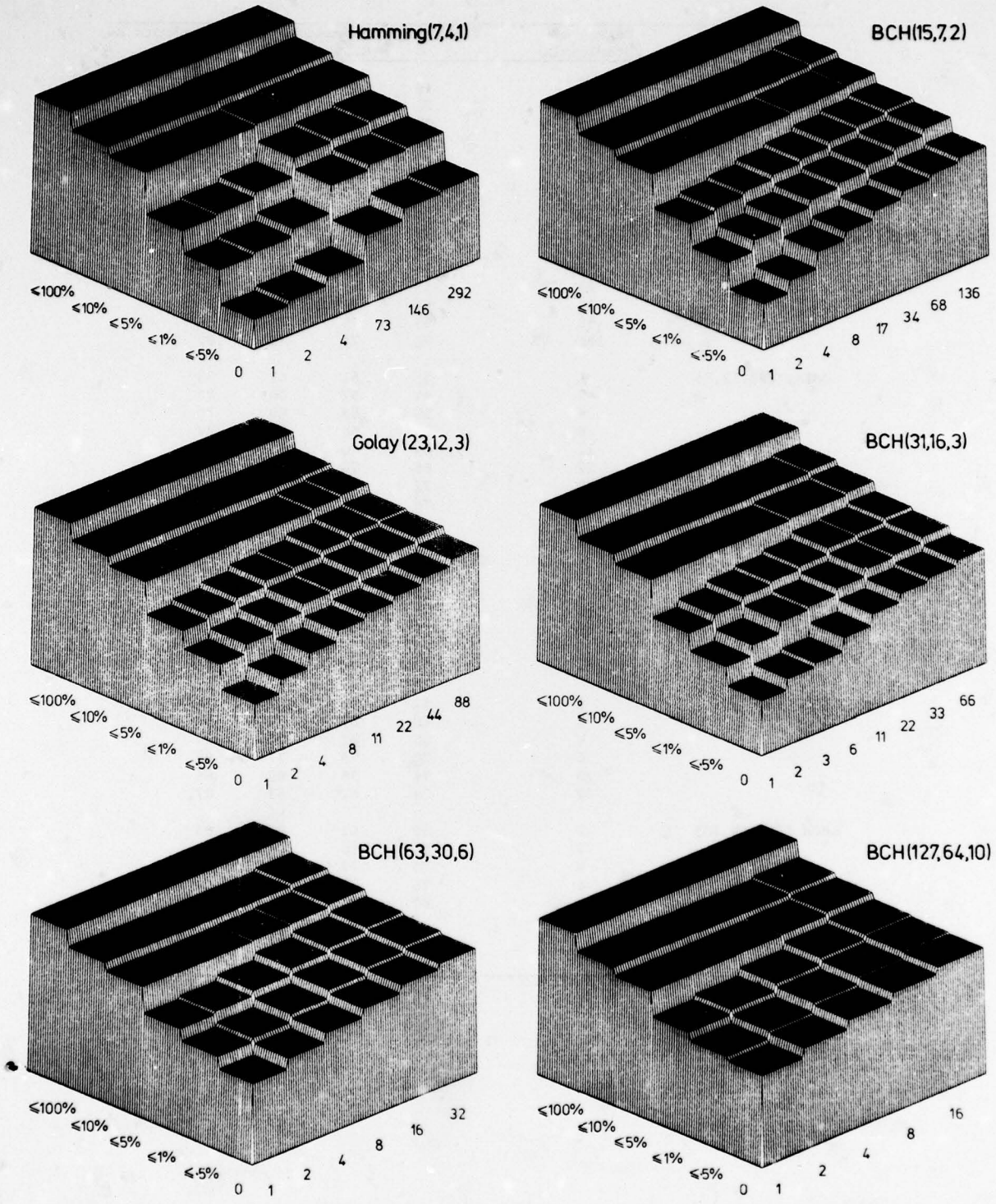


Fig.1 A comparison of interleaved block codes

AN ASYNCHRONOUS DATA TRANSMISSION SYSTEM WITH LOW ERROR PROBABILITY FOR THE SETAC LANDING AID

by
Wolfgang Beier
Standard Elektrik Lorenz AG (ITT)
7000 Stuttgart 40
Germany

Summary

This paper describes a serial code which is used in the TACAN compatible landing aid SETAC. This code uses the randomly spaced TACAN pulses as data carrier. The special need in this application is that the error correcting device must correct mainly for lost bits besides correcting wrong information.

1. Introduction

The SETAC landing aid transmits serial data from the ground station to the aircraft. The TACAN pulse groups are used as data carrier. Depending on the mode of operation TACAN or SETAC two to three pulses are transmitted respectively in a group. The third pulse in the group is amplitude modulated by 15 Hz and 135 Hz, when it is demodulated it gives high precision bearing information near the runway center line, in addition to the normal TACAN function.

The position of the third pulse in the group can be changed to two different spaces. So each pulse group can carry one bit of information. The position shift of the third pulse is used for data transmission. This channel transmits 2700 bits of information on the average, however, the clock rate varies so that the time space between two bits being transmitted changes from about 100 µs to 1 ms. Due to this highly randomly distributed bit positions it is not possible to synchronize the receiver on board.

The IFF transponder works in the same rf band as TACAN systems. Each time the IFF transponder is transmitting the TACAN receiver is disabled and if there is a TACAN or SETAC pulse group it is suppressed and the bit of data which was carried upon this group is lost in the serial string.

Hence there is a need for some sort of error correcting scheme which is able to correct the lost bit.

2. Description

2.1 Hardware

2.1.1 Transmitter

The SETAC ground station transmits on the average 2700 pulse triplets which can have two different formations. The first two pulses in both states are in accordance with the TACAN standard and the third pulse changes its relative position in respect to the first two pulses (Fig. 1). For each SETAC pulse which means a pulse triplet the data modulating device, called SETAC DATA TERMINAL, gets a trigger. The output of this device controls with a logical signal the type of SETAC pulse which is transmitted by the SETAC transponder.

Data is sent in a serial string of 6 bit ASCII characters. Included are parity bits and synchronisation pattern to enable a receiving device to synchronize.

If the data terminal is sending a logical "0" first it changes the pulse type by negating the control input of the transponder and then counts the trigger of the sent SETAC pulse up to five. To send another logical "0" again the terminal changes the pulse type and counts five transmitting triggers. A following logical "1" is sent by changing the pulse type again and counting 13 transmitting triggers. So the information is coded into the number of equal pulse types transmitted in a sequence. Hence a logical "0" is coded as five and a logical "1" as thirteen equal pulses (Fig. 2).

2.1.2 Receiver

On board of the aircraft receiving the SETAC signal there is a decoding and correcting device within the SETAC attachment. This device gets two triggers as an input signal. One trigger for each received type of SETAC pulse (Fig. 3). The output of this correcting device is a clock and data for each received logical bit i.e. one clock and a "0" data for 2 to 6 equal pulse types and one clock and a "1" for 7 to 13 equal pulse types in a sequence.

So if from the ground station a set of 5 pulses is transmitted for a logical "0" it results to a clock and a "0" data at the receiver output even if only 2, 3, or 4 pulses have arrived. If due to some noise 6 pulses are detected it will result also in a correct output. If only one single pulse is received in a group of opposite pulse types than it will be ignored in the decoding device.

If more than 6 pulses of the same type are received in order it will result in a clock and a "1" data output. Single pulses are ignored in this string also.

2.2 Theory

The major problem in this data link is to be able to detect and count each transmitted pulse. So the first task of an error correcting device is to ensure the right bit count. A second step then would be the correction of the information of each bit.

At the maximum rate of IFF replies the probability that a TACAN pulse is suppressed is 0.044 for each single pulse. So if a set of 5 pulses is transmitted the probability that only one pulse of it arrives in the decoder is $18.7 \cdot 10^{-6}$ or once in about 100 seconds (at 2700 SETAC pulses per second) (Fig. 4). This error rate would be acceptable.

Each bit of information either a logical "0" or a logical "1" must be coded into more than one pulse to ensure that at least a fragment of each logical bit is received. This fragment must carry enough information to detect it as a single logical bit and to decode its logical state to either a "1" or a "0". Such a set of pulses is a string of 5 equal pulses. There each combination of 3 suppressions gives the same remaining pattern. But the beginning and the end can only be detected if the pattern before and after this group is of the opposite type. So the correct logic bit count is achieved but no information can be decoded.

The problem to modulate an information to a given clock in this case is the same as it is with a normal fixed frequency clock. There are several methods available. The method with the minimum required pulses which is applicable in this case is the Delta Distance Method. That means that the duration between two pulse type changes is used as modulation. The logical "0" is coded in the above mentioned pulse type change after five SETAC pulses. In the other logical state the number of SETAC pulses between two pulse type changes must be higher. The minimum pulse count for a logical "1" is calculated so that the probability that its pulses are suppressed to 7 pulses is about the same as that 5 pulses are suppressed to 2. This requires 12 pulses for a logical "1". Due to the not purely randomly distributed TACAN and IFF pulses 13 pulses are used. So five SETAC pulses are used to transmit a logical "0" and 13 pulses for a logical "1" that is on the average 9 pulses per bit.

3. Conclusion

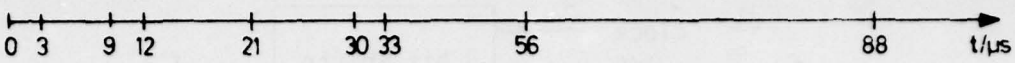
The above described code is specially designed for application in the SETAC systems. Its features are optimized for the expected error rate and types of errors as well as for the data transmission rate and the accepted error rate after correction.

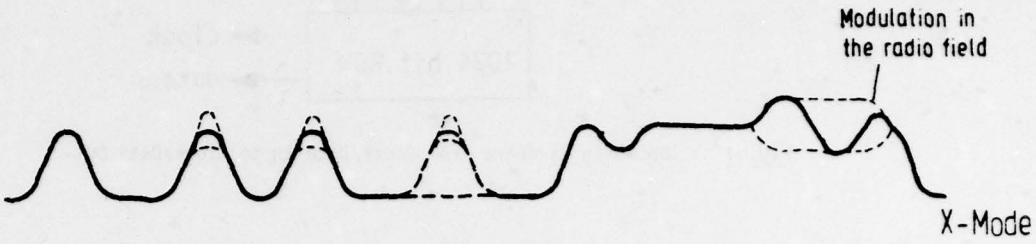
This code requires nine SETAC pulses on the average for one bit and is able to transmit 50 ASCII characters (6 bit) per second. In this data transmission channel the data correction has to account for missing pulses or additional pulses with the limitations as described in the following.

Suppressed pulses due to the IFF replies or any other reason can be corrected: 1.) If not more than every other pulse is lost, and 2.) in a normal randomly suppressed case not more than one uncorrectable error occurs in a 100 sec time interval at maximum IFF rate. At normal IFF reply rate the mean time between uncorrectable error is usually much higher than 100 sec.

Furthermore errors are corrected if more or erronous pulses are received, provided if the errors do not occur more often than once in every nine SETAC data bit. If an error cannot be corrected it can be detected with a parity word which is sent after each line of characters. An advantage of this type of code is the easy method of correcting for the type of errors described above. The hardware requirement consists of only one shift register (9 bit) and a 1 kbit ROM (Read Only Memory).

Hence with a minimum of hardware on board a sufficient error free data link is established to provide the aircraft with the information that is necessary for a safe instrument approach.





Pulse	1	2	3_0	3_1	4
Carrier	SROB	SROB	Sector	Sector	SETAC-E
Side frequency	—	SROB	Sector	Sector	SETAC-E
Δ f		15 Hz	15÷135 Hz	15÷135 Hz	100 kHz

SETAC-A

|Ref.- |Carrier-|Measuring- |Part



Pulse	1	3_1	3_0	2	4
Carrier	SROB	Sector	Sector	SROB	SETAC-E
Side frequency	—	Sector	Sector	SROB	SETAC-E
Δ f		15÷135 Hz	15÷135 Hz	15 Hz	100 kHz

SETAC-A

Fig. 1 SETAC Signal Format of Pulse Quartet

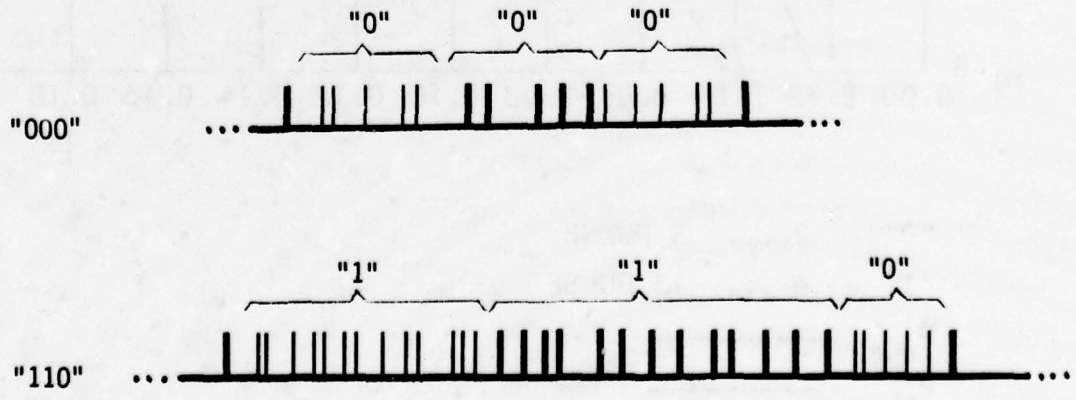


Fig. 2 Example of Bit Coding at Pulse Level

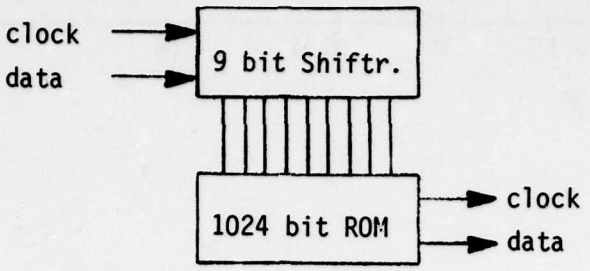
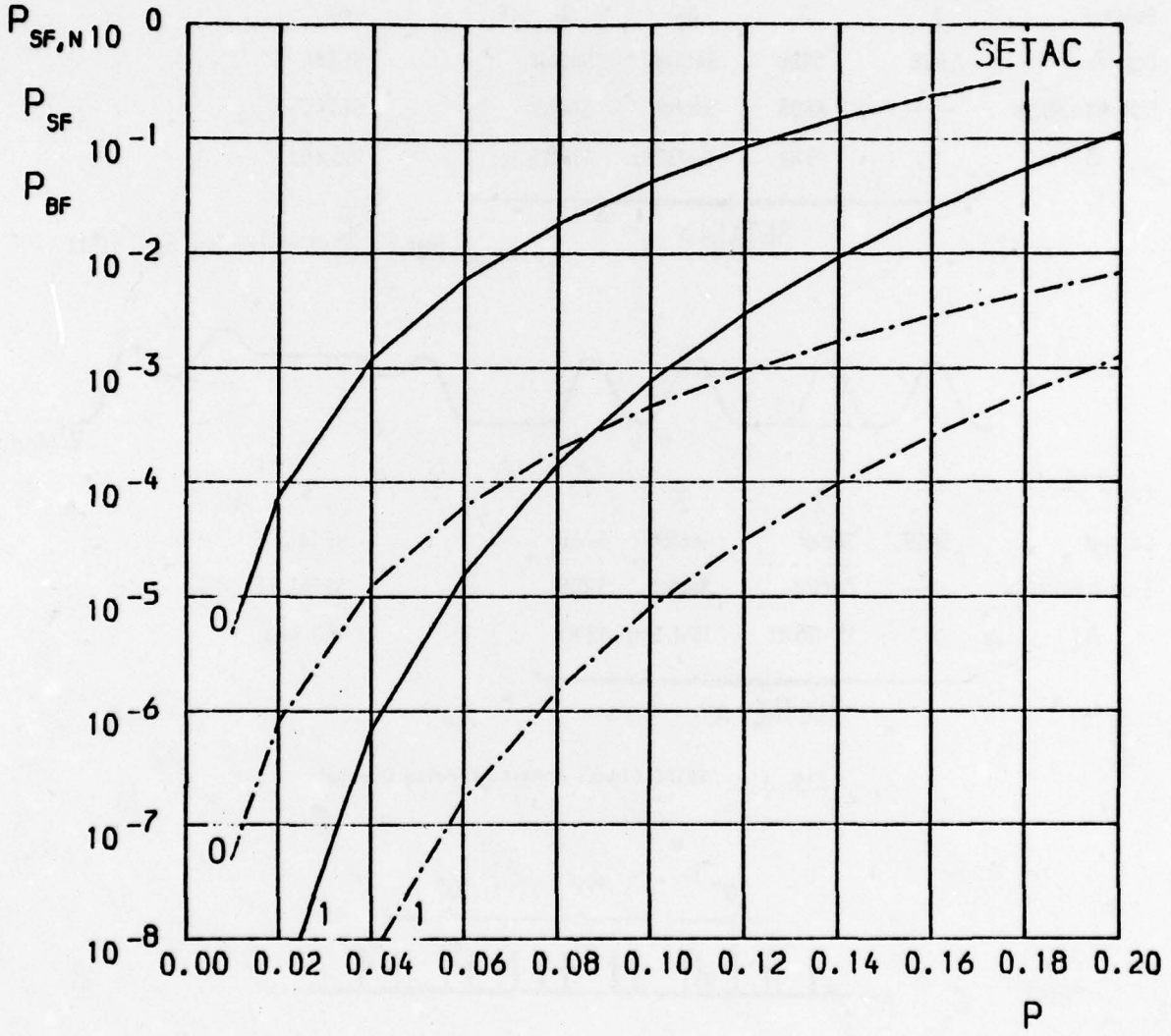


Fig. 3 Decoding Hardware from Clock/Data In to Clock/Data Out



"0" = 2 ... 5 PULSE

"1" = 6 ... 13 PULSE

$P_{SF,N}$ ————————

P_{SF} – – – – – – –

P_{BF} —·—·—·—

Fig. 4 Bit Error Probability P_{BF} v. Suppression Probability P

DISCUSSION

W.Skupin, Ge

Has the type of coding – 5 and 13 pulses of the same type – been optimized? And if so, what are the desired optimal parameters?

Author's Reply

The optimization of the number of equal pulses – now 5 and 13 for the '0' and 1 – was done so that with the expected suppression probability (0.044 with max. IFF rate) the remaining error rate is acceptable. In the SETAC application it is with this 5/13 groups $18.7 \cdot 10^{-6}$. If the expected suppression rate is lower, then may be a 3/5 group is sufficient.

The number of pulses for the "0" has to be chosen so that at a maximum of suppression 2 pulses remain. The number of pulses for a "1" must be high enough that at maximum suppression the remaining number of pulses is greater than the normal number for a logical "0".

ON THE PERFORMANCE OF A MAXIMUM LIKELIHOOD DECODER FOR CONVOLUTIONAL CODES

J.P.M. Schalkwijk

March 15, 1978.

The author is with the Department of Electrical Engineering, Eindhoven University of Technology, P.O. Box 513, Eindhoven, The Netherlands.

SUMMARY

Maximum likelihood (ML) decoding of short constraint length convolutional codes became feasible with the invention of the Viterbi decoder. Several authors have since upper bounded the performance of ML decoders. No one has as yet given an exact evaluation of the error rate. This paper describes a method to compute the event error probability of an ML decoder for convolutional codes.

I. INTRODUCTION

Maximum likelihood (ML) decoding of short constraint length convolutional codes became feasible with the invention of the Viterbi decoder. Several authors (Viterbi, A.J., 1971; Meeberg, v.d., L., 1974; Post, K.A., 1977) have since upper bounded the performance of ML decoders. No one has as yet given an exact evaluation of the error rate. This paper describes a method to compute the event error probability (Viterbi, A.J., 1971) of an ML decoder for convolutional codes.

The concept of event error probability will be elaborated on in Section II. It will be explained why to us this appears to be a more meaningful measure of performance than is the bit error probability.

The performance evaluation will be illustrated using the binary rate $\frac{1}{2}$ convolutional code generated by the simple constraint length $\nu=1$ encoder of Fig. 1. Simular calculations have been carried through for the code generated by the standard constraint length $\nu=2$ encoder having connection polynomials $1 + D^2$, and $1 + D + D^2$. However, like the practical implementation of the ML decoder becomes infeasible with increasing constraint length ν, so do the calculations necessary to exactly evaluate the performance of such a decoder.

II. SYNDROME DECODING

To evaluate the performance of convolutional codes with ML decoding we consider the syndrome decoder implementation (Schalkwijk, J.P.M., and Vinck, A.J., 1975; 1976) of the ML decoding algorithm. The classical Viterbi decoder (Viterbi, A.J., 1971) recursively finds the trellis path (codeword) closest to the received data. Given the received data $x(D) = m(D)+n_1(D)$, $y(D) = (1+D)m(D) + n_2(D)$, see Fig. 1, the syndrome decoder first forms a syndrome $w(D)$, instead. A recursive algorithm like Viterbi's is used to determine a noise sequence pair $[\hat{n}_1(D), \hat{n}_2(D)]$ of minimum Hamming weight that can be a possible cause of this syndrome. Given this estimate of the noise, one derives an estimate, $\hat{m}(D)$, of the original message sequence.

It is easily verified (Schalkwijk, J.P.M., and Vinck, A.J., 1975; 1976) that the syndrome sequence $w(D)$ is determined by the channel noise pair $[n_1(D), n_2(D)]$ only, i.e. the syndrome is independent of the message sequence $m(D)$. Hence, in Fig. 2, only the noise digits $[n_1, n_2]$, and the syndrome digit w are indicated. Each input $[n_1, n_2]$ to the syndrome-former causes a state transition (the state of the syndrome-former can be equated with the content of its memory cell), and an associated output w. The upper left hand part of Fig. 3 gives the state-diagram of the syndrome-former of Fig. 2. Indicated along the edges are the input digits $[n_1, n_2]$, and syndrome outputs w=0, and w=1 correspond to solid, and dashed edges, respectively.

Having observed a sequence of syndrome digits, one is to determine a corresponding sequence of state-transitions (solid, and dashed edges for syndrome digits 0, and 1, respecively) for which the accumulated Hamming weight of the associated noise digits is minimal. To find a proper sequence of state-transitions Viterbi introduces (Viterbi, A.J., 1971) a "metric-function". A metric-function is defined as a nonnegative integer-valued function of the states. Starting with a metric function f_0, given a syndrome sequence $w_1, w_2, w_3, \ldots$ new metric functions $f_1, f_2, f_3, \ldots$ can be computed recursively by means of the metric equations (Schalkwijk, J.P.M. and Vinck, A.J., 1975; 1976 ; Schalkwijk, J.P.M., Vinck, A.J., and Post, K.A., 1978). For the state-diagram of Fig. 3 the metric equations are

$$f_k(0) = \bar{w}_k \min [f_{k-1}(0) \ , \ f_{k-1}(1)+1] + w_k \min [f_{k-1}(0) \ , \ f_{k-1}(1) \ ] \tag{1a}$$

$$f_k(1) = \bar{w}_k \min [f_{k-1}(0)+2, \ f_{k-1}(1)+1] + w_k \min [f_{k-1}(0)+1, \ f_{k-1}(1)+2] \ , \tag{1b}$$

where $\bar{w}$ is the modulo 2 complement of w. The states (preimages) $s_k(0)$, and $s_k(1)$ associated with the minimum within the relevant pair of square brackets in (1a), and (1b), respectively, are called the "survivors". In the case where one has more candidates for survivor, the choice is determined by considerations regarding the complexity of the decoder (Schalkwijk, J.P.M., Vinck, A.J., and Post, K.A., 1978). The survivor sequence, finally, determines the noise estimate.

We are now ready to discuss our performance criterion, i.e. the "event error probability".

DEFINITION: An "error event" is an excursion from the decoded survivor sequence away from the true state sequence traced by the syndrome-former.

This research was supported by the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

It is easily seen that this definition of an error event coincides with the usual one (Forney, Jr., G.D., 1970). In either case the error event corresponds to a short codeword.

A final remark concerns the relation between event error probability and bit error probability. Observe that a single error event can cause several bit errors. Hence, bit errors occur in bursts and not independently as the concept of bit error probability suggests. Furthermore, in practice the information is sent in blocks, and a meaningful quantity is the block error probability. Given the block length N, the block error probability, P[block], is given by

$$P[\text{block}] = 1 - (1-P[\text{event}])^N \quad , \tag{2}$$

i.e. it follows directly from the event error probability!

III. METRIC x STATE-DIAGRAM

Using (1) one can construct Table I. The second column gives the values $f_{k-1}(0)$, and $f_{k-1}(1)$ of the old metric function $f_{k-1}(.)$. The columns further to the right list the values $f_k(0)$, and $f_k(1)$ of new metric functions $f_k(.)$, and the values $s_k(0)$, and $s_k(1)$ of the survivor function $s_k(.)$ for both the case where $w_k = 0$, and for the case where $w_k = 1$. If there are more candidates $s_k(0)$, or $s_k(1)$ for survivor, these candidates are put between parentheses in the survivor columns. In the remainder we arbitrarity select the "0" survivor in case of ambiguity.

TABLE I

METRIC-TRANSITIONS

row number	$f_{k-1}(0)$, $f_{k-1}(1)$	$w_k = 0$: $s_k(0)$, $s_k(1)$	$w_k = 0$: $f_k(0)$, $f_k(1)$	$w_k = 1$: $s_k(0)$, $s_k(1)$	$w_k = 1$: $f_k(0)$, $f_k(1)$
0	0 , 2	0 , 0	0 , 2	0 , 0	0 , 0
1	0 , 0	0 , 1	0 , 1	0 , 1	0 , 1
2	0 , 1	0 , (0 , 1)	0 , 2	(0 , 1) , 0	0 , 0

Note that there are three metric functions, i.e. $f^{(0)}(.) \triangleq \{f^{(0)}(0) = 0, f^{(0)}(1) = 2\}$, $f^{(1)}(.) \triangleq \{f^{(1)}(0) = 0, f^{(1)}(1) = 0\}$, and $f^{(2)}(.) \triangleq \{f^{(2)}(0) = 0, f^{(2)}(1) = 1\}$.
The "metric-diagram", see lower right hand part of Fig. 3, depicts the possible metric function transitions, where a solid edge again corresponds to a syndrome digit w = 0, and a dashed edge to a syndrome ditit w = 1. For example, a syndrome digit w=1 causes the transition $f^{(0)}(.) \rightarrow f^{(1)}(.)$.

The "metric x state-diagram", see upper right hand part of Fig. 3, is the cartesian product of the metric-diagram, and the state-diagram. This diagram, finally, allows to precisely describe the operation of the core (estimator) part of the syndrome decoder. The survivor selection $s_k(0)$, $s_k(1)$ according to Table I implies that the decoder can trace the subgraph with double edges of the metric x state-diagram only. For example, assume that in the metric x state-diagram we are in state 0 0 , compare Fig. 3. A syndrome-former, see Fig. 2, input $[n_1, n_2] = [1,0]$ causes the transition 0 0 → 1 1 in the metric x state-diagram. However, as we only observe a syndrome digit w = 1, all we can conclude is a move to the right in the metric x state-diagram to either state 1 0 , or to state 1 1 . Let the next syndrome-former input be $[n_1, n_2] = [0,0]$. The resulting syndrome output digit is again w = 1. Now, as in the subgraph with double edges only state 1 1 allows a syndrome digit w = 1, i.e. a dashed edge, the ambiguity between the states 1 0 , and 1 1 is resolved!

For a binary symmetric channel (BSC) with crossover probability p, $0<p<\frac{1}{2}$, the steady state probabilities for the metric x state-diagram can be easily computed. For example, for p = 0.1 we have P(00) = 0.655, P(01) = 0.010, P(10) = 0.095, P(11) = 0.075, P(20) = 0.150, and P(10) = 0.015. The event error probability is now the probability of leaving the subgraph with double edges in the metric x state-diagram of Fig. 3. Let Q(10) and Q(11) be the probabilities of reaching the states 1 0, and 1 1, respectively, via a double edge of the subgraph, and let $p_w(ij,k\ell)$ be the transition probability of the solid, w = 0, or dashed, w = 1, edge from state ij to state kℓ. $0 \leq i,k \leq 2$, $0 \leq j,\ell \leq 1$. Then the event error probability is

$$\begin{aligned} P[\text{event}] = {} & P(00)p_0(00,01) + P(10)p_0(10,21) + Q(10)\,[p_1(10,20) + p_1(10,21)] + \\ & + P(11)p_1(11,21) + Q(11)\,[p_0(11,20) + p_0(11,21)] + \\ & + P(20)p_0(20,01) \end{aligned} \tag{3}$$

Fig. 4 is a plot of P[event] versus the transition probability p of the BSC. Note that the Post-, and v.d. Meeberg upper bounds (Meeberg, v.d., L., 1974; Post, K.A., 1977) are tight for small values of p.

IV CONCLUSIONS

A method of evaluating the event error probability for ML decoding of convolutional codes is presented. To illustrate this method calculations are carried through for the simple rate ½ convolutional code generated by the encoder of Fig. 1. Similar calculations have been preformed for the standard constraint length $\nu = 2$ code with connection polynomials $1 + D^2$, ad $1 + D + D^2$. However, as in this case the metric x state-diagram has 48 states symmetry considerations (Schalkwijk, J.P.M., Vinck, A.J., and Post, K.A., 1978) are used to first reduce the metric x state-diagram to a simpler 36 state configuration. These results will be published later.

ACKNOWLEDGEMENT

The author wants to thank A.J.P. de Paepe for helpful discussions during the preparation of this manuscript, J.P.J.C. Aarts for performing the necessary calculations, and Mrs. G.H. Driever-van Hulsen for the accurate typing of the manuscript.

REFERENCES

Forney, Jr., 1971, "Convolutional codes I: Algebraic structure", IEEE Trans. Inform. Theory, vol. IT-16, pp. 720-738, and vol. IT-17, p. 360.

Meeberg, v.d., L., "A tightened upper bound on the error probability of binary convolutional codes with Viterbi decoding", IEEE Trans. Inform. Theory, vol. IT-20, pp. 389-391.

Post, K.A., 1977, "Explicit Evaluation of Viterbi's Union Bounds on Convolutional Code Performance for the Binary Symmetric Channel", IEEE Trans. Inform. Theory, vol. IT-23, pp. 403-404.

Schalkwijk, J.P.M., and Vinck, A.J., 1975, "Syndrome Decoding of convolutional codes", IEEE Trans. Commun. (Corresp.), vol. COM-23, pp. 789-792.

Schalkwijk, J.P.M., and Vinck, A.J., 1976, "Syndrome Decoding of Binary Rate-½ Convolutional Codes", IEEE Trans. Commun., vol. COM-24, pp. 977-985.

Schalkwijk, J.P.M., Vinck, A.J., and Post, K.A., 1978, "Syndrome Decoding of Binary Rate k/n Convolutional Codes", IEEE Trans. Inform. Theory, to appear.

Viterbi, A.J., 1971, "Convolutional codes and their performance in communication systems", IEEE Trans. Commun. Technol. (Special Issue on Error Correcting Codes-Part II), vol. COM-19, pp. 751-772.

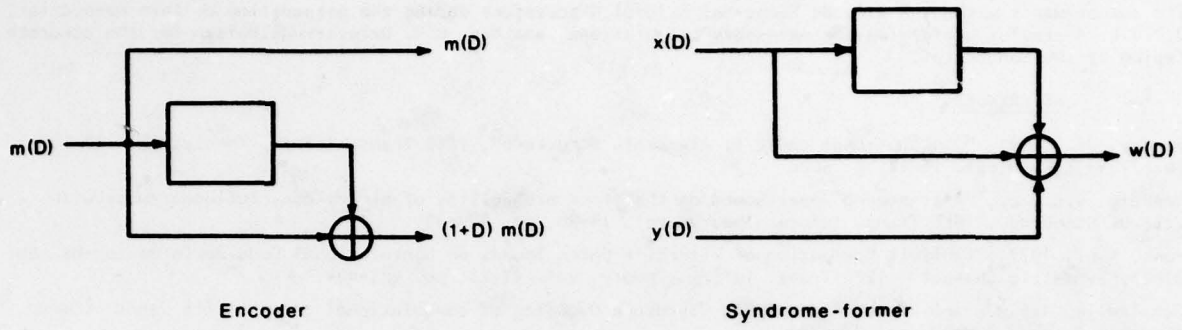


Fig. 1. Encoding and syndrome forming for a rate $\frac{1}{2}$ code.

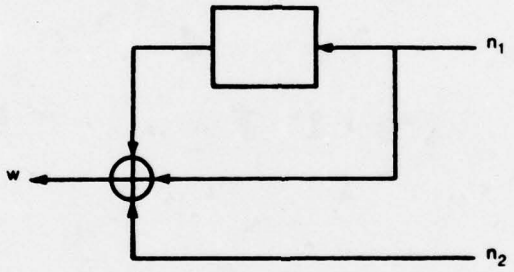


Fig. 2. Syndrome former with noise digits n_1, n_2, and syndrome digit w.

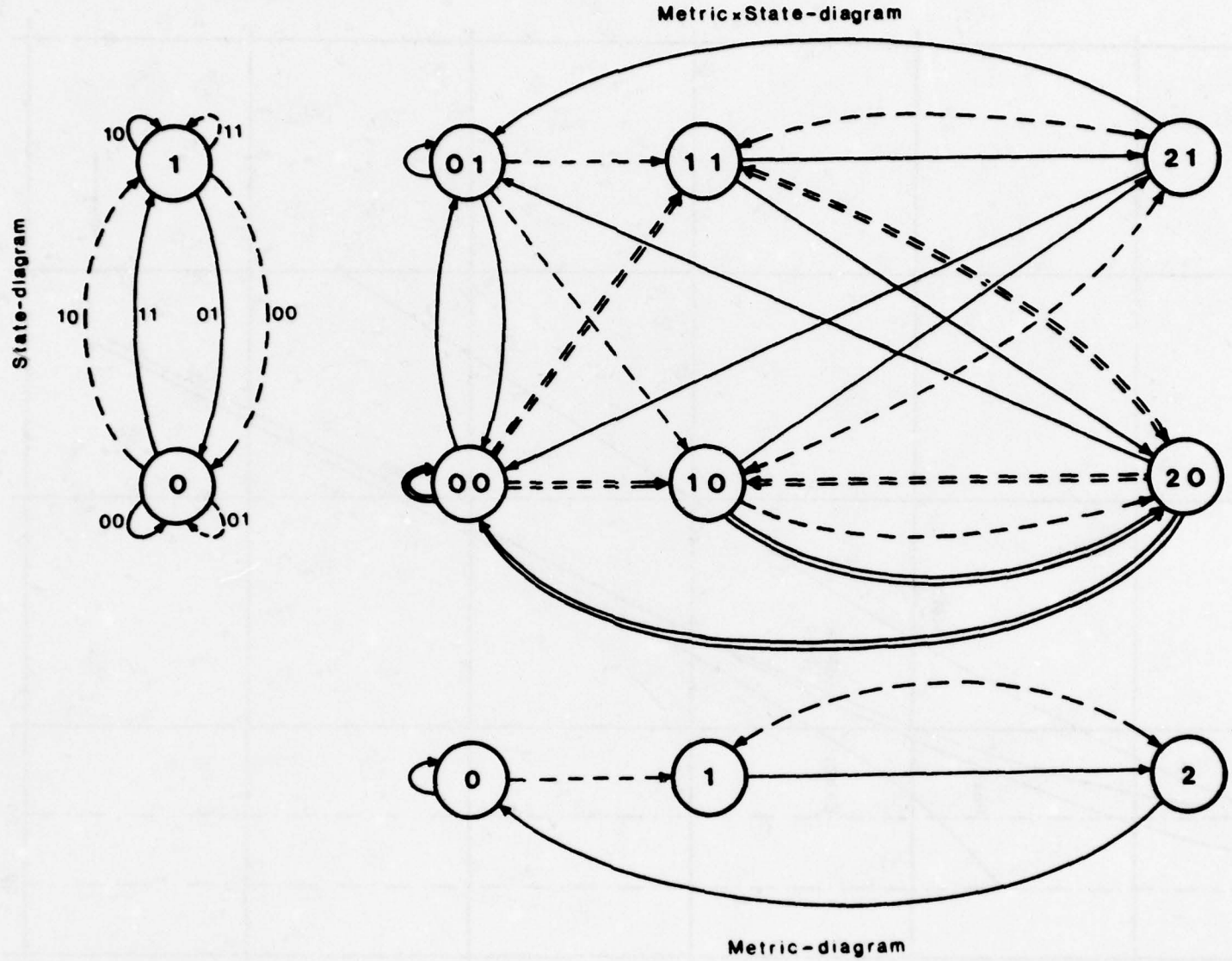


Fig. 3. State-diagram, metric diagram, and metric x state-diagram.

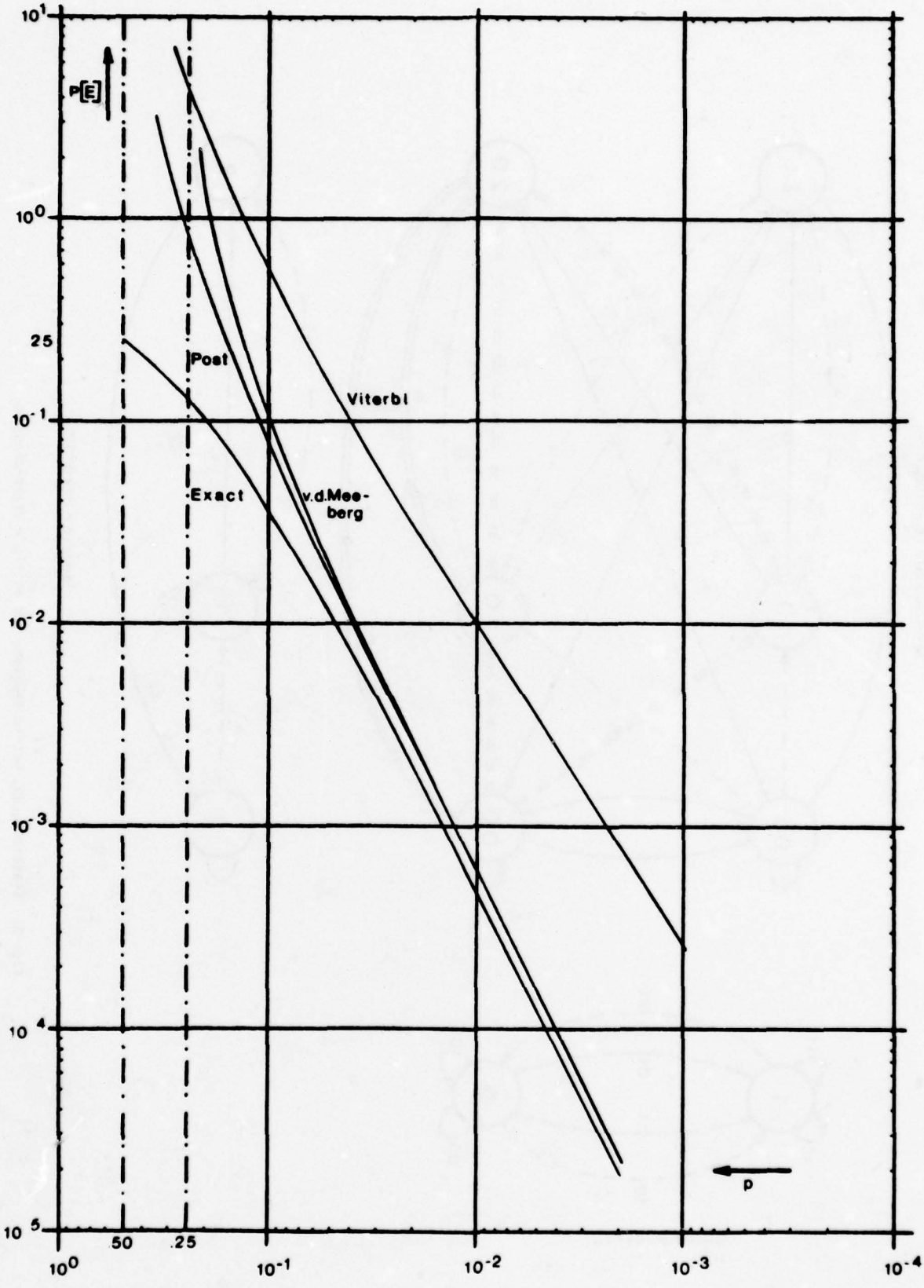


Fig. 4. The event error probability P[E]

DISCUSSION

J.K.Wolf, US

For longer constraint length codes, how many metric (states) are required?

Author's Reply

Conjecture, approximately (12) $N_{n,h,1/3}$ where $N_{n,h,1}$ is the number of metric equivalence classes of the code (see Schalkwijk, Vinck and Post, IT-Trans., Sept. 1978).

DIGITAL COMMUNICATIONS USING SOFT-DECISION DETECTION TECHNIQUES

P.G. FARRELL, E. MUNDAY & N. KALLIGEROS

University of Kent at Canterbury,
The Electronics Laboratories,
Canterbury,
Kent, CT2 7NT,
England.

SUMMARY

The paper begins by reviewing briefly the history of soft-decision techniques, and outlines the potential advantages of using such techniques. Soft-decision (probabilistic) detection has been applied with success to the decoding of convolutional (non-block) codes. Recent developments in semiconductor devices have now made the application of soft-decision detection to block coding systems both possible and practical.

Two major sections of the paper describe and evaluate experimental implementations of soft-decision schemes. First the results of some studies of quite simple block soft-decision decoding schemes are presented. Then a practical implementation of a data transmission system incorporating full minimum-distance soft-decision decoding is described. The system is based on a transparent product block code, with interleaving; and is suitable for use on HF, VHF and UHF channels. The results of tests on simulated and real channels are presented and commented on. A considerable improvement in system performance is achieved by means of soft-decision methods.

After a brief comment on the relevance of soft-decision detection to spread-spectrum systems, the paper concludes by suggesting that really efficient operation of data transmission systems with channel (redundant) coding is impossible without soft-decision detection, particularly in non-Gaussian environments.

1. INTRODUCTION

It is well known (e.g. BENNETT & DAVEY, 1965) that an optimum method of detection (demodulation and decoding), for a data transmission system with channel (error-correction) coding, is coherent correlation detection (or matched filtering) of the sequence of signal elements corresponding to the block length, in the case of a block code, or to the decoder search length, in the case of a convolutional code (see fig. 1(a)). In practice, unless the block or search length (and therefore the constraint length) is very short, this ideal detector is too complex to realise, because of the difficulty of generating, storing and correlating the large number of analogue signal elements required. Thus most practical detectors consist of an analogue demodulator, possibly coherent, operating on individual signal elements, followed by a purely digital decoder operating on blocks of the digits produced by the "hard" decisions of the demodulator (see fig. 1(b)). However, some of the information which would be lost by only correlating over a signal element can be used to assist and improve the decoding process, and vice-versa. Additional information can be fed forward from the demodulator to improve operation of the decoder, or fed back from the decoder to improve operation of the demodulator (see fig. 1(c)). The advantage of these forms of partially combined demodulation and decoding (or inter-active demodulation and decoding) is that they are much less complex to implement than fully combined forms of demodulation and decoding such as coherent correlation detection or matched filter detection. In addition, under certain circumstances the performance of some inter-active demodulation and decoding methods (or probabilistic decoding methods, as they were collectively called (WOZENCRAFT & KENNEDY, 1966)) is asymptoticaly close to that of ideal detection (e.g., LEE, 1976; and FRITCHMAN, et al, 1977).

Null-zone, or forced erasure detection, or failure correction decoding (BLOOM, 1957; PETTIT, 1965· JAYANT, 1966; MARQUART, 1967; WHITE, 1967; HELLER, 1967; KAZAKOV, 1968; SMITH, 1968; SULLIVAN & HEATON, 1969) are all ways of implementing feedforward between the demodulator and the decoder: processed signal elements with values lying near the threshold level of the demodulator, and thus of doubtful worth, are passed forward to the demodulator labelled as erasures. The decoder now has some knowledge of where errors are likely to be in the block, and can decode accordingly. In this way the error-correcting power of a code can be approximately doubled (a code with Hamming distance d can correct d-1 errors in a block transmitted over a binary erasure channel, but only $\lfloor (d-1)/2 \rfloor$ on a BSC, where $\lfloor x \rfloor$ is the largest integer $\leq x$). Feedback from the decoder could be used to adaptively adjust demodulator threshold levels, for example. In a binary symmetric system, if errors seem to occur more often in ones than in zeros, then the threshold could be altered to restore symmetry (i.e., approximately equal numbers of one and zero errors).

Null-zone detection can be extended to double-null-zone detection with an improvement in performance (CAHN, 1969), and can be generalised to more than two "null-zones", though with gradually diminishing rise in performance as the number of zones is increased. This general form of null-zone detection is called soft-decision decoding (CHASE, 1972; EINARSSON & SUNDBERG, 1976; SUNDBERG, 1975; SUNDBERG, 1977; HARRISON, 1977; GOODMAN & GREEN, 1977). Thus, strictly speaking, soft-decision decoding is a type of probabilistic decoding. Because of its importance, however, the term "soft-decision" has come to replace the word "probabilistic", which has fallen into disuse. Thus soft-decision decoding not only refers to the particular method, but also has come to mean the whole field of decoding with confidence (reliability) information. This is fortunate in a way, because the term probabilistic decoding has also been applied to sequential decoding of convolutional codes (FANO, 1963). Work in the field of soft-decision (probabilistic) decoding was initiated by BALSER & SILVERMAN (1954 & 1955), and some of the early work in this field is summarised in SCHWARTZ (1961).

2. SOFT-DECISION DECODING

Instead of making a hard decision, on each binary signal received, a soft-decision demodulator first of all decides whether it is above or below the decision threshold, and then computes a "confidence" number which specifies how far from the decision threshold the demodulator output is. This number could in theory be an analogue quantity, but in practice, if it is to be useful it must be quantised. Thus the output of the demodulator is still quantised, but into many more than the two regions of a hard-decision device.

An example of an 8-region device is given in figure 2. In this case the input to the demodulator is a binary signal, and the signal space is quantised into eight regions, deliniated by one decision threshold and three pairs of confidence thresholds. Each input signal is thus demodulated into an output character consisting of one binary hard decision digit and two binary confidence digits. In general, each binary signal is demodulated into a character consisting of $\log_2 Q$ binary digits (see figure 3) at the output of the demodulator, if there are Q regions in the quantised output signal space (normally Q is a power of 2). A binary code word of n digits is thus represented by $\log_2 Q = r$, say, binary code words, each of n digits. One of these consists of the (hard) decision digits, the others consist of the confidence digits of appropriate weighting. If the signal falls in a region of complete confidence, then the confidence digits of the corresponding output character are all ONES. The soft-decision-distance, d_s, between an output character and each of the two highest confidence output characters is then the Euclidean distance between the output character and each of the highest confidence characters; the distance between characters corresponding to adjacent regions being unity. Thus, for example, if a signal is demodulated as 0 0 0, then the soft-decision distances to 1 1 1 and 0 1 1 are respectively 4 and 3 (see figure 2). A convenient way of computing the soft-decision distance between two characters, when the regions and characters are mapped as in figure 2, is to invert the confidence digits of one of the characters if their decision digits are different, then modulo-2 add the resulting characters, and finally interpret the result as a binary number. Thus

$d_s(0\,0\,0,\ 1\,1\,1) = 0\,0\,0 \oplus 1\,0\,0 = 1\,0\,0 \equiv 4$, and

$d_s(0\,0\,0,\ 0\,1\,1) = 0\,0\,0 \oplus 0\,1\,1 = 0\,1\,1 \equiv 3.$

The soft-decision distance (SDD) between a received code word and a possible transmitted code word may be computed by summing the appropriate soft-decision distances for each digit (character) of the code word. Thus, by computing the soft-decision distances to all possible code words, the one most likely to have been transmitted may be determined by selecting the one with lowest SDD, in a manner exactly analogous to minimum Hamming distance decoding. For example, using the demodulator arrangement of figure 2, together with a single repetition code, assume that the code word 1 1 is transmitted. Imagine that the first ONE is demodulated as 0 0 0, and the second ONE as 1 1 0. Then the SDD to 1 1 1 ¦ 1 1 1 is 4 + 1 = 5, and the SDD to 0 1 1 ¦ 0 1 1 is 3 + 6 = 9 (see figure 4). The minimum SDD value is 5, so the soft-decision decoder outputs the correct code word, 1 1, in spite of the error in the first digit. A hard-decision decoder, given 0 1, could only detect the error, without being able to correct it. The soft-decision method of decoding described above is called minimum soft-decision distance (MSDD) decoding. It is equally applicable to block and convolutional codes (see below).

It should be noted that the confidence thresholds need not be linearly (equally) spaced; MASSEY, 1974; LEE, 1976; and HARRISON, 1977 amongst others, have shown that a non-linear spacing array may be optimum. Alternatively, or in addition, soft-decision distances may be calculated as weighted sums of the demodulator output characters, as in generalised minimum distance decoding (FORNEY, 1966) and threshold decoding (MASSEY, 1963). Finally, a different mapping of characters on to regions than that given in figure 2 may be used, either to pre-weight the characters or to make SDD computation easier. Alternative distance functions may also be used (REDDY, 1974).

The above example shows that a simple single-error-detecting repetition code (d = 2) is capable of single-error-correction when used with soft-decision demodulation ($d_s = 14$). This confirms the statement in section 1 that the error control power of a code is almost doubled by the use of soft-decision detection: an e-error-detecting code approximates to an e-error-correcting code, or a t-error-correcting code approximates to a 2t-error-correcting code. In general

$$d_s = d(Q-1)$$
$$\approx d.Q \quad \text{for large } Q.$$

In practice values of $\log_2 Q$ greater than 4 or 5 (16 or 32 regions) are unnecessary, as the increase in performance is only marginal (BATSON, et al, 1972). In terms of decoder output error rate, the improvement due to soft-decision demodulation depends on the particular code and channel error statistics, but one to two orders of magnitude or more decrease in output error rate, for white Gaussian noise channels with error rates in the range 10^{-2} to 10^{-4}, is typical. This corresponds to 1.5 to 2 dB improvement in signal-to-noise ratio. In non-Gaussian noise the improvement is considerably greater (CHASE, 1976).

Soft-decision decoding became of practical importance with the discovery of the Viterbi algorithm (VA) for maximum likelihood (minimum distance) decoding of convolutional codes (VITERBI, 1967). Use of soft-decision demodulation does not significantly increase the complexity of a VA decoder, which is a function of the rate and constraint length of the code. The present development of integrated circuit micro-electronics permits implementation of half-rate convolutional codes with encoding constraint lengths of up to about 12, with soft-decision decoding. More powerful convolutional codes, particularly if for use on bursty channels, require sequential decoding (WOZENCRAFT, 1957). Use of soft-decision sequential decoding was initially found to be impractical, but pioneering work by JORDAN, 1966; the advent of stack decoding algorithms (JELINEK, 1969; HACCOUN & FERGUSON, 1975); and more recent research into algorithms which make efficient use of the structural and distance properties of convolutional codes (NG & GOODMAN, 1978), indicate that soft-decision sequential decoding is feasible. Soft-decision threshold decoding of convolutional codes is also possible (GOODMAN & NG, 1977).

Soft-decision techniques were less generally applicable to block codes until comparatively recently. The early work previously mentioned was concerned with quite simple block codes (e.g., the Wagner code - a single-parity-check code with soft-decision demodulation - of BALSER & SILVERMAN, 1954). Hamming single-error-correcting codes with soft-decision decoding have been studied by SUNDBERG (1977) and HARRISON (1977). Work by FORNEY (1966) on generalised minimum distance decoding led to the application of soft-decision techniques to iterated and concatenated codes (REDDY & ROBINSON, 1972; JUSTESEN, 1972); to product (two-coordinate) codes (WAINBERG, 1972; CHASE, 1973); and to algebraic decoders (EINARSSON & SUNDBERG, 1976) and error-trapping decoders (GOODMAN & GREEN, 1977) based on successive erasure decoding. WELDON (1971), developed a method of weighted erasure (multiple syndrome) decoding, a soft-decision decoding technique applicable in principle to any block code for which a decoding procedure is known. This work was extended by WAINBERG & WOLF (1972) for burst errors, and by REDDY (1974). MASSEY's (1966) work on threshold decoding has led to the combination of soft-decision techniques with majority logic decoding (e.g., SUNDBERG, 1975). A quite different approach was discovered by HARTMANN & RUDOLPH (1976), which may be called soft-decison dual-code-domain decoding. It is an optimum decoding method in a symbol-by-symbol sense in that it minimises the symbol error probability, rather than the code-word error probability. It is important because it applies to codes of high rate, unlike most of the methods mentioned previously. More general application of soft-decision decoding to block codes is possible if full(comparison of received word with all possible code words) minimum distance decoding or MSDD, is used. This has recently become feasible in practice because of the availability of cheap integrated circuits and microprocessors. It is particularly feasible if the code used has some internal structure which can simplify MSDD decoding; for example, if the code is a product or concatenated code (DORSCH, 1974; FARRELL & MUNDAY, 1976; FARRELL, 1977 and section 4 of this paper). Also WOLF (1977) has shown that any linear block code can be soft-decision decoded using the Viterbi algorithm. Thus a very wide range of block error-correcting codes can be decoded efficiently by means of soft-decision techniques.

The discussion so far has concentrated on error-correcting codes; it is of interest to note that soft-decision techniques can also be used to improve the efficiency of error detection codes, and therefore of automatic request for repeat (ARQ) systems (SUNDBERG, 1976).

3. SIMPLE SOFT-DECISION DECODING WITH AND WITHOUT RETRANSMISSION

This section describes and evaluates the performance of a binary single-parity-check code with soft-decision decoding capable of attempting correction of either single errors, or single and double errors, with or without the assistance of retransmission requested via a feedback channel. The block length of the code is 8, so its rate is 7/8. Binary polar baseband signals were amplitude modulated (ASK), transmitted through an additive Gaussian white noise channel, and received with (in the absence of noise) a nominal amplitude of ± 1V before soft-decision processing in the demodulator. Six soft-decision confidence thresholds were used, set at ± 0.25 V, ± 0.5 V and ± 0.75 V; thus the soft-decision quantiser is an 8-region device (Q = 8), and the demodulator output for each binary signal is a hard-decision digit and two confidence digits (as in fig. 2).

The algorithms for error-correction are as follows:-

(i) Single-error-correction, without retransmission (SECA)

(a) Compute the hard-decision (HD) and confidence digits for the 8 binary signals in each block;
(b) using the HD digits of the seven information signals, re-calculate the parity check, and compare with the received parity check;
(c) if the same (parity holds), take no further action (i.e., assume that no error has occurred); if different (parity fails), invert the HD digit with least confidence number in the block, thus hopefully correcting the most likely single error. If there is more than one digit with the same confidence number, then correct the first one in the block.

(ii) Single-and-double-error-correction, without retransmission (SDECA)

(a) and (b) as above;
(c) if parity fails, invert the HD digit with least confidence number (the first if two or more the same);
(d) if parity holds, then there are either no errors, or two errors: if none of the confidence numbers has the lowest possible value (0 0) then assume no errors; if at least one confidence number is 0 0, then assume a double error has occurred, and invert the HD digit with 0 0 confidence and the HD digit with least (or equal) confidence among those remaining (again, the first if two or more are the same).

(iii) Single-and-double-error-correction, with retransmission (SDECAR)

(a) - (c) as above;
(d) if parity holds, and no confidence number is 0 0, then no further action required;
(e) if parity holds, and at least one confidence number is 0 0, then request retransmission of that signal;
(f) if the retransmitted HD digit is the same as the first one, and the confidence number the same or higher, take no action (no errors assumed); if the HD digits are different, and the confidence numbers the same, no action is again appropriate, since a decision for or against a double error would be arbitrary; if the HD digits are different, and the retransmitted confidence number is higher (it can't be lower) then a double error is indicated, and correction as in (ii) (c) above is applicable.

An experimental soft-decision SECA system was constructed in hardware and signal-to-noise ratio versus error rate measurements were made, see fig. 5; the details will be found in KALLIGEROS, 1977. All three systems (SECA, SDECA and SDECAR) were also simulated on a computer; performance measurements for the experimental and simulated SECA system were in close agreement. The performance curves are shown in figure 6. SDECA without retransmission is clearly worse than both the SDECAR and the SECA. This is because of the poor double-error-correction properties of the SDECA; if the decision that two errors are present in the block is wrong, then two extra errors are created. Re-transmission of the signal upon which the decision is based clearly improves the reliability of the final decision, and this emerges in the improved performance of the SDECAR system. There is a theoretical justification for this result: a single-parity-check code is incapable of correcting two errors, even with soft-decision decoding, because

$$d_s = d(Q - 1) = 2(8 - 1) = 14$$

and to correct 2 errors $d_s = 17$ is required. There is some spare soft-decision distance, however, since to correct only a single error $d_s = 9$ is needed; this spare distance can be put to good use if retransmission is possible (compare with SUNDBERG, 1976). The rate penalty of re-transmission is quite small, as the curve shows, because only a single digit (signal) in the block is retransmitted (housekeeping problems arise only in the feedback link). For signal-to-noise (SNR) ratios less than about -1.5db the performance of the SDECAR is better than the SECA; this is because of the predominance of double errors at low SNR. The advantage of SECA at relatively high SNR is small, however (0.5-1 db). The performance of the SECA is approximately that of the Wagner code investigated by BALSER & SILVERMAN (1954). The curves indicate that coding gain is achieved for SNR > -2.5db. This is clearly much better than the performance of a Hamming single-error-correcting code with similar block length or even similar rate. It is interesting to conjecture whether an ARQ system based only on soft-decision-error-detection; that is, with no redundant digits transmitted on the forward link; is feasible and effective.

4. DATA TRANSMISSION WITH ADAPTIVE SOFT-DECISION DECODING

In this section, a binary data transmission system which incorporates an error-correcting product code, interleaving, and adaptive soft-decision decoding, is described and evaluated. A transparent product (row-and-column) code is used because it is relatively simple to implement and decode efficiently using soft-decision methods. The code is formed from the product of two linear binary cyclic codes with (n, k, d) = (15, 11, 3); thus the code has parameters (225, 121, 9), and is approximately half-rate. The transparency permits superimposition of a 10-times slower data stream, giving a flexible two-channel system. Interleaving is required to combat the bursts of errors which occur on the HF channel (3-30MHz) for which the system was designed. Each binary data digit is transmitted as a 15-bit m-sequence, inverted or non-inverted to convey a ONE or a ZERO. This spreads the spectrum of the baseband signal, with potential advantage on the HF channel which is perturbed by frequency selective fading and multi-path effects. The m-sequence is detected at the receiver in a digital correlator, after hard-decision demodulation of the ASK RF signal. The output of the correlator is a number ranging in value from 0 to 15, depending on whether a data 0 or 1 was transmitted, and on how many errors occurred in the m-sequence bits. This number may be interpreted as the hard-decision digit and confidence digits of a 16-region soft-decision demodulator. Thus each data digit is demodulated into a 4-digit binary number: a HD digit and 3 confidence digits. This novel way of obtaining soft-decision information avoids the problems of demodulator output quantisation. The soft-decision decoder is an implementation of MSDD decoding as described in section 2. The structure of the transparent product code makes this practical.

4.1 System Transmitter and Receiver

A block diagram of the transmitter is given in fig. 7. Channel 1 is for binary data at $\leq$ 100 bits/sec (the faster channel) and channel 2 for data at $\leq$ 10 bits/sec (the slower channel). Digits on channel 1 are fed simultaneously into the cyclic outer encoder (a 4-stage feedback shift-register) the inner cyclic encoder (basically an 11-stage feedback shift-register) and a random-access memory (RAM) which is used for bit interleaving. The RAM is organised into four sub-frames, each consisting of 15 rows and 15 columns, as shown in Fig. 8.

Data from channel 2 is fed into the first position in each row; it is also used to invert (if it is a ONE) or non-invert (if a ZERO) the 10 information digits from channel 1 being fed into the outer and inner encoders. The channel 2 digit is also fed into the encoders, which require 11 information digits in all. Positions 2-11 in the first row are filled with channel 1 data, and then the remaining 4 positions with the parity checks derived by the outer cyclic encoder; this completes one code word row of 15 digits. Rows 1-11 are then filled with outer code words in the same way; and the final 4 rows are completed with the parity checks derived by the inner cyclic encoder. Thus the rows of the sub-frame are outer code words, and the columns are inner code words. The remaining 3 sub-frames are filled in the same way. Digits can now be read out of the RAM, but not just vertically, because this would mean that consecutive inner code word digits would not be interleaved. Instead, the RAM is read out in a diagonal pattern, in the digit order shown in the diagram: first the main diagonals in each sub-frame; then the ones immediately below, which since they have less than 15 digits, are completed with the digits in the top-right-hand-side of the sub-memories; and so on, taking one digit from each sub-frame in turn. In this way an interleaving factor of 60 is achieved for the inner code, and 56 for the outer code (they are different because the sub-frames are square, as the codes have the same block lengths). While the RAM is being read out, encoding continues in a second RAM, organised in the same manner; thus complete frames are encoded in each RAM alternately.

After sequence inversion keying, the waveform is passed through a baseband equaliser, adjusted so as to minimise symbol distortion and intersymbol interference arising in the system due to the various filter characteristics. The equalised sequence waveform is then amplitude-keyed onto a 1.6KHz sub-carrier, and then fed into a SSB-ASK HF transmitter with associated wide-band linear HF amplifier. The nominal transmission bandwidth of the system was 2.7KHz, as determined by the transmitter filter. There is provision in the system for inserting a frame synchronisation sequence, so that the decoder can be correctly aligned before actual data transmission begins, and checked afterwards.

A block diagram of the receiver is given in fig. 9. After RF and IF filtering and amplification, carrier is extracted and regenerated in a P.L.L. circuit, and applied with the IF waveform to a product detector. Clock is also regenerated at the appropriate rate. The sequence waveform is effectively sampled by the clock as it enters the correlator; this minimises noise and intersymbol interference effects. The correlator, by comparing the received sequence, correctly phased, with a locally generated 15-digit m-sequence, removes the sequence-inversion-keying and quantises each received encoded digit into 16 levels, represented by a four-digit binary character: a decision digit and three confidence digits. The characters are then fed into a RAM, organised into four sub-frames holding 15 x 15 = 225 characters each, in the same way that the corresponding coded digits are read out from the transmitter RAM; so that when the receiver RAM is full, the characters in it are in exactly the same position that the corresponding coded digits were in in the transmitter RAM. Frame synchronisation for the de-interleaving and decoding operations is established from the hard-decision digit.

Decoding of the received characters can begin as soon as the RAM is full and synchronisation has been achieved; subsequent demodulated characters are fed into a second RAM, identical with the first, so that reception can continue uninterrupted. Decoding of columns 12-15 of each sub-frame is carried out first, as these columns contain checks on checks, and are required for decoding columns 1-11. The confidence digits of the rows, and of columns 1-11, are then summed, and the rows and the columns separately ranked in order of confidence. The row with highest confidence value is decoded next, followed by the column with highest confidence and so on. Rows and columns are thus decoded alternately in order of confidence value, as this minimises the probability of erroneous decoding. In this sense decoding is an adaptive process. Decoding is done by computing the soft-decision distance between each received - possibly erroneous - row or column (consisting of 15 4-digit characters) and all the possible correct code words. The correct code words are stored in a programmable read only memory (PROM). As the distance computations are done sequentially in a systematic order, it is only necessary to store the parity checks of the code words. Also, the code is transparent, so that only half the words need be stored anyway, as the remainder are inversions of the first half. When the nearest code words is found, it is read into the RAM, via a buffer, replacing the decision digits of the appropriate row or column, the confidence digits being all reduced to zero if an error has been detected. Since only half the code book is stored, the modulo-225 threshold circuit inverts the nearest code word in the buffer if necessary.

After the first-pass decoding operation, a second pass is done. In this second operation the first eleven rows are passed again sequentially through the soft-decision distance processor, without regard for their rank. This further improves the reliability of the information digits on both channels. Once all rows and columns have been decoded, then the corrected decision digits can be read out of the RAM. All the processing will take place in the time it takes to fill the second RAM, so that the first RAM then becomes free for storing demodulated characters while those in the second RAM are being processed. Further details of the system may be found in FARRELL & MUNDAY (1976), FARRELL (1977), FARRELL & MUNDAY (1978) and MUNDAY (1979).

4.2 System Performance

The digital parts of the transmitter and receiver were tested by adding (modulo-2) random errors and burst error patterns, from an experimental error generator developed at Kent (ROCHA, 1976), to the output of the sequence inversion keyer. Error rates were measured before m-sequence correlation, after correlation (on the hard-decision digits), and after the first and second passes of the decoding process (with and without soft-decision). The results for random errors were as follows:-

channel error rate	1.8×10^{-1}
error-rate at output of correlator (hard-decision or data digit error rate)	3×10^{-3}
(approx. decoded error-rate of (15, 1, 15) repetition code at error-rate 1.8×10^{-1}	2.4×10^{-3})
error-rate after hard-decision product decoding first pass	2×10^{-5}
second pass	2.5×10^{-6}
(approx. decoded error-rate of (225, 121, 9) code at error-rate 3×10^{-3}	2.3×10^{-5})
error-rate after soft-decision product decoding first pass	1.2×10^{-6}
second pass	4×10^{-7}

These results for the system, when compared with those in brakets, show that both the correlator and the hard-decision product decoder are performing as expected, and that soft-decision decoding decreases the output error-rate by about one order of magnitude. It is also clear that the adaptive (alternate row and column) decoding technique is efficient even in the absence of soft-decision information, particularly when a second pass is used. Results for various burst lengths are shown in figure 10 (the burst error densities were approximately 0.56). In a bursty situation, the improvement in error rate due to soft-decision decoding is much more than one order of magnitude.

The complete transmitter and receiver (without the linear HF power amplifier) were connected back-to-back, the equaliser was adjusted for optimum operation, and error rates were measured, as detailed above, with white Gaussian noise added to the transmitted signal, at a sequence digit rate of 2K bit/sec. These tests were done with the sub-carrier frequency adjusted to 100KHz. The results are given in figure 11. It can be seen that soft-decision decoding is worth about 1.5db decrease in S/N ratio, or an order of magnitude decrease in error-rate. This compares well with the performance predicted in section 2. The overall performance of the system, based on a consideration of the overall rate ($0.5 \times 1/15 = 1/30 \equiv 14.8$db), indicates that there is a short-fall of about 2db. This is due to the hard-decision correlation demodulation of the m-sequence (see comments in the following sections).

The complete system was tested in three sets of HF trials, two over approximately 120 miles, and one over approximately 300 miles. HF frequencies of about 4.5MHz were used, and various sequence digit rates. Error rates as above were measured, and a chart recording was made of the input signal strength during each trial. Some of the results obtained are given in figure 12, which shows plots of error rate against percentage of lowest output-error-rate frames received, at 1K bit/sec over a 300 mile path. As a preliminary to these tests, the transmitter equaliser was re-adjusted to match, as far as possible, the actual HF transmitter characteristic. It is interesting to note how little the channel error rate varies, and how rapidly the soft-decision decoding error rate falls, as the worst frames are deleted. Clearly the system performs well even in a very bursty environment, and degrades gracefully.

It was not difficult to implement the digital parts of the system. Complexity is moderate (34 cards, or approximately a total of 418 IC chips) and the cost surprisingly low (in the region of £500). Apart from some difficulties connected with the synchronisation of the correlator, arising from the unstable nature of the HF channel, the experimental system operated quite satisfactorily and reliably during the field trials.

Further work with this system will seek to quantify more precisely the advantage of soft-decision decoding when applied to HF transmission, and will extend its use to VHF and UHF channels as well. DPSK modulation will be investigated, as an alternative to ASK; and the m-sequence inversion spectrum spreading technique will probably be abandoned or modified (see next section). Microprocessor implementation of the decoder will be also studied.

5. SOFT-DECISION TECHNIQUES FOR SPREAD-SPECTRUM SYSTEMS

The results of the work both reviewed and reported in this paper indicate that spread-spectrum systems of the direct-sequence type (DIXON, 1976; FARRELL & ANDJARGHOLI, 1976) would perform better with soft-decision decoding. Direct-sequence modulation is a form of repetition coding, to which soft-decision decoding is applicable. Preliminary calculations for white Gaussian noise (TOZER, 1978) show that an advantage of 1.5 - 2db is achieved. In a non-Gaussian environment the advantage is likely to be very much greater. This corresponds very well with the calculations and results reported above. In particular, the short-fall of about 2db in Gaussian noise noted for the adaptive data transmission system of section 4 could well be mainly due to the fact that the spectrum spreading m-sequence was not soft-decision detected, and similarly this will have had a severely deleterious effect on the performance of the system in the HF environment. It is conjectured that errors arising from noise and interference effects suffered by spread-spectrum systems, particularly the self-interference effects that occur in multi-user spread-sprectrum systems, are amenable to effective control with soft-decision techniques.

6. CONCLUSIONS

The performance results for the experimental soft-decision decoding systems studied in this paper confirm the results predicted by theory, and presented by other researchers. The advantages of using soft-decision techniques are clear, and may be listed as:-

(i) soft-decision decoding is applicable to a wide range of error-correcting codes, both block and convolutional;
(ii) soft-decision techniques are also appropriate for use with error-detection/ARQ systems;
(iii) there is a substantial increase in performance (~ 2db in SNR) where soft-decision techniques are applied to the Gaussian channel, but the increase is even more marked in a non-Gaussian environment;
(iv) the performance of the best soft-decision algorithms asymptotically approaches that of the equivalent optimum detector;
(v) in the case of a non-Gaussian (e.g., impulse noise) channel, the optimum detector may be unknown or unrealisable: a soft-decision detector may then be the best practical device to use;
(vi) the implementation of a soft-decision decoder is not substantially more complex than that of the corresponding minimum distance decoder;
(vii) use of a soft-decision demodulator may be traded for additional decoding complexity.

It is of particular value to have demonstrated that soft-decision technqiues can be effectively applied to block codes. Though in many circumstances convolutional codes outperform block codes, there are certain situations in which block codes are more appropriate, such as when relatively short messages are to be transmitted, or when system synchronisation has to be achieved very rapidly, or when a relatively simple coding method is sufficient. Use of soft-decision decoding in these cases enables achievement of the highest possible performance.

A reason often quoted for rejecting the use of soft-decision decoding is that it requires modification or replacement of the hard-decision demodulator in a receiver. This modification, however, is normally quite simple; merely the provision at an additional output terminal of the demodulated signal (suitably buffered if necessary) before hard-decision, limiting or pulse regeneration. The reward for this modification could be a doubling of coding gain, since many practical hard-decision decoding schemes can only offer up to about 2db of coding gain. As MASSEY (1974) has pointed out, to use a hard-decision demodulator can, in overall system performance terms, cancel out most or all of the gain provided by the coding scheme. Thus soft-decision demodulation should be adopted or provided for wherever possible.

7. REFERENCES

BALSER, M. & SILVERMAN, R.A. (1954); Coding for Constant-Data-Rate Systems, Part I: A New Error-Correcting Code; Proc IRE, Vol 42, No 9 (September), p 1428; (1955), Part II: Multiple-Error-Correcting Codes; Proc IRE, Vol 43, No 6 (June), p 728.

BATSON, B.H., MOOREHEAD, R.W. & TAQVI, S.Z.H. (1972); Simulation Results for the Viterbi Decoding Algorithm, NASA Report No TR R-396.

BENNETT, W.R. & DAVEY, J.R. (1965); Data Transmission; McGraw-Hill.

BLOOM, F.J. et al (1957); Improvement of Binary Transmission by Null-Zone Reception; Proc IRE, Vol 45, p 963.

CAHN, C.R. (1969); Binary Decoding Extended to Nonbinary Demodulation of Phase Shift Keying; IEE Trans, Vol COM-17, No 5 (Oct.), p 583.

CHASE, D. (1972); A Class of Algorithms for Decoding Block Codes with Channel Measurement Information; IEEE Trans, Vol IT-18, No. 1 (Jan.), p 170.

CHASE, D. (1973),;A Combined Coding and Modulation Approach for Communication over Dispersive Channels; IEEE Trans, Vol COM-21, No 3 (March), pp 159-174.

CHASE, D. (1976); Digital Signal Design Concepts for a Time-Varying Rician Channel; IEEE Trans, Vol COM-24, No 2 (Feb), pp 164-172.

DIXON, R.C. (1976); Spread-Spectrum Systems; Wiley.

DORSCH, B. (1974); A Decoding Algorithm for Binary Block Codes and J-ary Output Channels; IEE Trans, Vol IT-20, No 3 (May), pp 391-394.

EINARSSON, G. & SUNDBERG, C.E. (1976); A Note on Soft-Decision Decoding with Successive Erasures; IEEE Trans, Vol IT-22, Nol 1 (Jan.), p 88.

FANO, R.M. (1963); A Heuristic Discussion of Probabilistic Decoding; IEEE Tans, Vol IT-9, pp 64-74.

FARRELL, P.G. (1977); Soft-Decision Minimum-Distance Decoding; Proc NATO ASI on Communications Systems and Random Process Theory, Darlington, England, Aug. 1977.

FARRELL, P.G. & ANDJARGHOLI, G. (1976); A Spread-Spectrum Digital Transmission System for Reliable Communication in the HF Band; Proc IEE Colloq. on HF Communication Systems, London, Feb. 1976.

FARRELL, P.G. & MUNDAY, E. (1976); Economical Practical Realisation of Minimum-Distance Soft-Decision Decoding for Data Transmission; Proc. Zurich Int. Seminar on Digital Communications, March, pp 135.1-6.

FARRELL, P.G. & MUNDAY, E. (1978); Variable Redundancy HF Digital Communications with Adaptive Soft-Decision Minimum-Distance Decoding; Final Report on MOD (ASWE) Res. Study Contract AT/2099/05/ASWE.

FORNEY, G. (1966); Generalised Minimum Distance Decoding; IEEE Trans, Vol IT-12, No 2 (April), pp 125-131, and in "Concatenated Codes", MIT Res. Memo. No 37, 1966.

FRITCHMAN, B.D., et.al (1977); Approximations to a Joint Detection/Decoding Algorithm; IEEE Trans, Vol COM-25, No2 (Feb), pp 271-278.

GOODMAN, R.M.F. & GREEN, A.D. (1977); Microprocessor Controlled Soft-Decision Decoding of Error-Correcting Block Codes; Proc. IERE Conf. on Digital Processing of Signals in Communications, No 37, pp 37-349, Loughborough, England.

GOODMAN, R.M.F. & NG, W.H. (1977); Soft-Decision Threshold Decoding of Convolutional Codes; Proc IERE Conf on Digital Processing of Signals in Communications, No 37, pp 535-546, Loughborough, England.

HABER, F. (1977); Spread-Spectrum Signals and Bandwidth Utilization; Proc NATO ASI on Communications Systems and Random Process Theory, Darlington, England, Oct. 1977.

HACCOUN, D. & FERGUSON, M.J. (1975); Generalised Stack Algorithms for Decoding Convolutional Codes; IEEE Trans, Vol IT-21, No 6 (November), pp 638-651.

HARRISON, C.N. (1977); Application of Soft Decision Techniques to Block Codes; Proc IERE Conf on Digital Processing of Signals in Communications, Loughborough, England, No 37, pp 331-336.

HARTMANN, C.R.P. & RUDOLPH, L.D. (1976); An Optimum Symbol-by-Symbol Decoding Rule for Linear Codes; IEEE Trans, Vol IT-22, No 5 (Sept.), pp 514-517.

HELLER, R.M. (1967); Forced-Erasure Decoding and the Erasure Reconstruction Spectra of Group Codes; IEEE Trans, Vol COM-15, No 3 (June), p 390.

HOBBS, C.F. (1967); Universality of Blank-Correction and Error Detection; IEEE Trans IT-13, No 2 (April), p 342.

JAYANT, N.S. (1966); An Erasure Scheme for Atmospheric Noise Burst Interference; Proc IEEE, Vol 54, No 12 (Dec.), p 1943.

JELINEK, F. (1969); A Fast Sequential Decoding Algorithm using a Stack; IBM Jour. Res. Dev., Vol 13, Nov., pp 675-685.

JORDAN, K.L. (1966); The Performance of Sequential Decoding in Conjunction with Efficient Modulation; IEEE Trans, Vol COM-14, No 3 (June), pp 283-297.

KALLIGEROS, N. (1977); Soft-Decision Error-Correction; M.Sc. Dissertation, University of Kent at Canterbury, England.

KAZAKOV, A.A. (1968); A Method of Improving the Noise Immunity of Redundant Binary Code Reception; Telecoms (trans of Elec. & Radioteknika) Vol 22, No 3 (March), p 51.

LEE, L.N. (1976); On Optimal Soft-Decision Demodulation; IEE Trans, Vol IT-22, No 4 (July), pp 437-444.

MARQUART, R.G. (1967); The Performance of Forced Erasure Decoding; IEEE Trans, COM-15, No 2 (June), p 397.

MASSEY, J.L. (1963); Threshold Decoding; MIT Press.

MASSEY, J.L. (1974); Coding and Modulation in Digital Communications; Proc. Zurich Int. Seminar on Digital Communications, pp E2(1)-(4).

MUNDAY, E. (1978); Soft-Decision Decoding for HF Data Communications; Ph.D. Thesis, University of Kent at Canterbury, England.

NG, W.H. & GOODMAN, R.M.F. (1978); An Efficient Minimum Distance Decoding Algorithm for Convolutional Error-Correcting Codes, Proc IEE, to be published.

PETTIT, R.H. (1965); Use of the Null-Zone in Voice Communications; IEEE Trans, Vol COM-13, No 2 (June), p 175.

RAPPAPORT, S.S. & KURZ, L. (1965); Optimal Decision Thresholds for Digital Signalling in Non-Gaussian Noise; IEEE Int. Conv. Rec., Part 2, p 198.

REDDY, S.M. (1974); Further Results on Decoders for Q-ary Output Channels; IEEE Trans Vol IT-20, No 4 (July), pp 552-4.

ROCHA, V.C. (1976); Versatile Error-Control Coding Systems, Ph.D. Thesis, Univ. of Kent at Canterbury, England.

SCHWARTZ, L.S. (1961); Some Recent Developments in Digital Feedback Systems; IRE Trans, Vol CS-9, No 1 (March), pp 51-7.

SMITH, J.S. (1968); Error Control in Duobinary Data Systems by Means of Null-Zone Detection; IEEE Trans. Vol COM-16, No 6 (Dec.), p 825.

SULLIVAN, N.J. & HEATON, A.G. (1969); Transient Frequency Response of Transmittance Peaked I.F. Filters with Application to Null Zone Detection; Elec. Letters, Vol 5, No 18, p 423, 4th Sept.

SUNDBERG, C.E. (1974); Soft-Decision Error-Detection for Binary Antipodal Signals on the Gaussian Channel; Dept. Telecom. Th., Lund Univ., Sweden, Tech. Rep. TR-65.

SUNDBERG, C.E. (1975); Reliability Numbers Matching Binary Symbols for Gray-Coded MPSK and MDPSK Signals; as above, TR-66.

SUNDBERG, C.E. (1975); One-Step Majority Logic Decoding with Symbol Reliability Information; IEE Trans, IT-21, pp 236-242, No 2 (March).

SUNDBERG, C.E. (1976); A Class of Soft-Decision Error Detectors for the Gaussian Channel; IEEE Trans, Vol COM-24, No 1 (Jan), pp 106-112.

SUNDBERG, C.E. (1977); Asymptotically Optimum Soft-Decision Decoding Algorithms for Hamming Codes; Elec. Letters, Vol 13, No 2, p 38, 20th Jan.

C.C.I.T.T. (1964); Control of Errors for Data Transmission on Switched Telephone Connections. "Blue Book" (Supplement No. 66).

THIEDE, E.C. (1972); Decision Hysteresis Reduces Digital Pe; IEEE Trans, Vol COM-20, No 5 (Oct.), p 1038.

TOZER, T.C. (1978); Comparative Performance of Digital and Analogue Correlators; Digital Communications Res. Group Internal Report.

VITERBI, A.J. (1967); Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm; IEEE Trans, Vol IT-13, No 2 (April), pp 260-269.

WAINBERG, S. & WOLF, J.K. (1972); Burst Decoding of Binary Block Codes on Q-ary Output Channels; IEEE Trans, Vol IT-18, No 5 (Sept.), p 684.

WELDON, E.J. (1971); Decoding Binary Block Codes on Q-ary Output Channels; IEEE Trans, Vol IT-17, No 6 (Nov.), pp 713-718.

WHITE, H.E. (1967); Failure-Correction Decoding; IEEE Trans, Vol COM-15, No 1 (Feb.), p 23.

WOLF, J.K. (1977); Soft-Decision Decoding of Linear Block Codes; private communication.

WOZENCRAFT, J.M. (1957); Sequential Decoding for Reliable Communication; IRE Nat. Conv. Rec., Part II, pp 11-25.

WOZENCRAFT, J.M. & KENNEDY, R.S., Modulation and Demodulation for Probabilistic Coding; IEEE Trans, Vol IT-12, No 4 (July), pp 291-297.

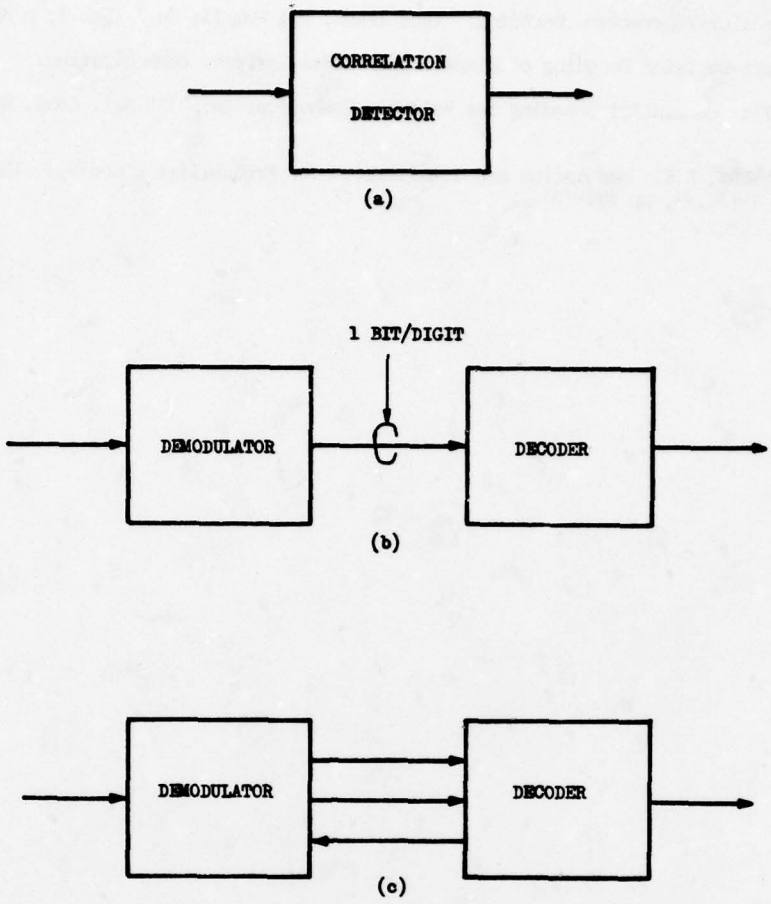


Fig.1 Detectors, demodulators and decoders

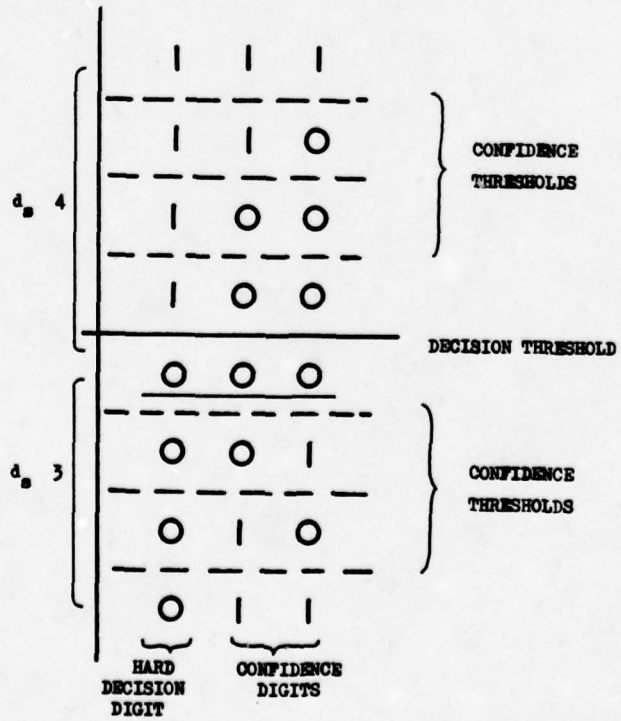


Fig.2 8-region soft-decision quantisation

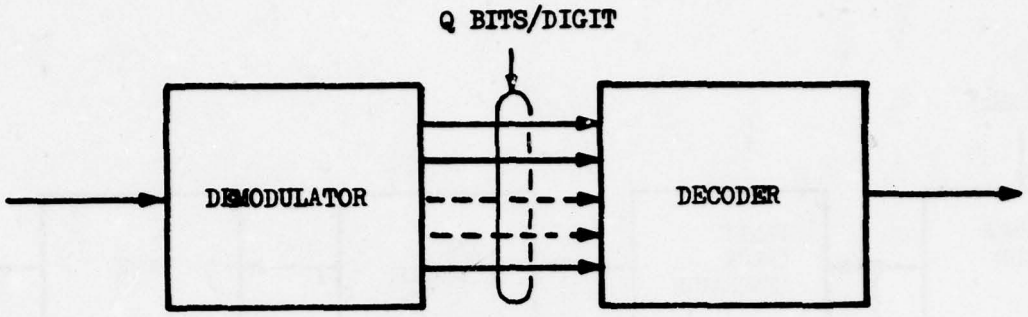


Fig.3 Soft-decision detector

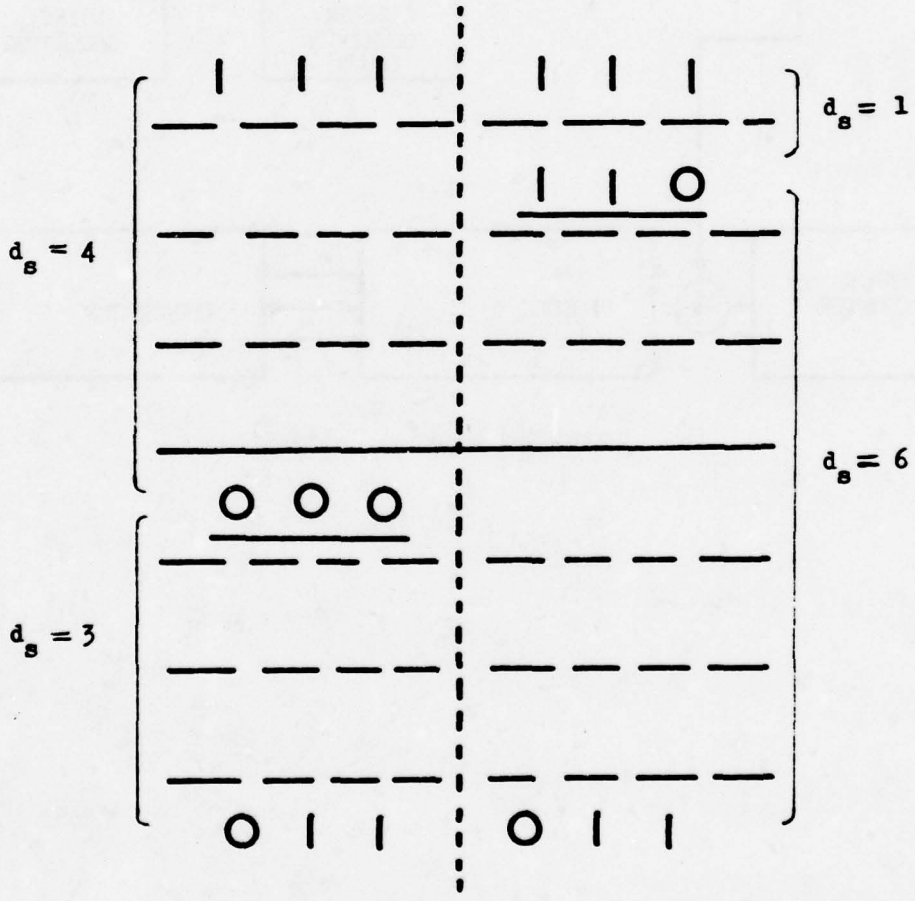


Fig.4 Soft-decision decoding example

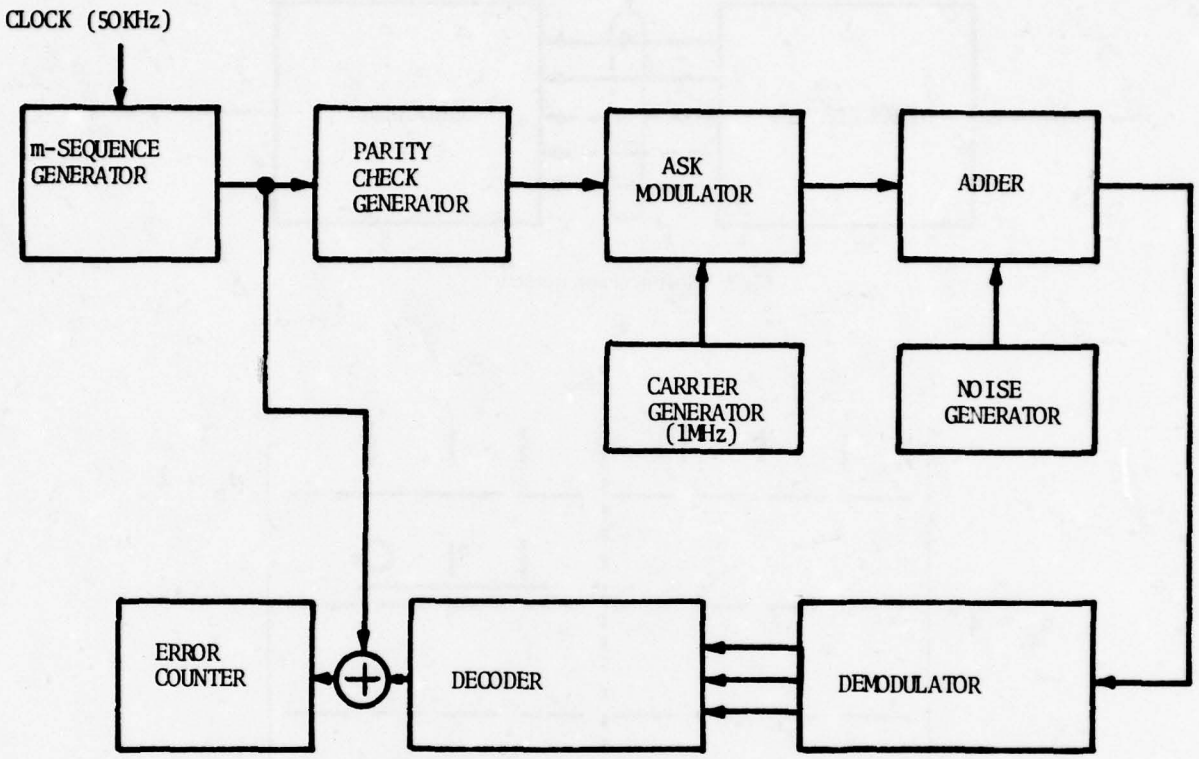


Fig.5 Experimental soft-decision SECA system

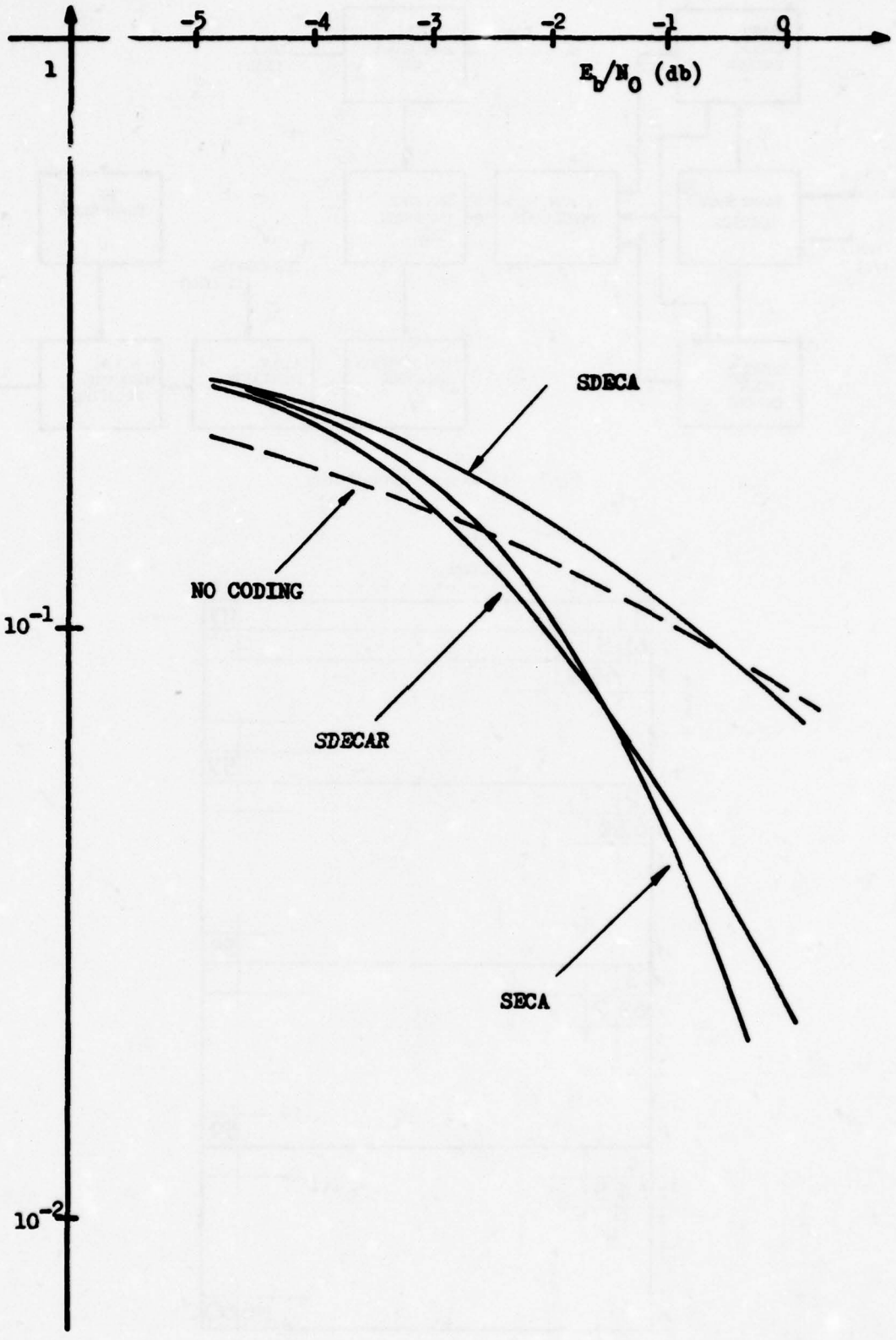


Fig.6 Performance of soft-decision algorithms

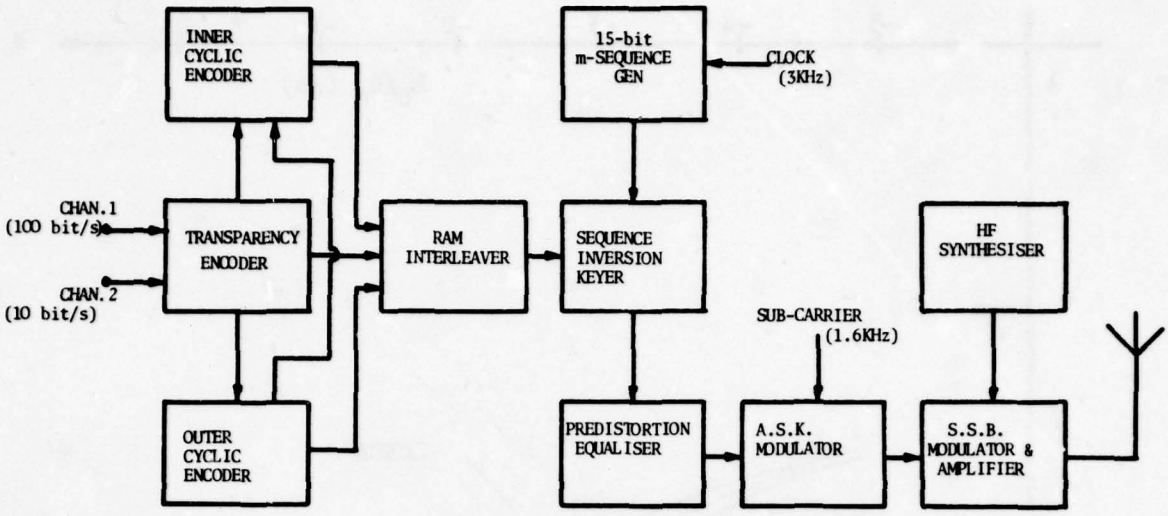


Fig.7 MSDD system transmitter

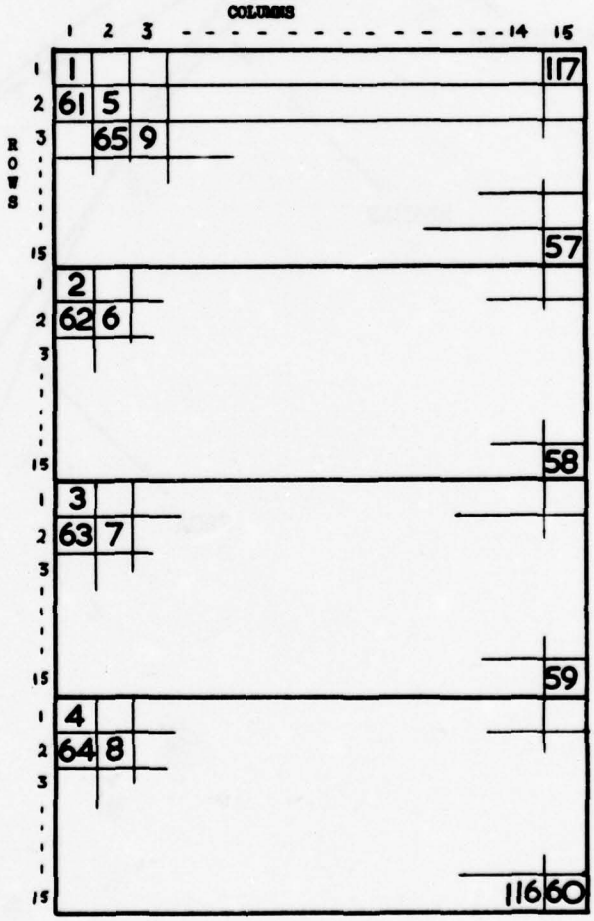


Fig.8 RAM organisation

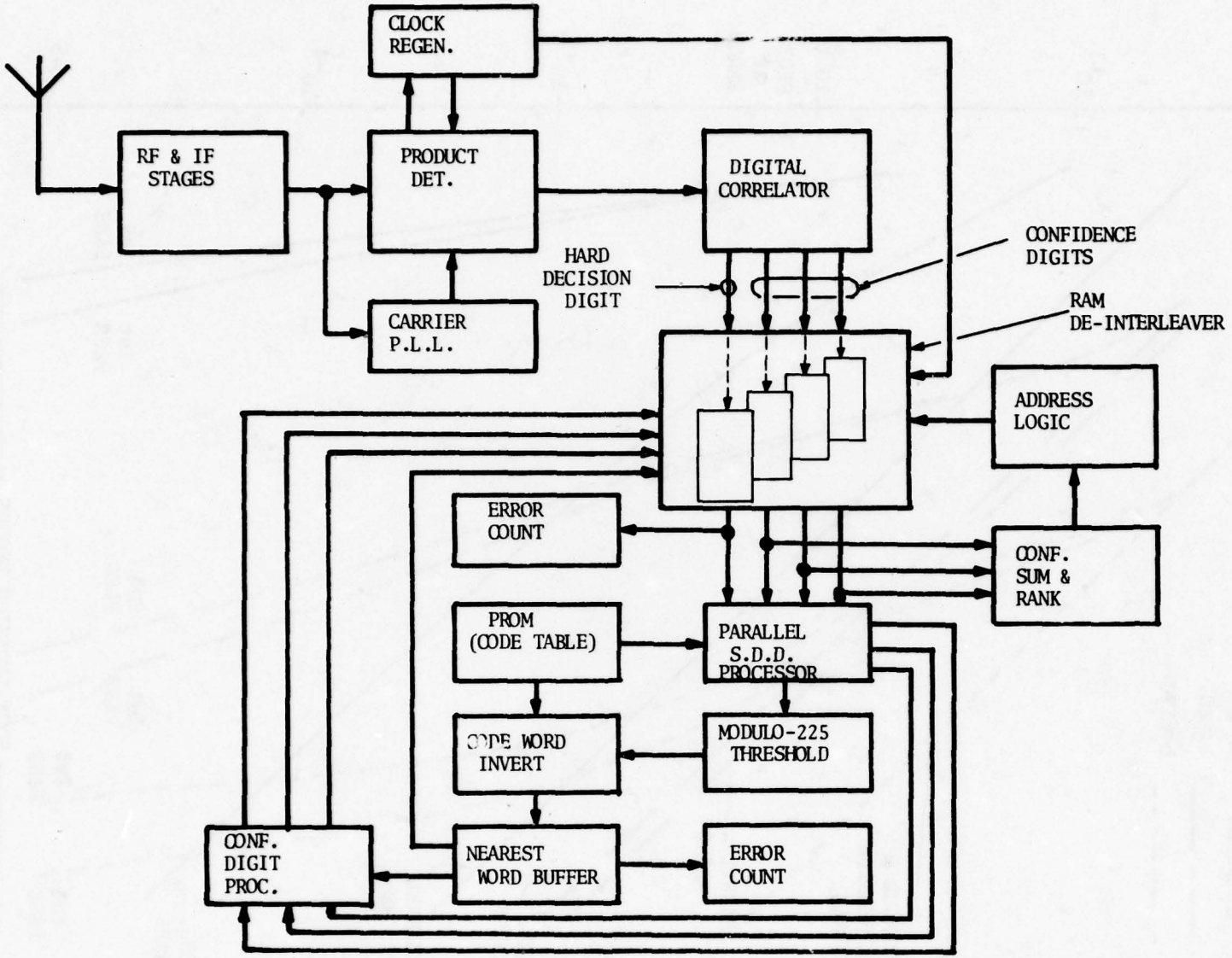


Fig.9 MSDD system receiver

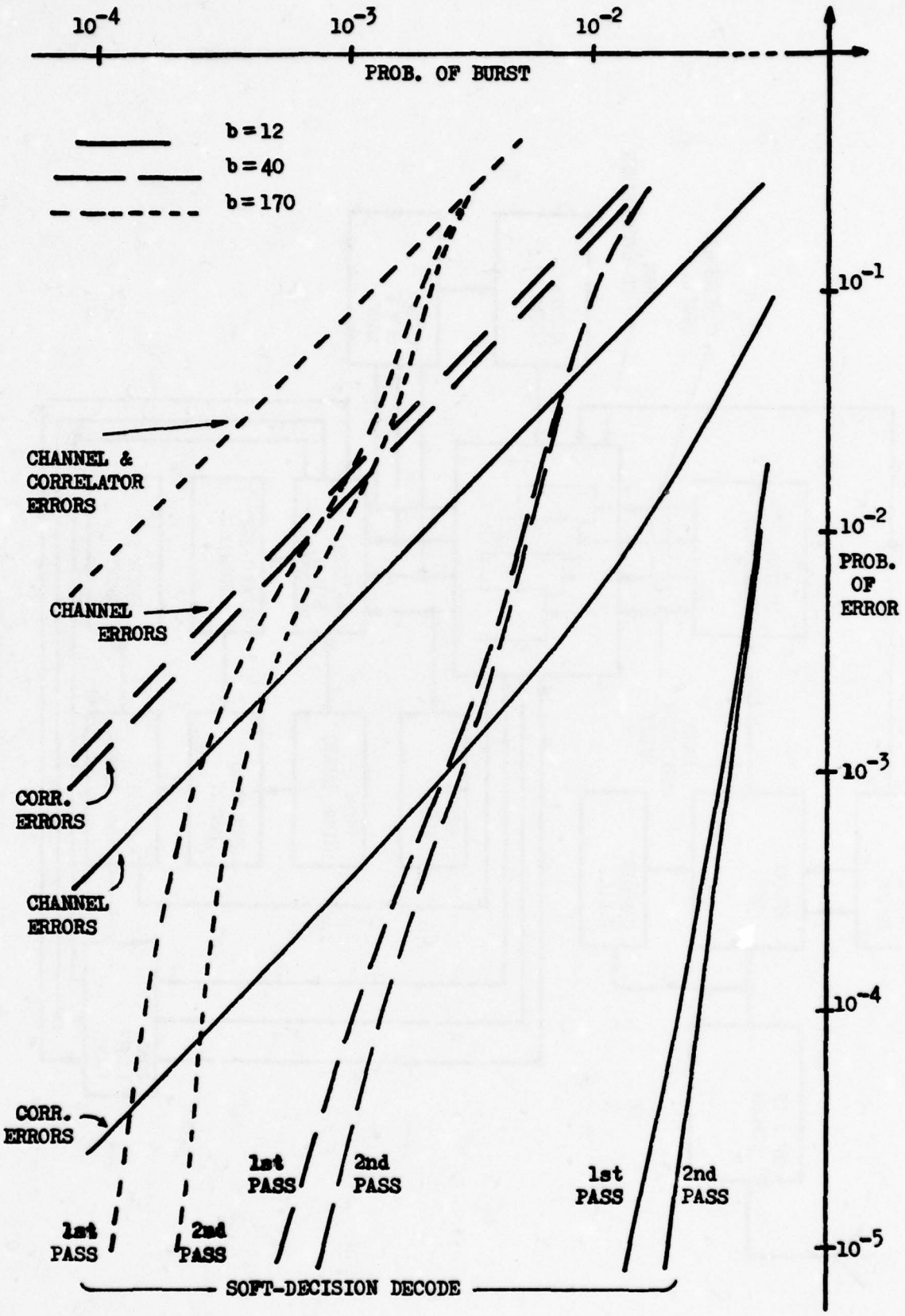


Fig.10 Digital tests: burst errors

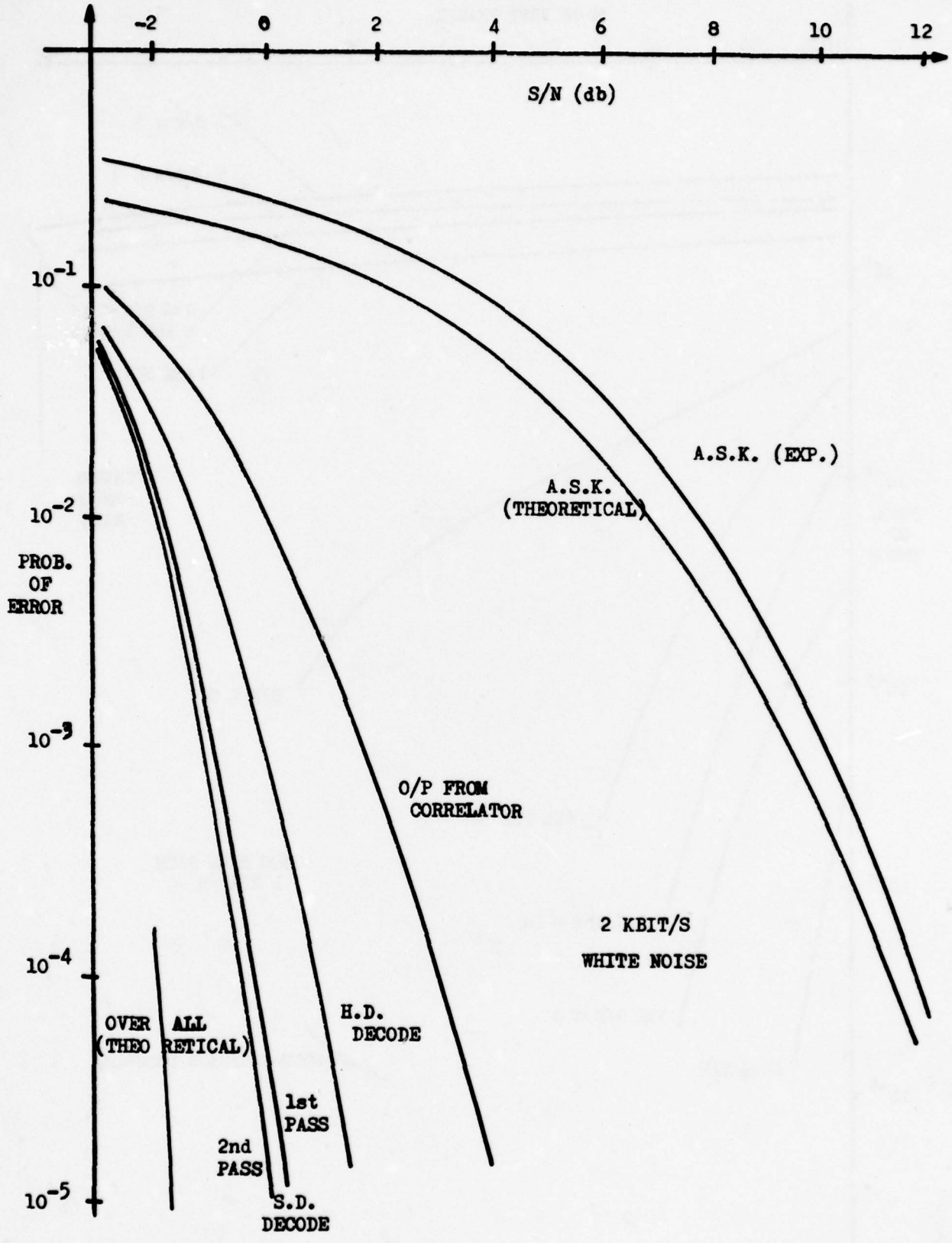


Fig.11 Whit noise tests

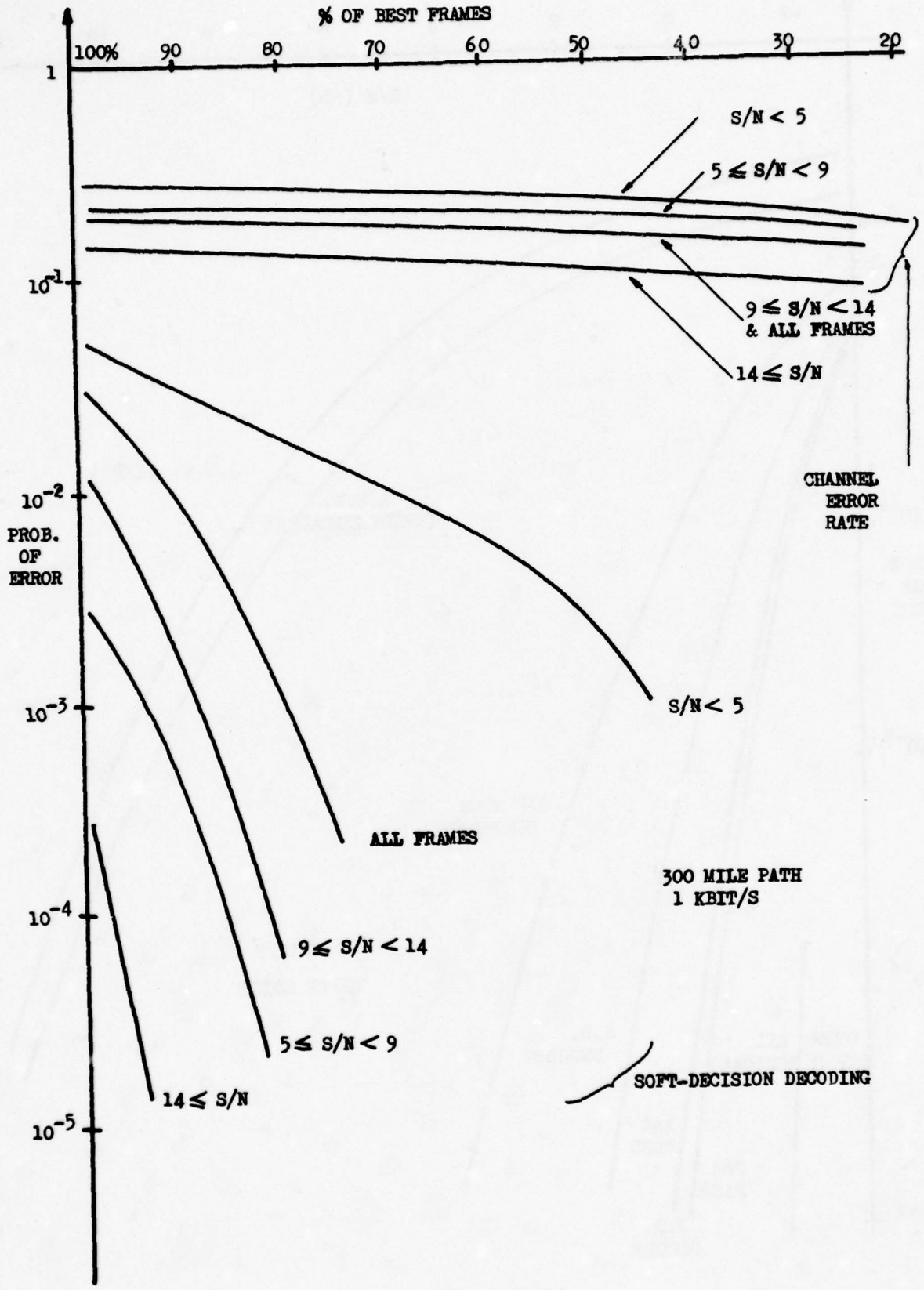


Fig.12 HF trials

DISCUSSION

D.Bosman, Ne

How does one calculate the upper bound of the number of soft decision bits, given a specific margin of confidence increase?

Author's Reply

There is no appreciable advantage in going to more than 4 or 5 soft-decision bits (i.e. of $\log_2 Q$ 4 or 5, or $Q = 16$ or 32). Use of 2 soft-decision bits achieves about 50–75% of the advantage of using 4 or 5 bits. Thus, in practice, if soft-decision can be used, 3 or 4 soft-decision bits are adequate. In other words, there is little need to calculate a bound; choice of code parameters is much more critical.

H.J.Matt, Ge

The performance of soft decision algorithms is shown as a function over (E_6/N_0). Since the Shannon capacity is at -1.6 dB how can it be explained that the curves go beyond that bound?

Author's Reply

Part of the answer is that the Shannon capacity bound assumes error-free operation at high signal-to-noise ratios, and is therefore not strictly applicable at low signal-to-noise ratios, and high error-rates. Further investigations at high error-rates of soft-decision algorithms is required.

THEORETICAL LIMITS ON CHANNEL CODING

UNDER VARIOUS CONSTRAINTS

by

B.G. Dorsch and F. Dolainsky

German Aerospace Research Establishment (DFVLR)
Institute of Communication Technology
D-8031 Oberpfaffenhofen

SUMMARY

Theoretical limits on coding under some fundamental constraints of practical schemes are calculated, which are not taken into account in Shannon's absolute limit of $E_b/N_o = -1.6$ dB for Gaussian noise. First channel capacity (zero error probability with codes of infinite length) is considered for finite rate codes (finite energy per symbol, finite bandwidth), binary input and quantized output signals. Then the influence of finite length and finite error probability on the theoretical limits of coding is investigated for three decoding philosophies (maximum likelihood decoding, bounded minimum distance decoding and decoding with an optimum fixed threshold). The performance of practical coding schemes under certain constraints should be compared to these limits rather than the absolute Shannon limit.

1. INTRODUCTION

The absolute theoretical limits of coding were given by Shannon in 1948 in his famous coding theorems, before serious coding work started. In the last few decades highly sophisticated coding/decoding procedures were developed, mainly for the additive white Gaussian noise channel AWGC, and compared to Shannons absolute limit. However, practical coding schemes do not approach this limit. The reason is simply that the comparison is not fair. All practical coding/decoding schemes are liable to some general and fundamental constraints which are not taken into account in Shannon's limit. For a fair comparison the theoretical limits under some of those fundamental constraints will be regarded in this paper. Mainly three constraints influence the channel capacity:

- In Shannon's derivation an infinite number of symbols has to be transmitted per time unit, resulting in an infinite bandwidth and a vanishing energy per symbol. In practical applications, besides bandwidth contraints, each transmission has to have a certain energy per symbol in order to maintain demodulation and symbol synchronization. Therefore coderates greater than zero will be regarded in Chapter 2.1.

- Under the assumption that only the mean power has to be limited, but not the power of each symbol, channel capacity for the AWGC is achieved by a Gaussian distributed input amplitude. But almost all coding schemes use binary input signals with constant power. Therefore systems with finite coderate and binary input are discussed in Chapter 2.2

- Practical decoding schemes do not use the received signal as an analog value but a finite quantization. The influence of output quantization on the channel capacity of finite rate binary input system is subject of Chapter 2.3.

While channel capacity considered so far does not regard the influence of finite codelength and finite error probability these constraints are considered in chapter 3.

2. CHANNEL CAPACITY OF THE AWGC UNDER VARIOUS CONSTRAINTS

First we will have a look on the curves and limits with which practical coding systems are usually compared. These are the Shannon limit of the AWGC and the biterror probability for binary input signals without channel coding. For binary transmission Binary Phase-Shift-Keying BPSK modulation with coherent demodulation is optimum with respect to transmission power (not constraint by bandwidth requirements), e.g. Wozencraft, Jacobs 1967. For transmission of binary signals x_1 and x_2 with

S = mean received signal power (= received power without noise)

T_s = *time per transmitted symbol*

E_s = $S \cdot T_s$ = mean received energy per symbol

N_o = onesided spectral noise power density

the output of the matched filter (integrate and dump) is a Gaussian distributed value y with mean $y_1 = +\sqrt{E_s}$, if x_1 is sent, resp. $y_2 = -\sqrt{E_s}$, if x_2 is sent, and variance $\sigma^2 = N_o/2$. The probability density function of y conditioned by x_1 resp. x_2 therefore ist

$$(1)\quad p(y/x_{1/2}) = (2\pi\sigma^2)^{-1/2} \cdot \exp\left((y-y_{1/2})^2/2\sigma^2\right)$$

The error probability $P_{e,\ uncoded} = Pr(y < 0/x_1) = Pr(y > 0/x_2)$ for equally probable x_1 and x_2 is plotted in fig. 1 as a function of E_s/N_o (here $E_b = E_s$). For example $P_{e,\ uncoded} = 10^{-5}$ can be achieved with $E_b/N_o = 9.6$ dB. E_b/N_o approaches infinity for $P_{e,\ uncoded} \to 0$.

Without channel coding each transmitted symbol can represent one bit of information. With T_s = time per information bit and E_b = mean received signal energy per information bit, we therefore have $T_b=T_s$, $E_b=E_s$ without channel coding. With channel coding, using blockcodes for example, instead of a block of K information digits (refered to as message) a longer codeword with $N>K$ symbols x is transmitted. The ratio $K/N =: R<1$ is called coderate. For a fair comparison with uncoded binary transmission a codeword (having N symbols) has to be transmitted with the same energy and in the same time as K information digit in the uncoded case. Therefore the time available for one symbol is $T_s=T_b \cdot R$ with 1/R times as much bandwidth required. The mean energy per symbol E_s also is reduced by the factor R, $E_s=S \cdot T_s=S \cdot T_b \cdot R=E_b \cdot R$. According to Shannon's coding theorem a zero error probability ($P_e \to 0$) can be achieved (with $N \to \infty$), if R is smaller than the channel capacity C (in bit per symbol), defined as (Shannon 1948)

$$(2)\quad C = \text{Max}\ (H(x)-H(x/y)) = \text{Max} \int_y \int_x p(x)p(y/x)\log \frac{p(y/x)}{p(y)}\, dx\, dy$$

This formula is based on random codes, where a codeword of length N is assigned randomly to each message. p(x) is the probability density function with which each of the N symbols of a codeword is selected from a given alphabet. The maximum is taken over all probability distributions which are possible within some given constraints. The constraint commonly used for the coding theorem is that only the expected (or mean) value of the transmitted signal power $\overline{S}$ is limited rather than the power S at each instant. Under this constraint channel capacity for the Gaussian channel is achieved, when the signal amplitude also has a Gaussian distribution. The capacity C in (2) then is, s. for example Gallager, 1968.

$$(3)\quad C = \frac{1}{2} \log_2 \left(1+\frac{2 \cdot S \cdot Ts}{N_o}\right) \text{ bit per symbol}$$

With $T_b = T_s \cdot R$, the requirement $R<C$ of the coding theorem yields

$$(4)\quad 1/T_b < (1/2T_s) \log_2(1+2 \cdot S \cdot T_s/N_o) =: C_{time}$$

the well known expression for the channel capacity per time unit (rather than per symbol, given in (3)) as upper limit for the number $1/T_b$ of error free transmittable bits of information per time unit.

For $T_s \to 0$ (or $R = T_s/T_b \to 0$) (4) results in the Shannon limit

$$(5)\quad S \cdot T_b/N_o = E_b/N_o > \ln 2 \quad \text{or} \quad -1.6 \text{ dB}$$

as absolute limit for any coding/decoding system for the AWGC.

1.1 Coderate $R>0$

Because the channel capacity given in (3) does not depend on the coderate, the requirement $R<C$ for finite rates $R = ST_s/ST_b > 0$ results in

$$(6)\quad S \cdot T_b/N_o = E_b/N_o > (2^{2R}-1)/2R$$

for the analog (Gaussian) input, AWGC, plotted as dotted line in fig. 2. For R=1/2 for example we get $E_b/N_o > 0$ dB. For R=1/4 the limit is $E_b/N_o = -.8$ dB, only .8 dB worse than the absolute Shannon limit of -1.6 dB for $R \to 0$.

1.2 Coderate $R>0$ and binary input

For binary input symbols x_1 and x_2 and analog output values y with conditioned probabilities p(y/x) given in (1) the channel capacity (2) is

$$(7)\quad C = \text{Max} \int_y \sum_x P(x)p(y/x) \log \frac{p(y/x)}{p(y)}\, dy$$

with $p(y) = P(x_1)p(y/x_1)+P(x_2)p(y/x_2)$. The maximum is achieved for equally probable x, $P(x_1) = P(x_2)=1/2$, because of symmetry. p(y/x) and therefore p(y) and C are functions of E_s/N_o. The requirement $R<C$ with $R = E_s/E_b$ yields

$$(8)\quad E_b/N_o > (E_s/N_o)/C$$

also a function of E_s/N_o. Because $R<C(E_s/N_o)$ and E_b/N_o are functions of E_s/N_o, R can be given as function of E_b/N_o using E_s/N_o as parameter. Since E_b/N_o as function of R cannot be evaluated analytically, the solution was computed numerically. The results for finite coderates, binary input, analog output are plotted as lower solid line in fig. 2. We see that the difference in minimum required E_b/N_o compared to Gaussian input symbols is negligible for small coderates. But for high coderates it may be worthwhile to search for Gaussian distributed input coding/decoding schemes, which do not yet exist.

1.3 Coderate $R>0$, binary input, quantized output

In practical applications the analog received values y are quantized. All values y in a range $y_{j-1} \leq y < y_j$ are represented by one of J integer numbers z_j, $j = 1,2,\ldots J$, and $y_o = -\infty$, $y_J = +\infty$. The number J and the arrangement of the quantization limits y_j is of much influence on channel capacity. Usually J is a power of 2. Optimum equidistant (with $y_j-y_{j-1}=\Delta y$ = const. for $1<j<J$) and optimum nonequidistant quantization spacing will be investigated. For binary input symbols x_1 and x_2 and $p(y/x_{1/2})$ given in (1) we have

$$P(z_j/x_{1/2}) = \Pr\ (y_{j-1} \leq y < y_j) = \int_{y_{j-1}}^{y_j} p(y/x_{1/2})\, dy \text{ and}$$

(9) $P(z_j) = P(x_1) \cdot P(z_j/x_1) + P(x_2)\, P(z_j/x_2)$

Instead of the analog capacity $C = C_{J=\infty}$ given in (7) we now have

(10) $C_J = \text{Max} \sum_z \sum_x P(x) P(z/x) \log \frac{P(z/x)}{P(z)}$

Because of symmetry of $p(y/x_{1/2})$ C_J is maximized by $P(x_1)=P(x_2)=1/2$ and a symmetrical arrangement of the quantization limits y_j around y=0. For equidistant quantization, C_J is furthermore a function of the quantization spacing $\Delta y = QF \cdot \sigma$, where $\sigma^2 = N_o/2$ is the variance of y and QF we call quantization factor. The optimum QF as a function of E_s/N_o with J as parameter is plotted in fig. 3. In practical applications QF will be a fixed value, adjusted for the lowest operational E_s/N_o. The loss in capacity for higher E_s/N_o then is negligible (Dolainsky, 1971).

The optimum nonequidistant quantization spacing has been calculated by trial and error.
Because the y_j are symmetric around y=0, equidistant and nonequidistant spacing are the same for $J \leq 4$.
For $J \geq 5$ the difference between the C_J of optimum equidistant and optimum nonequidistant quantization turns out to be less than 1/100 dB (!) for $E_s/N_o \geq -3$ dB (Dorsch, 1971) and therefore is negligible.

The minimum required E_b/N_o as function of the coderate R for optimum equidistant quantization was calculated in the same way as for analog output signals in 2.2. The results for J = 2,4 and 8 quantization levels are plotted in fig. 2. It can be seen that J = 8 is almost as good as $J = \infty$. The difference in E_b/N_o between J = 2 (i.e. binary decisions) and $J = \infty$ is roughly 2 dB. The limit for $R \to 0$ and J = 2 is $E_b/N_o = +.4$ dB (rather than -1.6 dB for $J = \infty$). Coderates R below 1/4 do not gain much in E_b/N_o for all J. Since demodulators and bitsynchronizers have operational limits in the order of $E_s/N_o \approx -1$ dB (due to some unavoidable phase jitter), no coderates R much below .5 may be used for systems working near the theoretical limit of E_b/N_o.

Finally fig. 2 displays the influence of finite coderates, binary input and finite output quantization on the theoretical limits of E_b/N_o for the AWGC.

3. CODING BOUNDS FOR FINITE CODELENGTH AND FINITE ERROR PROBABILITY

In this chapter the influence of finite codelength N and finite error probability P_e on theoretical bounds of coding will be investigated. First the Binary Symmetric Channel BSC with symbol error probability p will be regarded. Then the results are applied to the AWBC (binary input, binary output only), when p is a function of E_b/N_o and the coderate.

For a simple notation let $\underline{x}$ be any codeword of length N, $\underline{x}'$ the transmitted codeword, $\hat{\underline{x}}$ the decoding decision, $\underline{y}$ the received binary word and $d(\underline{x},\underline{y})$ the Hamming distance between $\underline{x}$ and $\underline{y}$. Coding bounds for finite P_e and N depend very much on the decoding philosophy. Three decoding philosophies will be considered:

- Maximum Likelihood Decoding (MLD) is optimum with respect to P_e, when all codewords are sent with equal a priori probabilities. For the BSC the decoding decision $\hat{\underline{x}}$ is the codeword with smallest distance $d(\hat{\underline{x}}, \underline{y}) \leq d(\underline{x},\underline{y})$ of all codewords $\underline{x}$. A decoding estimate $\hat{\underline{x}}$ is made for each received $\underline{y}$. The coding theorem usually is proofed by MLD arguments for random codes.

- Bounded Minimum Distance Decoding (BMD) is used by most algebraic block decoding procedures for codes with a guaranteed minimum Hamming distance $\geq D$ between codewords. A decoding estimate $\hat{\underline{x}}$ is made only if there exists a codeword $\underline{x}$ within distance $d(\underline{x},\underline{y}) < D/2$. Otherwise a detected error, i.e. no decoding estimate $\hat{\underline{x}}$, is made. The disadvantage of BMD is that detected errors contribute so much to the error probability that channel capacity cannot be achieved, as will be seen in detail later.

- With Optimum Threshold Decoding (OTD), (Dorsch 1977), a decoding decision $\hat{\underline{x}}$ is made if any codeword can be found within a distance $d(\underline{x},\underline{y}) < T$. For random codes there is an optimum threshold T as will be shown later. Threshold arguments are used in Shannon's paper, 1948, on the coding theorem with a more detailed proof given by Massey 1977. A very simple proof of the coding theorem for the BSC with threshold arguments was given by Van Lint, 1973.

For the BSC and a certain coding/decoding principle the error probability P_e depends on p,R,N. If $P_e(p,R,N)$ is written in the form

(11) $P_e = 2^{-N \cdot E(p,R,N)}$ then the limit value

(12) $\lim_{N\to\infty} E(p,R,N) := \lim_{N\to\infty} (-1/N) \log P_e =: E(p,R)$

is called the error exponent E.

Upper and lower bounds on E(p,R) can be derived. These bounds mean that there are no codes with $E > E_{upper}$, but there must exist codes with $E \geq E_{lower}$, or: The error exponent of the best code lies in the range $E_{lower} \leq E_{best\ code} \leq E_{upper}$. For finite but large N the error exponent gives an estimate $P_e \approx 2^{-N \cdot E(p,R)}$ of the error probability. The bounds of the error exponent now will be investigated for the three decoding principles mentioned before.

3.1 Maximum Likelihood Decoding

Upper and lower bounds of the error exponent E for MLD can be found for example in W.W. Peterson, E.J. Weldon 1972. Let the coderate R be given in parametric form by

(13) $R = 1-H(v)$ with v as parameter in the range $p \leq v < 1/2$

where $H(v) = -v \log v - (1-v)\log(1-v)$ is the binary entropy function. Then (13) represents R in the interesting range between $R(v=1/2) = 0$ and $R(v=p) = 1-H(p)=C(p)$, the channel capacity of the BSC. With the definitions

(14) $\gamma(v,p) := \log_2 \left[(p/(1-p))^v \ ((1-p)/(1-v))^{1-v} \right]$ and

(15) $q := 1-p$

the best known lower bound of E for MLD is given in three ranges

(16) $$E_{lower} = \begin{cases} E_1 = y \cdot \log\ (1/\sqrt{4pq}) & \text{for } 1/2 > v \geq v_1 := \sqrt{4pq}\ /(1+\sqrt{4pq}\ ) \\ E_2 = \log(2/(1+\sqrt{4pq}\ )) - R(v) & \text{for } v_1 \geq v \geq v_2 := \sqrt{p}\ /(\sqrt{p}+\sqrt{q}) \\ E_3 = V(v,p) & \text{for } v_2 \geq v \geq p \end{cases}$$

The best upper bound of E for MLD is

(17) $E_{upper} = \text{Min}\left[V(v,p),\ \text{straight line through } E_1(v=1/2) \text{ tangential to } V(v,p) \right]$

Those upper and lower bounds of E for MLD as functions of R with p as parameter are plotted in figs. 4,5,6 (for $p = 10^{-1}, 10^{-2}, 10^{-3}$). Upper and lower bound are fairly close together for rates $0 < R < R(v_2)$. For $R=0$ and $R(v_2) \leq R \leq C(p)$ upper and lower bound are the same. E_{MLD} is positive for all coderates $R < C(p)$ resulting in $P_e \to 0$ for $N \to \infty$.

3.2 Bounded Minimum Distance Decoding

With BMD a decoding error occurs, if and only if the received word $\underline{y}$ has D/2 or more symbol errors. The error probability therefore is given by the binomial expression

(18) $$P_e = \sum_{i=D/2}^{N} \binom{N}{i} \cdot p^i \cdot q^{N-i}$$

with V(v,p) defined in (14), P_e in (18) can be bounded by a Chernoff technique, e.g. W.W. Peterson, E.J. Weldon 1972.

(19) $P_e \leq 2^{-N \cdot V(D/2N,p)}$ valid for $D/2N > p$ and all N.

This bound is asymptotically tight for high values of N resulting in

(20) $E_{BMD} = V(D/2N,p)$

Upper and lower bounds of E_{BMD} as functions of the coderate R can be calculated using upper and lower bounds of R as functions of D/N. The best known lower bound on R(D/N) given by Varshamow, 1957, and Gilbert, 1952, guarantees that there exist codes with rate

(21) $R \geq R_{VG} = 1-H(D/N)$

for high values of N. An upper bound on R(D/N) , given by Elias, 1960, establishes, that all long codes have rate

(22) $R < R_{E1} = 1 - H(.5 - \sqrt{.25-D/2N}\ )$

The corresponding upper and lower bounds of E_{BMD} as functions of R with p as parameter are plotted in figs. 4,5,6 (for $p=10^{-1}, 10^{-2}, 10^{-3}$). The Varshamow-Gilbert lower bound of E_{BMD}, based on (21), seems to be a more realistic estimate of the coderate of real codes than the Elias upper bound: Almost all long random codes have $R \approx R_{VG}(D/N)$, Wyner 1969. Furthermore a binary BCH-Code of length N = 1023 has $R \approx R_{VG}$. For high values of p, as can be seen from fig. 4, E_{BMD} is much worse than E_{MLD} and becomes zero for rates R far below capacity C(p). Therefore with BMD, which is used in most practical decoding procedures for algebraic block codes, channel capacity can never be achieved.

3.3 Optimum Threshold Decoding of Random Codes

OTD is a suboptimum decoding rule which takes any codeword within a fixed distance T from the received word as decoding decision rather than the most probable one. For random codes of rate R there is an optimum threshold T which minimizes the error probability $P_e \leq P(T) := P_1 + P_2$, where

(23) $$P_1 := \Pr\left[d(\underline{x}', \underline{y}) \geq T \right] = \sum_{i=T}^{N} \binom{N}{i} p^i q^{N-i}$$

is the probability that the distance between the received word $\underline{y}$ and transmitted codeword $\underline{x}'$ becomes larger than threshold. An error also may be caused by OTD, when any of the $(2^{N \cdot R}-1)$ other random codewords $x \neq x'$ fall within a distance $d(\underline{x},\underline{y}) < T$. This probability is overbounded with a union bound by

$$(24)\quad P_2 := 2^{N\cdot R}\cdot \sum_{i=0}^{T-1} \binom{N}{i}/2^N$$

The sum $P(T)=P_1+P_2$ becomes minimum, when

$$(25)\quad P(T+1)-P(T)=\binom{N}{T}\left[2^{NR-N}-p^T q^{N-T}\right]=0$$

resulting in an optimum $T = T_o$ for

$$(26)\quad T_o/N =: t_o = (C(p)-R)/\log_2(q/p) + p$$

Bounding (23) and (24) for $T=T_o$, using Chernoff techniques (19), results in

$$(27)\quad P_e \leq P(T_o) \leq 2^{-N(V(t_o,p)+1/N)} \text{ valid for } t_o > p \text{ and all } N$$

From (26) we see that $t_o > p$ for all $R < C(p)$, $p < 1/2$. For $N \to \infty$ and the error exponent E defined in (12), eq. (27) gives as lower bound of E

$$(28)\quad E_{OTD,\ lower} = V(t_o,p)$$

also plotted in figs. 4,5,6. As upper bound for OTD the upper bound of MLD may be used, which is an upper bound for all decoding schemes.

For rates R close to capacity C(p) the error exponent E_{OTD} approaches E_{MLD}. That means that for rates close to capacity the suboptimum decoding principle OTD is as good as the optimum MLD for the BSC. Compared to BMD we see that OTD is superior for high values of p or R. For $p > .08$ we have $E_{OTD} > E_{BMD}$ for all coderates. The reason of $E_{OTD} < E_{BMD}$ for low p and R is due to the randomness of the codes used for the OTD estimates. For codes with guaranteed minimum distance D of course an optimum threshold T_o results in a better error exponent than the fixed threshold D/2 of BMD.

For example N=1000, $P_e \approx 2^{-N\cdot E} \leq 10^{-5}$ (i.e. $E \geq .0166$) may be achieved for $p = 10^{-1}$ (fig. 3) by MLD (using random codes) with $R \leq .39$ (lower and upper bound), by OTD (random codes) with R = .37 (lower bound), but with BMD and the more realistic lower (Varshamow-Gilbert) bound only with R = .12 (e.g. using BCH-Codes, where $R \approx R_{VG}$ for $N \approx 1000$)

There is not much hope that MLD algorithms will be found which practically can be applied to long codes of medium rate. But there is some hope that constructive OTD procedures will be invented which extend BMD algorithms beyond the threshold D/2 by allowing a certain ambiguity of the decoding decision $\hat{x}$. The performance of asymptotically good classes of well structured algebraic codes, which are BMD decodable and are assumed to have fairly random distance properties, could become close to capacity with OTD.

3.4 Results for the AWGC with hard binary decisions

With binary input, binary quantized output, the AWGC also is a BSC where the symbol error probability p is a function of E_s/N_o, $p = f(E_s/N_o)$, plotted as P_e (uncoded) in fig. 1. Because this function is monotonic, the inverse function $E_s/N_o = f^{-1}(p)$ is unique. So $E_b/N_o = E_s/N_o \cdot R = f^{-1}(p)/R$ and the error exponent $E(p,R)$ are functions of p and R. For fixed rates R = 0.8, 0.5, 0.1 the error exponent E as function of E_b/N_o is plotted in figs. 6,7,8 for the three decoding principles MLD, BMD, OTD.

For R=.5 (fig. 8) E=.0166 (i.e. $P_e = 2^{-N\cdot E} = 10^{-5}$, N=1000) may be achieved by MLD with E_b/N_o=3.2 dB, by OTD with 3.5 dB but with BMD (using the more realistic Varshamow-Gilbert bound) only with 5.7 dB.

There is an optimum coderate (with minimum E_b/N_o) for BMD. Errors can be corrected effectively by BMD only if the maximum number of correctable errors D/2 is greater than the expected number of errors $p\cdot N$ in a codeword. Therefore $E_s/N_o = f^{-1}(p) \geq f^{-1}(D/2N)$ or

$$(29)\quad E_b/N_o = E_s/N_o \cdot R \geq f^{-1}(D/2N)/R$$

is required. With R as function of D/N there is an optimum D/N, which minimized E_b/N_o. For $R = R_{VG}(D/N)$ given in (21), resp. $R = R_{E1}(D/N)$ in (22), the minimum values of E_b/N_o are

$$(30)\quad \begin{aligned} (E_b/N_o)_{VG,opt} &= 4.1 \text{ dB} \quad \text{for } R = R_{VG,opt} = .54 \quad \text{resp.} \\ (E_b/N_o)_{E1,opt} &= 2.2 \text{ dB} \quad \text{for } R = R_{E1,opt} = .43 \end{aligned}$$

The Varshamow-Gilbert bound (VG), resp. Elias bound (E1), here mean: There exist long codes with $E_{BMD} > 0$ (i.e. $P_e \to 0$ for $N \to \infty$) using BMD for $E_b/N_o \leq 4.1$ dB but not for $E_b/N_o < 2.2$ dB. According to (30) coderates $R \approx .5$ are near optimum for BMD.

For MLD and OTD the required E_b/N_o becomes smaller for smaller R, if E is close to zero. But for finite E (i.e. finite N, P_e) there are optimum coderates also for MLD and OTD. For example for E = .0166 (corresponding to $P_e = 10^{-5}$, N=1000) the optimum coderate for MLD is R = .25 resulting in E_b/N_o = 3.1 dB. (whereas E_b/N_o (R=.1)=3.7 dB, E_b/N_o (R=.01) = 7.5 dB!)

4. CONCLUSION

The influence of some combinations of binary input, quantized output, coderate, codelength, error probability and various decoding philosophies on the minimum required E_b/N_o for the AWGC was calculated and plotted in figs. 4-9. Practical coding/decoding schemes under such constraints should be compared to those limits rather than the absolute Shannon limit of -1.6 dB. Some of the results are summerized in the following table for a concised survey.

INPUT	OUTPUT	N	P_e	R	DECOD.	E_b/N_o min.
Gaussian	analog	$\to\infty$	$\to 0$	$\to 0$	MLD,OTD	-1.6 dB
binary	"	"	"	"	"	-1.6 dB
binary	binary	"	"	"	"	+ .4 dB
Gaussian	analog	"	"	.5	"	0 dB
binary	analog	"	"	.5	"	.2 dB
"	J=8	"	"	.5	"	.3 dB
"	J=4	"	"	.5	"	.6 dB
"	binary(J=2)	"	"	.5	"	1.8 dB
"	"	≤ 1000	$\leq 10^{-5}$	.5	MLD	3.2 dB
"	"	"	"	"	OTD	3.5 dB
"	"	"	"	"	BMD_{VG}	5.7 dB

Limits of E_b/N_o under some constraints

REFERENCES

DOLAINSKY, F., DORSCH, B.G., 1971 — "Transmission Limits for the Gaussian Channel with Finite Rate Codes" Second Internat. Symposium on Information Theory, Tsachkadsor, Armenia, UdSSR, 1971

DORSCH, B.G., 1971 — "Optimum Quantization for the Gaussian Channel" Second Internat. Symposium on Information Theory, Tsachkadsor, Armenia, UdSSR, 1971

" " 1977 — "Decoding of Random Codes with an Optimum Threshold" IEEE-Internat. Symp. on Inf. Theory Ithaca, New York, Oct. 1977

GALLAGER, R.G., 1968 — "Information Theory and Reliable Communication" J. Wiley and Sons, New York

MASSEY, J.L., 1977 — "Shannon's "Proof" of the Noisy Coding Theorem" IEEE-Internat. Symposium on Information Theory, Oct. 1977, Cornell University, Ithaca, NY, USA

PETERSON, W.W., WELDON, Jr., E.J., 1972 — "Error Correcting Codes" The M.I.T. Press, Cambridge, Mass.

SHANNON, C.E., 1948 — "A Mathematical Theory of Communication" Bell System Tech. J., 27; Math. Rev., 10.

VAN LINT, J.H., 1973 — "Coding Theory" Springer-Verlag, Berlin

WOZENCRAFT, J.M., JACOBS, I.M., 1967 — "Principles of Communication Engineering" J. Wiley and Sons, New York

WYNER, A.D., 1969 — "On Coding and Information Theory" SIAM Review, Vol. 11, No. 3, July 1969

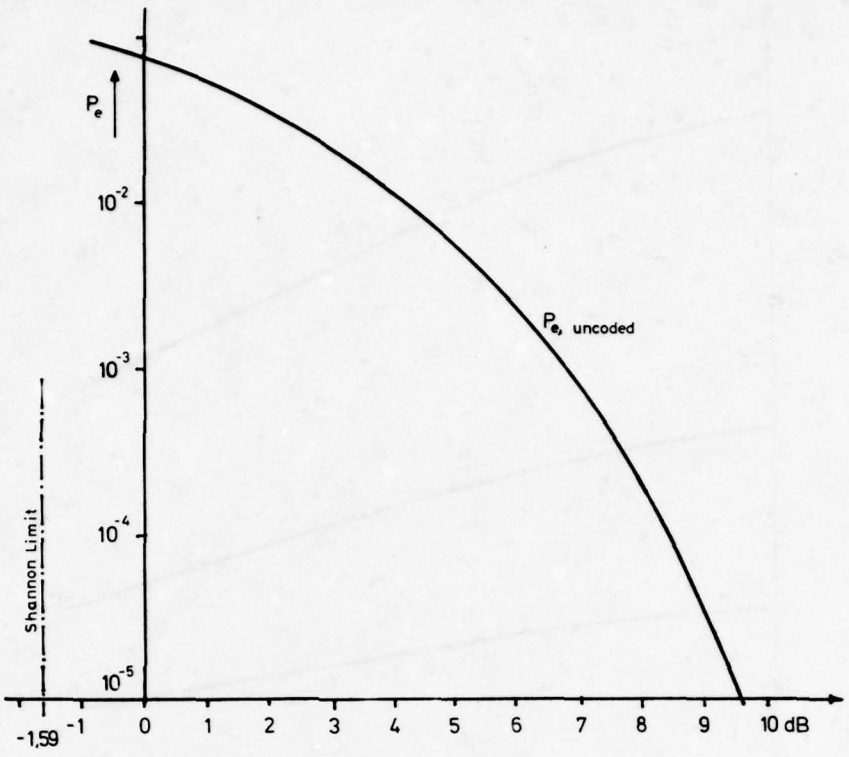


Fig. 1: Shannon's Limit and $P_{e,\ uncoded}$ vrs. E_b/N_o for the AWGC

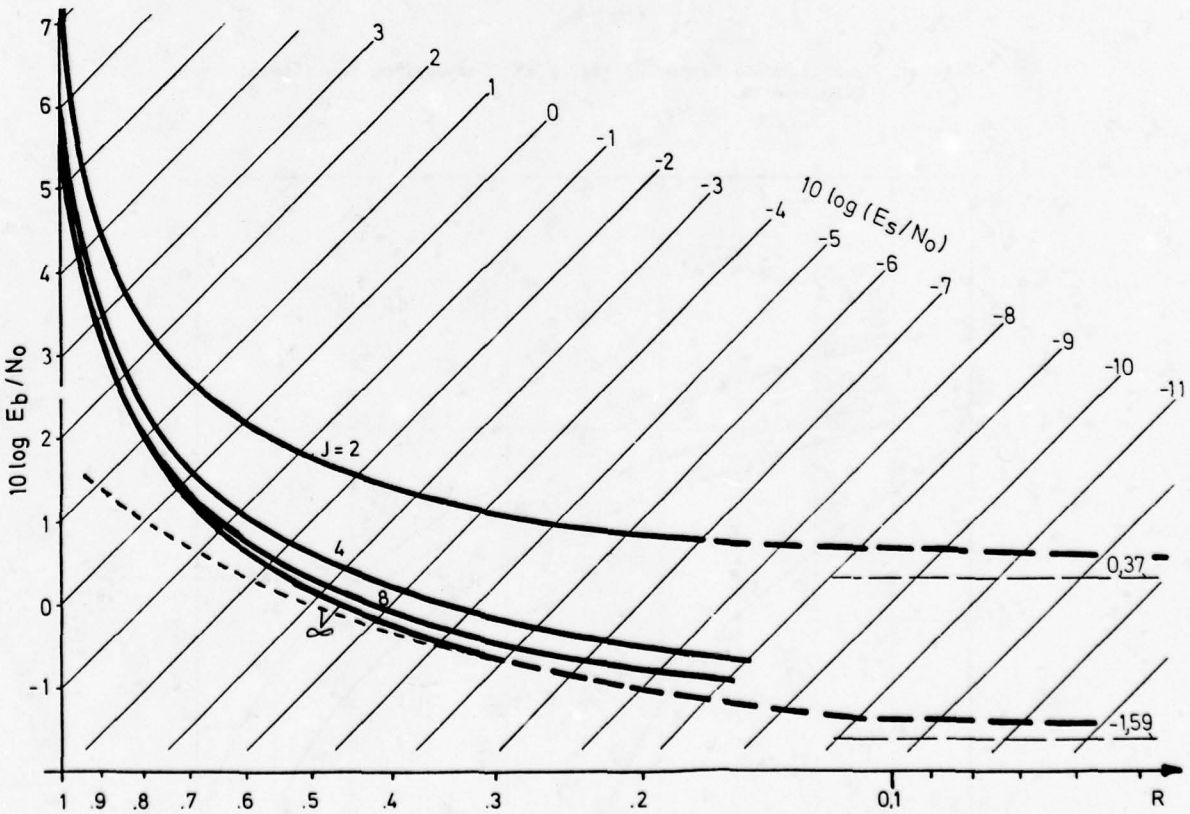


Fig. 2: Bounds of E_b/N_o vrs. Coderate R for the AWGC with Binary Input (solid lines), Gaussian Input (dotted line), J-ary Output (Optimum Equidistant Quantization)

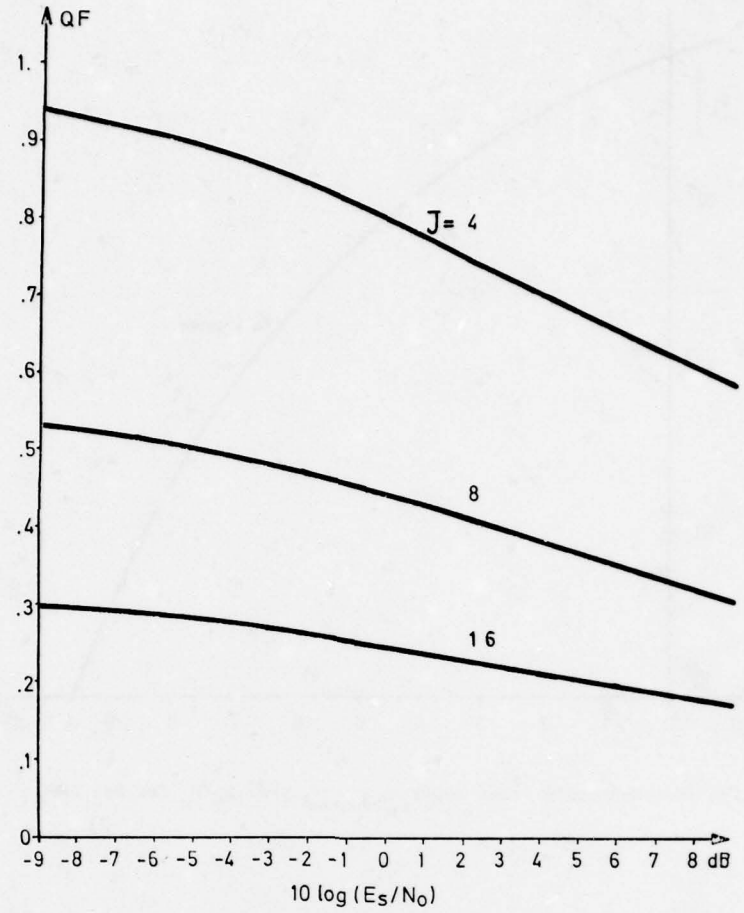


Fig. 3: Quantization Factor QF vrs. E_s/N_o for Optimum Equidistant Quantization

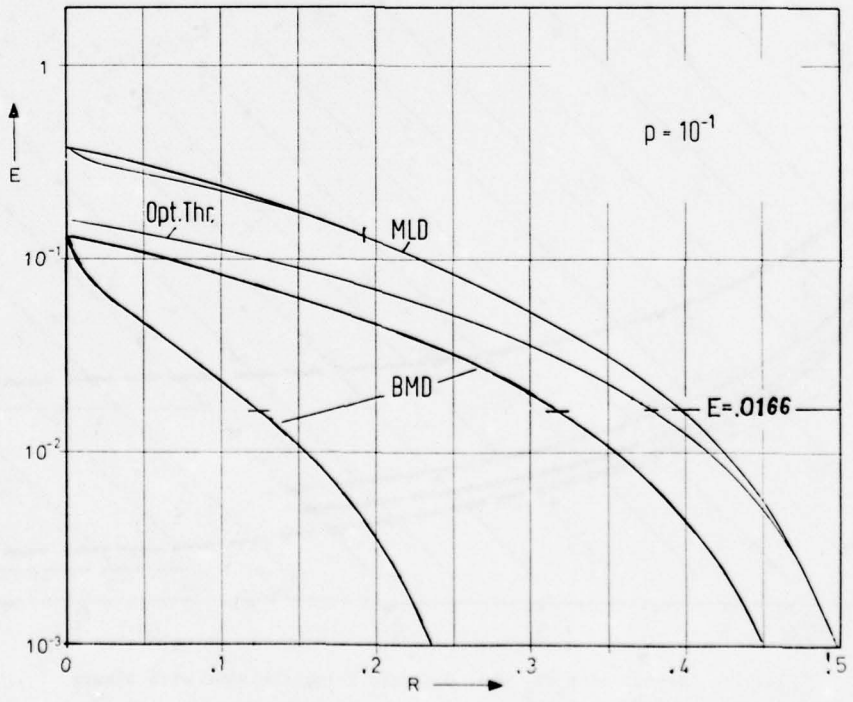


Fig. 4: Error Exponents E (for MLD, BMD, OTD) vrs. Coderate R for $p = 10^{-1}$

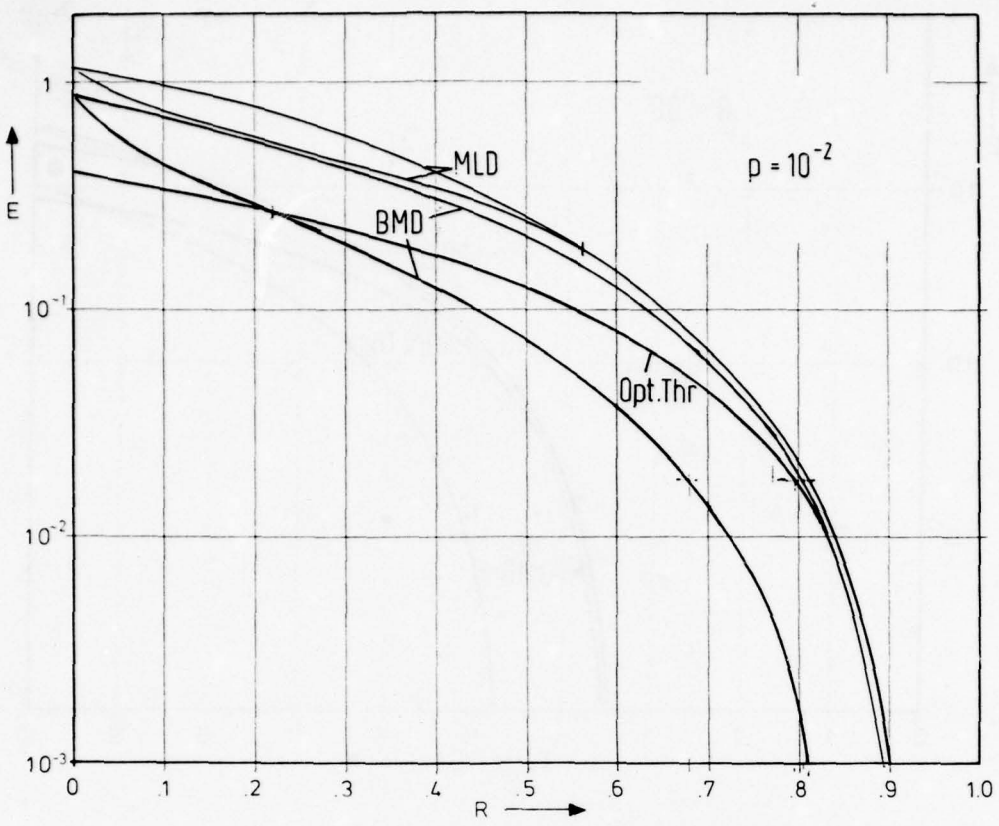


Fig. 5: Error Exponents E (for MLD, BMD, OTD) vrs. Coderate R for $p = 10^{-2}$

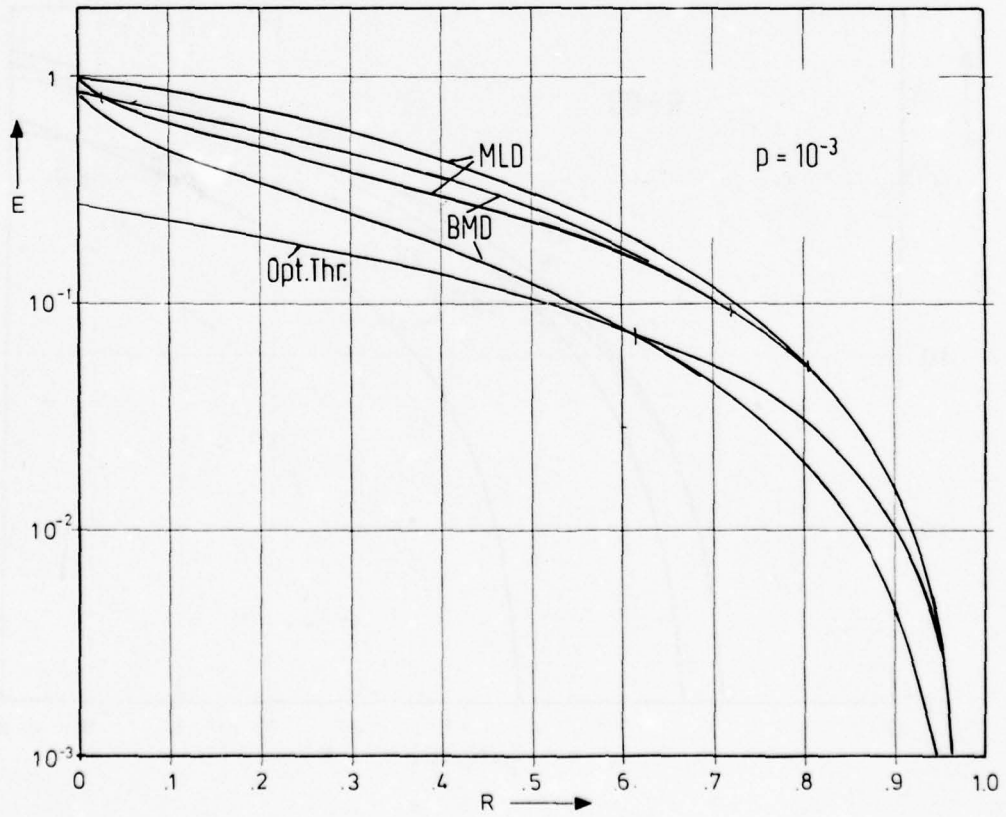


Fig. 6: Error Exponents E (for MLD, BMD, OTD) vrs. Coderate R for $p = 10^{-3}$

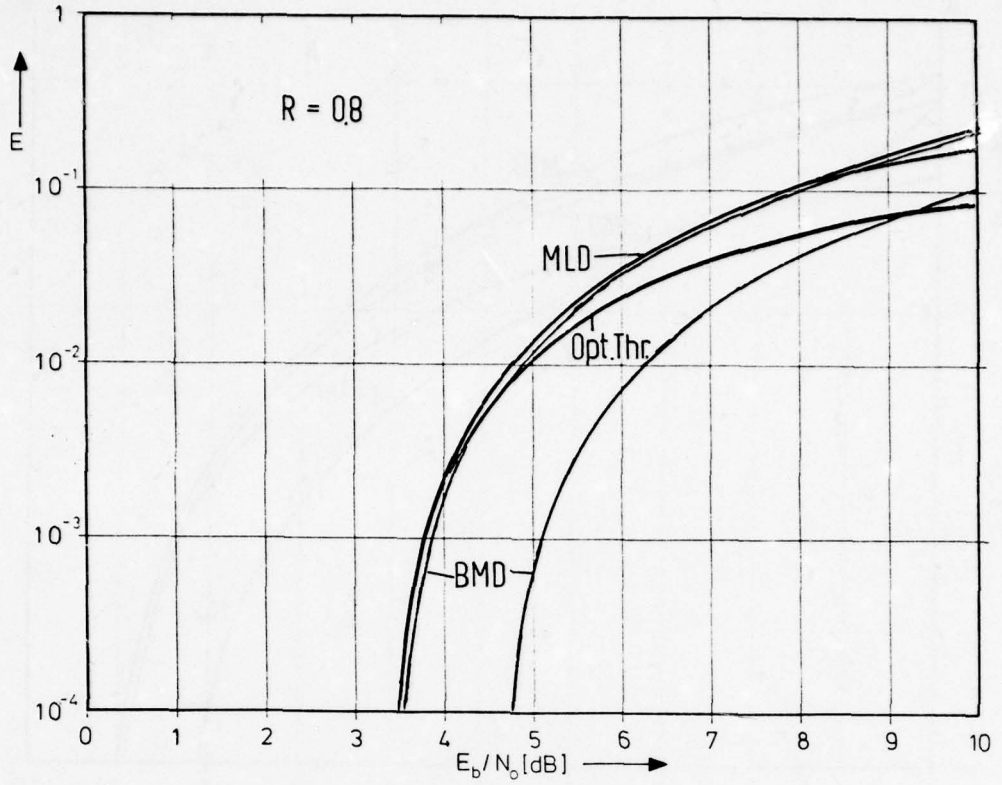


Fig. 7: Error Exponents E (for MLD, BMD, OTD) vrs. E_b/N_o for R = 0.8

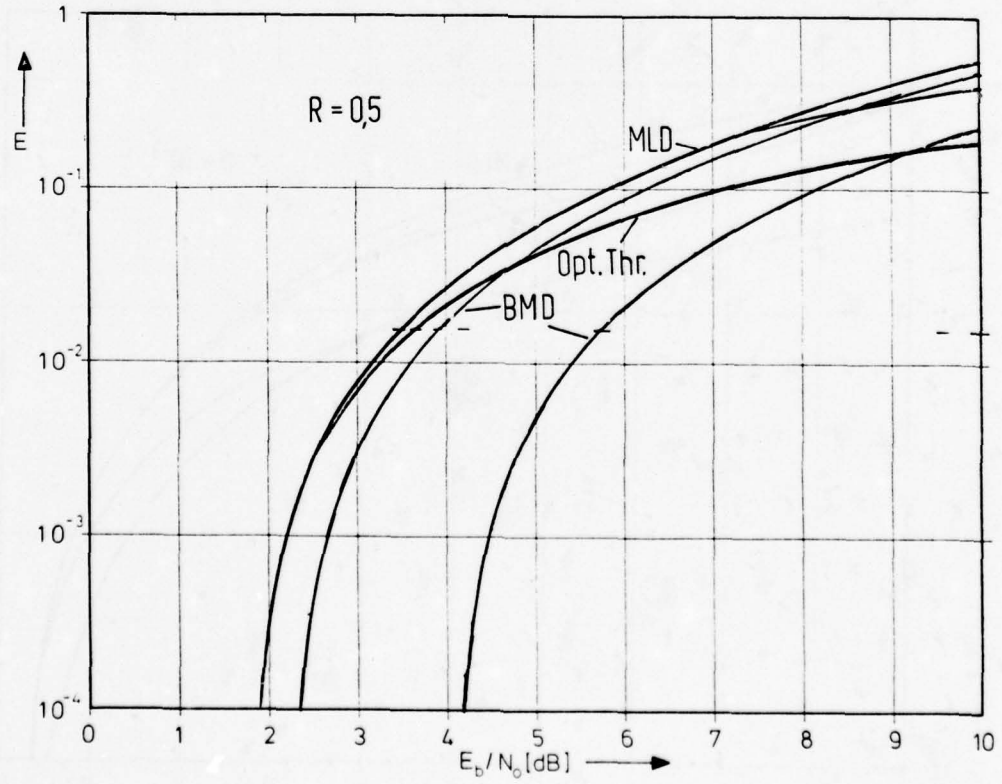


Fig. 8: Error Exponents E (for MLD, BMD, OTD) vrs. E_b/N_o for R = 0.5

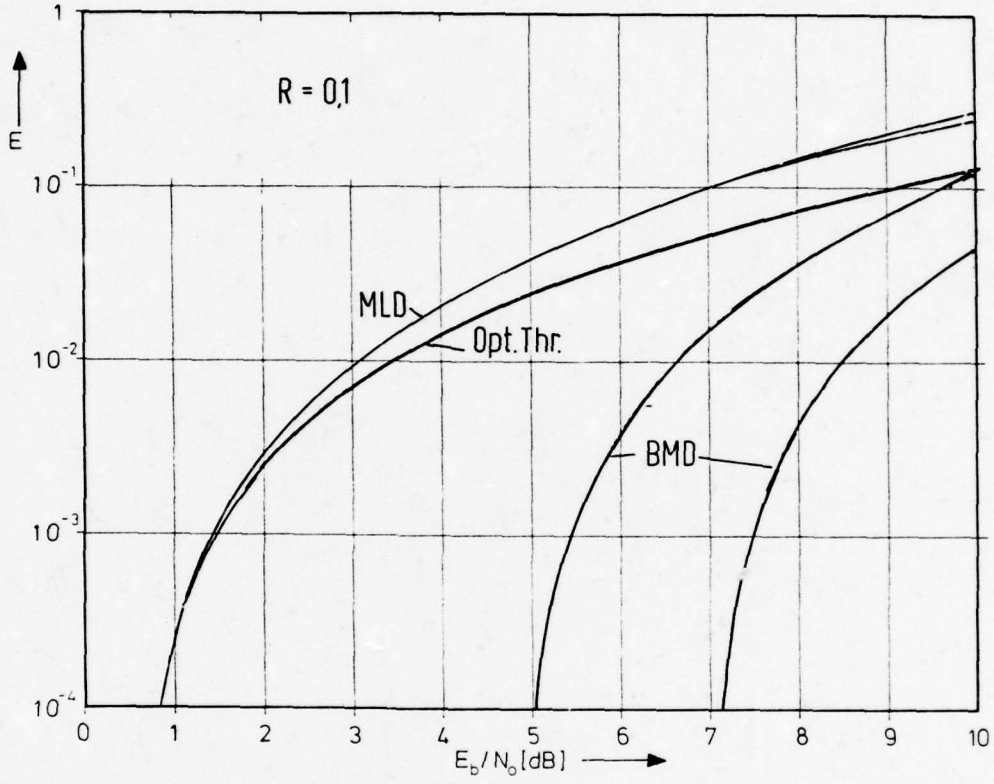


Fig. 9: Error Exponents E (for MLD, BMD, OTD) vrs. E_b/N_o for R = 0.1

DISCUSSION

P.G.Farrell, UK

How can the DOS algorithm be implemented in practice?

Author's Reply

Not yet! But due to the fine structure of e.g. "asymptotically good" blockcodes there is some hope to find a decoding algorithm, which answers the question: Is there any code word within a fixed Hamming-distance $T \geqslant D/2$ from the received binary vector. Find it and become even more famous!

AN ERROR-RATE MEASUREMENT SET-UP OPERATING AT 1 GBIT/S

Ulrich Wellens

Institut für Elektronik, Ruhr-Universität Bochum,
Postfach 10 21 48, D-4630 Bochum, F. R. Germany

SUMMARY

An error-rate measurement set-up is described which directly operates at information rates near 1 Gbit/s. The transmitter delivers NRZ/RZ pseudo-random sequences which are generated by high-speed base-coupled logic gates working in conjunction with two delay lines. In the error-rate detector the bit stream, transmitted via the communications system under test, is compared with a reference bit stream from the transmitter in an exclusive OR gate. The discovered error bits are registered in a counter chain and divided by the total number of bits transmitted during an adjustable measurement time. The accuracy achieved with this equipment is presently better than 10^{-9}. Two application examples in utilizing the measurement set-up are described. In one of these the error rate versus the input signal power of a coaxial-cable transmission path is determined. The other application is concerned with a special problem in an optical transmission system. Moreover, the detector can be used as a normal pulse counter at gigahertz clock frequencies.

1. INTRODUCTION

In high-resolution radar and broadband PCM communication systems, information rates with gigahertz clock frequencies are under discussion. To obtain the overall characteristics of such systems the error rate versus several significant parameters must be measured.

Recently, some error-rate measurement equipments were described which operate at 1.28 Gbit/s (HANKE, G., 1974, and HANKE, G., STEINER, M., 1977). These instruments employ a pseudo-random sequence (PRS) with the high bit rate formed by multiplexing four PRS's with respect to the characteristics of the m-sequences. In the error-rate detector these four sequences are at first recovered by demultiplexing circuits and then compared with four reference sequences which are synchronously restored in the detector. Only in case of a statistical error distribution it would be sufficient to inspect one of the channels. Thus, the discovery of the errors in all the channels leads to very extensive electronic circuitry.

In this contribution an error-rate measurement set-up is pointed out which operates directly at 1 Gbit/s both on the transmitting and on the receiving end. In this way the necessary expenditure is greatly reduced. Under this aspect, also the reference sequence is not derived from the detected signal in the error-rate receiver. First, this implies a generator which delivers a test and a reference word, and secondly, this requires that the transmission system under test is arranged in a loop. The former does not cause particular difficulties. The latter means a limitation but is in many cases fulfilled in the laboratories where complete communications systems or the like have to be investigated at high bit rates.

2. GENERAL INFORMATION

The basic block diagram of the complete error-rate measurement set-up is shown in Fig. 1. The transmitter delivers two identical bit patterns. One of these, the test word, is transmitted via the system under test and may show bit errors, e.g. as a result of noise introduced in the communication channel. The second bit pattern, the reference word, has to be delayed in such a manner that it is in phase with the test word at the input of the error-rate detector. The discovered error bits are counted and divided by the total number of the pulses transmitted during a fixed measurement time. The desired error rate results from this operation. In the following sections the transmitter, the detector, and two application examples are described.

2.1 The transmitter

A digital signal that is to be transferred by a communication system has mainly two requirements to meet: the changes of state of the signal should be as frequent as possible to facilitate the regeneration of the clock, and the spectral power density should be as uniform as possible. Both requirements are fulfilled by binary pseudo-random signals. Multi-level codes are out of question, since they cannot be generated presently at the high bit rate involved. Therefore, the transmitter generates a binary PRS of a maximum word length $L=2^n-1$. The exponent n indicates the section number of an appropriate shift register with feedback loops connected to the input via an exclusive OR gate, Fig. 2a. At bit rates with gigahertz clock frequencies this concept relies on very fast flipflops which at present are not yet commercially available. Therefore, a different realization was chosen (BALL, J. R. et al., 1975, and MEYER, F., 1976) in which the shift register elements are replaced by a delay line, Fig. 2b.

In order to generate pseudo-random sequences, this line is divided into two parts, corresponding to m and n with regard to the construction rules of the PRS. The transit time τ_1 of the first delay line is reduced by the delay time of the exclusive OR gate simulating a modulo-2 adder of zero delay. Thus, the maximum possible bit rate is only limited by the rise and fall times of the exclusive OR gate. In the special case of m=n-1, which allows the generation of many PRS's, the transit time τ_2 fixes the bit rate and the sum $\tau_1 + \tau_2$ determines the generated word length, comparable to a choice of the cell number n of a shift register. If high values n are desired, the first transmission line may be required to be so long that the attenuation and frequency-dependence of the cable call for a regeneration, which can be achieved by an additional gate within this transmission line. Then the transit time τ_1 must be reduced by the gate delay time. A generator working in the above mentioned manner is a delay line oscillator and will freely run if no external clock is applied. Fig. 3 shows a section of a PRS with NRZ (no-return-to-zero) pulses at 1.12 Gbit/s.

The extension of this basic unit to the complete transmitter is depicted in Fig. 4. Synchronization of the generator to an external clock is possible if the clock frequency comes close to the free-running bit rate of the generator. One way of obtaining synchronization is to inject the clock signal in one input of an additional gate in cascade with the first cable (BALL, J. R. et al., 1975). Another way, which was used here, is to inject a small amount of the clock signal directly into the common base of two transistors in the exclusive OR gate (see remarks below). Experiments have shown that the synchronization range of about 1 % does not exceed the clock tolerance allowed in digital transmission systems.

The AND gate in which the PRS is combined with the clock frequency transforms the NRZ signal stream into a RZ (return-to-zero) bit pattern. The output circuit provides the test and reference word which are derived from the PRS by two inverting circuits.

The logic configuration of the high-speed gates was implemented by base-coupled logic (BCL) (MEYER, F., 1975 and 1976). Fig. 5 shows the BCL exclusive OR/NOR gate with the values of the supply and reference voltages normally used. BCL circuits switch from ground (logical one level) to -0.6V, thus giving a typical swing of 0.6V. A change of the logical gate function can easily be obtained by omission of appropriate transistors. Therefore, BCL possesses a high flexibility. Since all gates have normal and inverted outputs, they are used in a push-pull operation of both the AND gate and the inverters in the output circuit. This improves the transfer characteristic and the noise immunity. The reference voltage is not necessary for these gates; the AND gate, however, needs an additional transistor parallel to the reference transistor Tr3. The design of a push-pull exclusive OR/NOR gate is possible but leads to some crossing of conductors, which can only be prevented by a complicated multilayer structure. The lengths of the delay lines were experimentally determined for PRS's with n=6 and 7 at different bit rates.

The circuits were fabricated on 1" x 1" glass fibre-reinforced teflon substrates using chip resistors, chip capacitors, and ceramic packaged transistors (BFR 35A, f_T=4.3 GHz; partly special multi-emitter transistors). Fig. 6 shows the measured eye diagram of a 127-bit RZ-PRS at 1.06 Gbit/s. Rise and fall times of the pulses amount to about 250 ps.

2.2 The error-rate detector

As mentioned above, the test and reference pulse streams are compared bit by bit in the error-rate detector, the simplified block diagram of which is shown in Fig. 7. For a first consideration disregard the dashed blocks. An exclusive OR gate, identical to the appropriate gate working in the m-sequence generator, serves as a digital comparator. The discovered error bits are registered in a counter chain when leaving the prescaler. In a second counter chain the clock pulses, synchronized with the reference word, are also registered. These two counting operations are controlled by the start-stop-unit and the AND gate. At the beginning of the measurement cycle the counts stored in the counter chains are cancelled, and simultaneously the AND gate is opened. The cycle ends as soon as the 10^nth clock pulse closes the AND gate. Thus, the exponent n, adjustable within a range from 3 to 12, determines the measurement time and the denominator D of the error rate. The error bits pass the AND gate during its opening time and are registered in the error counter which determines the numerator of the error rate = $N/D = N \cdot 10^{-n}$, n = 3...12.

The error digits from 10^3 to 10^{12} are directly indicated on a 7-segment LED display. The information stored in the prescaler, however, is not immediately at disposal. Therefore, as the AND gate closes, at the end of the measurement time, a 1 MHz generator is started (dashed blocks). The pulses of this generator enter both the prescaler and a count-down counter which is at first adjusted to one thousand. For example, the error number, stored in the prescaler, may be E with $0<E<1000$. The 1 MHz generator delivers F=1000-E pulses, before the first pulse then leaving the prescaler output stops the generator. Simultaneously, after F pulses, the count-down process will be finished. In this way a reproduction of the now displayable error number E is obtained by the count-down counter. During this process a gate (not shown in the figure) separates the output of the prescaler from the following counter to avoid a change of the previous counting result.

For reasons of compatibility with the developed word generator, the high-speed gates are base-coupled logic gates, too. Since the BCL exclusive OR gate needs both the normal input pulses of the test and reference channel and their negations, two basic BCL gates are connected in front of the exclusive OR, to reduce the necessary number of inputs to

only two. Moreover, these gates effect a regeneration of the input pulses. This fact may be especially important in the reference channel if the indispensable delay line causes some attenuation. The regeneration in the test channel is usually made by a repeater which is a part of the communications system under test. In the prescaler and in the clock counter, respectively, the first two dividers with 4:1 and 10:1 division factors are based on ECL (Fairchild 11C05 and 95H90), followed by TTL gates for the other circuits. The MHz pulse generator is separated from the prescaler input by a Schottky diode.

The accuracy of the error-rate detector depends on two effects. A first disturbing influence is the fact that the first ECL divider cannot be reset, but this leads only to an uncertainty in the last digit of the measured error number. The second influence is caused by the following: Both the measurement time T_M and the denominator of the error rate are fixed by the choice of the above mentioned exponent n. The discovered error bits, however, are counted during the time T_{AND}, e.g. when the AND gate is opened. Since the 10^nth pulse has to switch several gates in the start-stop-unit before the AND gate is finally closed, T_{AND} lasts a certain time longer than T_M. During this delay time additional error bits falsify the measurement results. In the worst case, the delay time of the error-rate detector amounts to 160 ns, corresponding to a measuring error of 160 at 1 Gbit/s. If the error number is large compared to this value, the measuring error is negligible. On the other hand, if the error number is small in comparison with 160, it is improbable that the error bits appear just during the delay time; i.e., at an error rate of 10^{-9} and a bit rate of 1 Gbit/s the averaged error period lasts one second. Therefore, the actual measuring error is much smaller than the worst case value. Moreover, the accuracy increases with increasing measurement time.

Since the errors are directly displayed, the influence of adjusting operations on circuits of the communications system can immediately be observed on the display. If only one input channel of the error-rate detector is used (with regard to the flexibility of BCL the exclusive OR gate then works as an inverting circuit), the instrument can be employed to count pulse events during an adjustable measurement time. Among other things this mode of operation was used to check the equipment by counting the known number of "1"-bits of a 127-bit pseudo-random sequence. This number could be measured with an accuracy of better than 10^{-9}.

2.3 Measurements

In the following section two applications of the complete error-rate measurement set-up are explained. The error rate versus the repeater input power of a simulated coaxial transmission system was measured in a first experiment. It is assumed that the coaxial cable has a frequency independent attenuation which is proportional to the cable length. This behaviour was simulated by a step attenuator. The communication source delivered binary NRZ pulses, which were generated by the transmitter (NRZ-RZ conversion omitted). Fig. 8 shows the computed error rate and the measurement result. The experiment was performed with a 127-bit PRS at 1.06 Gbit/s. The theory predicts an error rate which is dependent on the signal-to noise ratio by means of the normal error integral (HÖLZLER,E., and HOLZWARTH,H., 1975). The difference between measurement and computation arises from a not perfect (as theoretically postulated) regeneration made by the communication system. The absolute value of the input power was not optimized by the transmission system. It was the purpose to show that the developed error-rate measurement set-up permits an optimization of real systems at high bit rates if the dependence of the error rate on several parameters is investigated and analysed.

In the second example parts of an optical transmission system were tested, Fig. 9. A GaAs/GaAlAs semiconductor laser diode was modulated by the RZ pulses of the PRS generator. After transmission over a short optical fibre the light pulses were detected by an avalanche photodiode, the output pulses of which were amplified and then fed into the error-rate detector. To avoid both the delay time and the so called "pattern effect", the laser diode carries a pre-current I_o. This current has to be equal to, or somewhat above, the threshold current I_{th} of the laser diode. At the high bit rate employed, this adjustment is rather critical. Fig. 10 shows the eye diagram of the amplified photodiode signals. The experiment was performed at 0.95 Gbit/s with a 63-bit pseudo-random sequence. The error rate versus a deviation ΔI_o was measured at an optical power of about -38 dBm, Fig. 11. The critical point certainly is the dependence of the threshold current on temperature (about 0.5 mA/K for many lasers). For reasons of simplicity, this dependence was simulated by the change of I_o leading to the same result. The measurement result demonstrates that in future optical communication systems of high bit rates only a small temperature drift is admitted for the laser diode. Moreover, the long-term drift of the laser threshold current still causes some difficulties today.

3. CONCLUSION

An error-rate measurement set-up was presented working at 1 Gbit/s. The transmitter delivers NRZ/RZ pseudo-random sequences which are directly handled by the error-rate detector at the high bit rate. Since no multiplexing technique is used, the expenditure could be greatly reduced compared to previously proposed solutions. A limitation of the present measurement principle is the necessary arrangement of the system under test in a loop. The accuracy of the set-up is better than 10^{-9}. The variety of the application range was demonstrated by means of two examples derived from different research fields.

Acknowledgements:
The author would like to thank Prof. B.G. Bosch for stimulating contributions to this work and Dr. F. Meyer, Siemens AG, for providing the multi-emitter transistors.

REFERENCES

BALL, J. R., SPITTLE, A. H., and LIU, H. T., 1975, "High-speed m sequence generation: a further note", Electron. Lett., 11, pp. 107 - 108

HANKE, G., 1974, "Entwurf eines 1.28 Gbit/s-PCM-Versuchssystems mit Bitfehlerratenmeß-einrichtung in integrierter Schaltungstechnik", Forschungsinstitut beim FTZ der Deutschen Bundespost, 442 TBr 52;also: International Conf. on Communications, San Francisco, June 16 - 18, 1975, Session 24

HANKE, G., and STEINER, M., 1977, "Schnelle Ereigniszähler in integrierter Schaltungs-technik für Bitfehlerratenmeßgeräte bis 1280 Mbit/s", Forschungsinstitut beim FTZ der Deutschen Bundespost, 51 TBr 9

HÖLZLER, E., and HOLZWARTH, H., 1975, "Pulstechnik Bd. I", Springer Verlag, Berlin - Heidelberg - New York, p. 384

MEYER, F., 1976, "Gigabit/s m-sequence generation", Electron. Lett., 12, p. 353

MEYER, F., 1975, "Subnanosecond base-coupled logic circuits", Conf. Publ. of the 1st European Solid State Circuits Conf., Canterbury, Sept. 2 - 5, pp. 32 - 33

MEYER, F., 1976, " Base-coupled logic circuits for high-speed digital systems", Nachrichtentechn. Z., 29, pp. 828 - 830

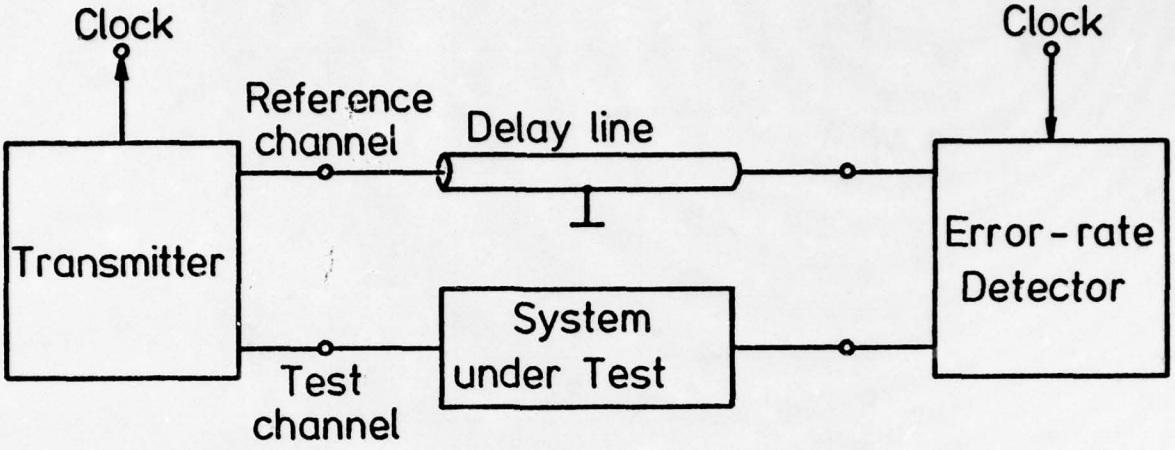


Fig. 1 Block diagram of the error-rate measurement set-up

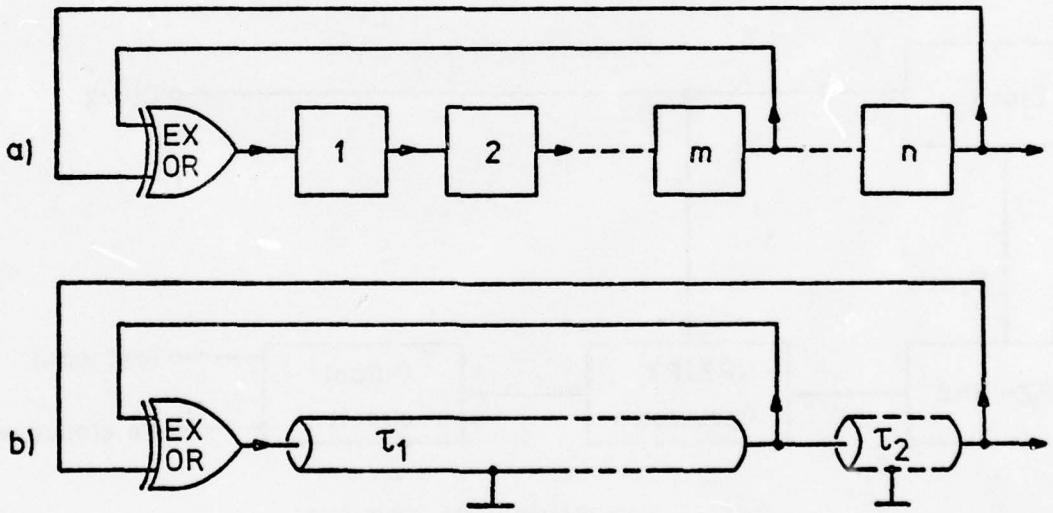


Fig. 2 m-sequence generator circuit: a) n-stage shift register,
b) realisation with delay lines

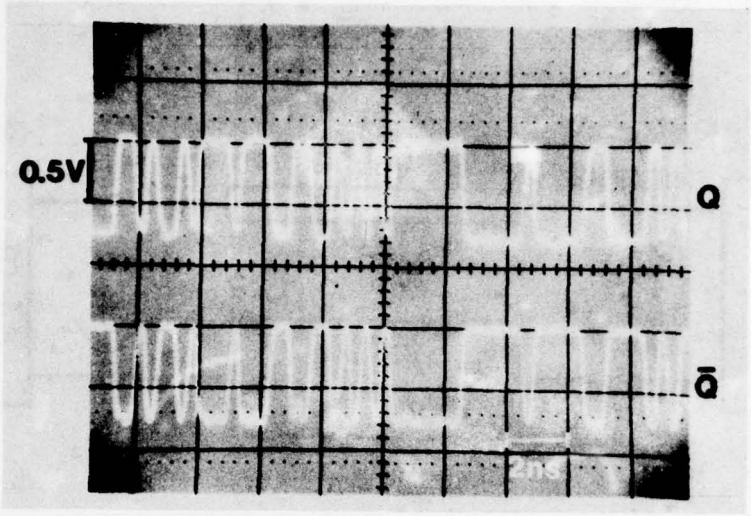


Fig. 3 Part of a 127-bit NRZ pseudo-random sequence at 1.12 Gbit/s

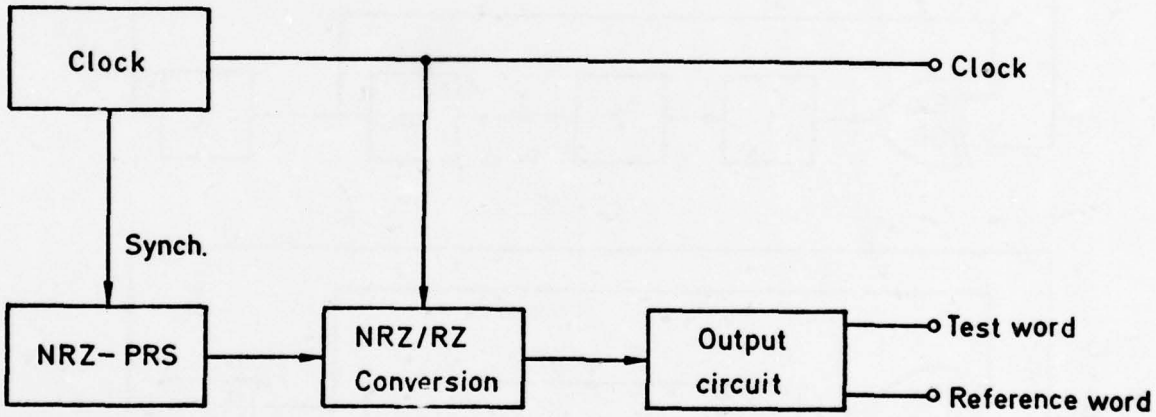
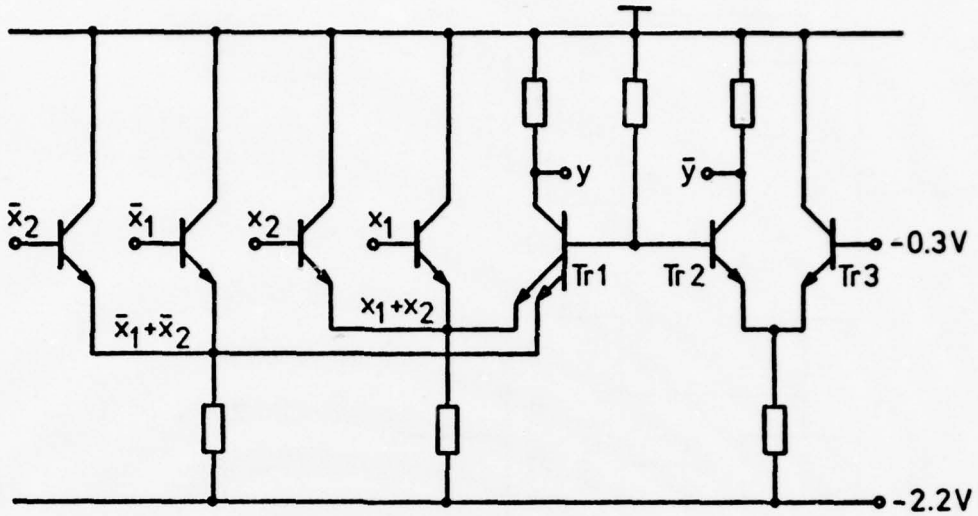


Fig. 4 Block diagram of the transmitter



$$y = (x_1+x_2)\cdot(\bar{x}_1+\bar{x}_2)$$

Fig. 5 Base-coupled logic exclusive OR/NOR gate

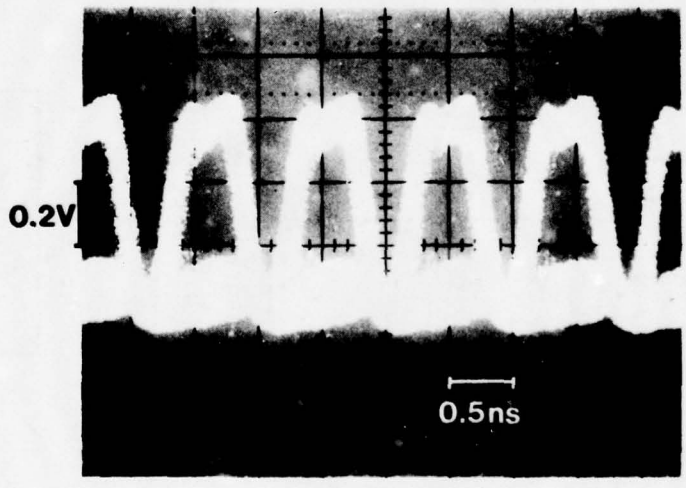


Fig. 6 Eye diagram of a 127-bit RZ pseudo-random sequence at 1.06 Gbit/s

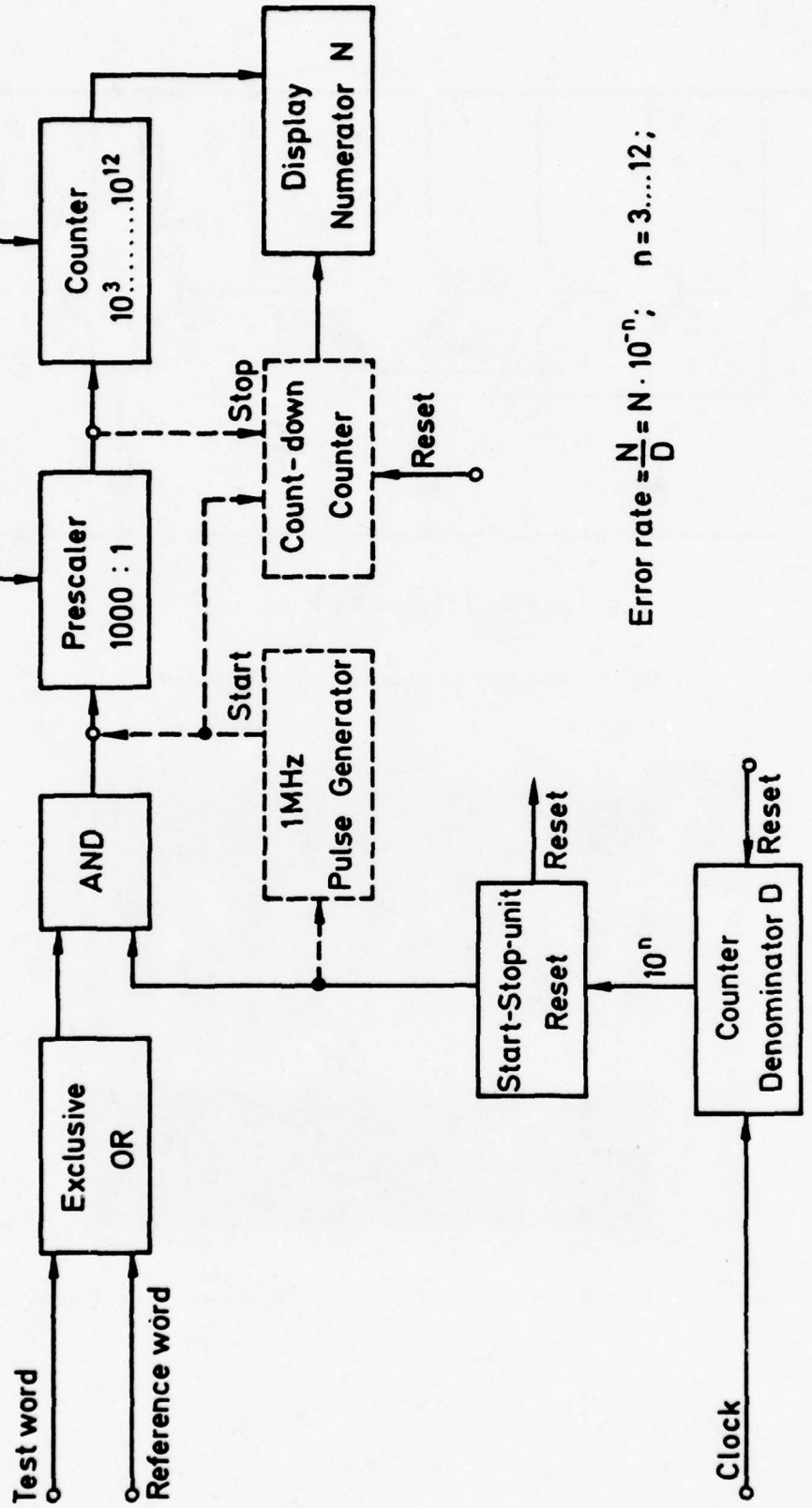


Fig. 7 Block diagram of the error-rate detector

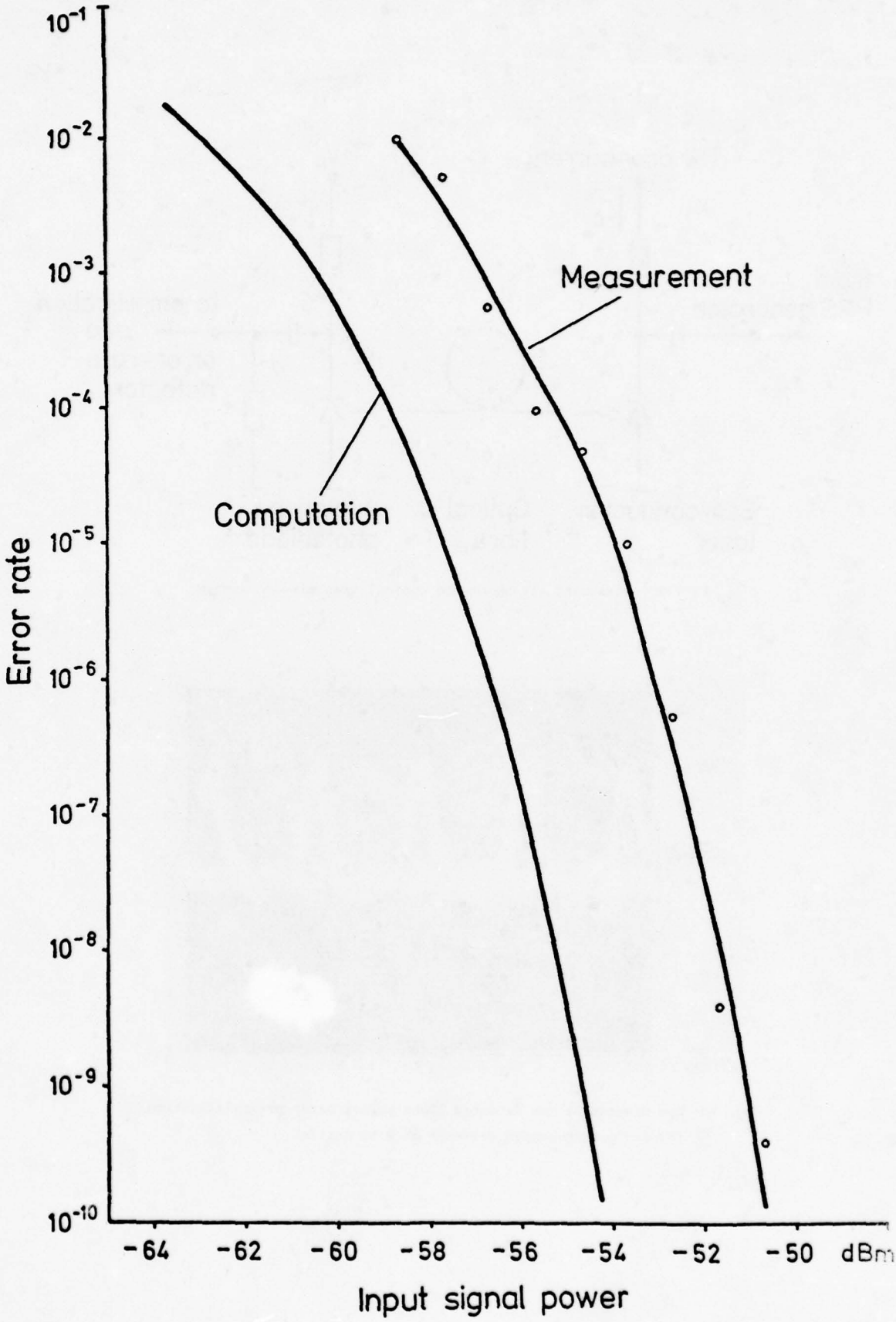


Fig. 8 Measured and calculated error rate versus repeater input signal power at 1.06 Gbit/s

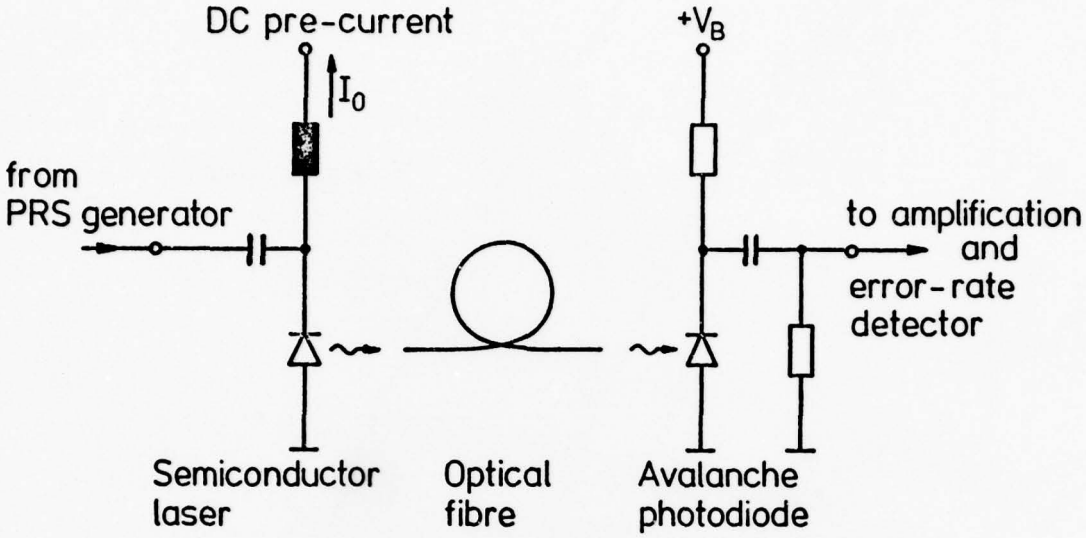


Fig. 9 Measuring set-up of the optical transmission system

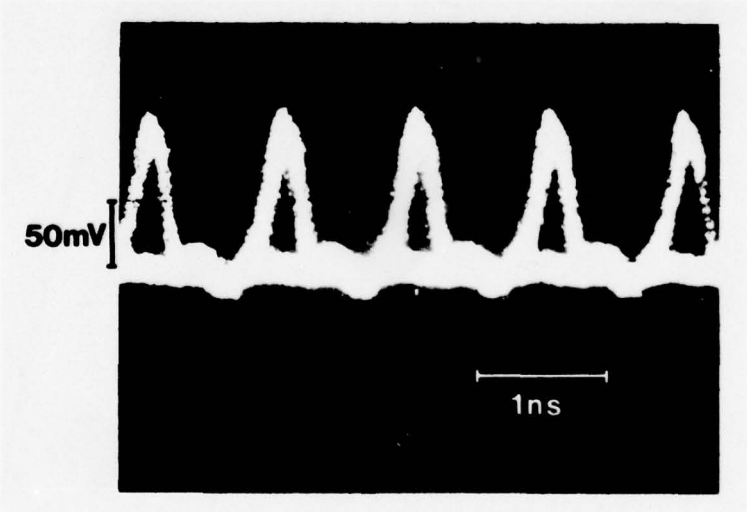


Fig. 10 **Eye diagram** of the detected light pulses after preamplification
(63-bit pseudo-random sequence at 0.95 Gbit/s)

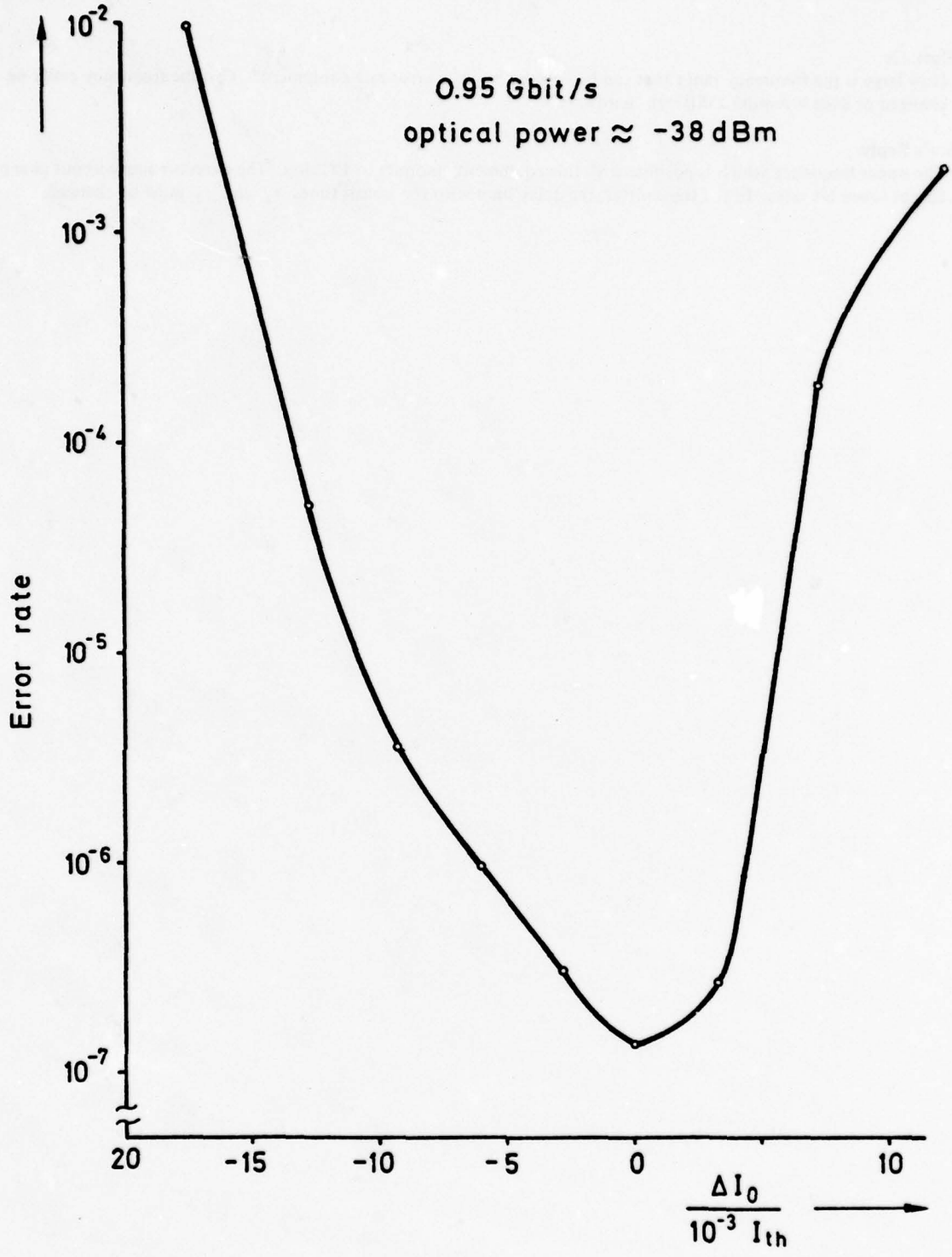


Fig. 11 Measured error rate versus normalized change of the laser pre-current

DISCUSSION

H.J.Matt, Ge

How large is the frequency range that can be covered by your error rate equipment? Can the frequency easily be changed or does it require a different hardware?

Author's Reply

The upper frequency which is possible with this equipment amounts to 1 Gbit/s. The detector acts without changes also at lower bit rates. In the transmitter, the delay lines with the transit times τ_1 and τ_2 must be changed.

INTRODUCTORY NOTES ON PROPAGATION EFFECTS AND RELATED ASPECTS

by
Dr H.J.Albrecht
FGAN
Königstrasse 2
D-5307 Wachtberg-Werthhoven, Germany

1. General Comment

In telecommunications generally, and particularly with digital communications in avionics as well as similarly specialized fields of application, criteria of electromagnetic wave propagation may represent most significant limitations to system design and system operation. Normal conditions of a propagation medium may result in an appropriately simplified consideration, or in reasonably valid assumptions when balancing theoretical system design against medium-dependend limitations.
To some extent the medium may be considered a "black-box" with characteristic behaviour. Such circumstances might even strengthen the trend towards ignoring the physics of propagation processes. As appealing as this possibility may be in technical work, a more serious approach in optimizing system design leads to a word of warning: Modern requirements of system reliability and capacity, and particularly those in avionics communications, can hardly be separated from the most stringent conditions with regard to wave propagation. Related aspects have to be taken into account up to a degree governed by the state of the art in relevant research efforts.

Recent research and development have been directed towards digital communication with and among airborne and spaceborne carriers. Whereas other sections of these Conference Proceedings refer to work in non-propagation fields, this session deals with propagation effects and related aspects and has been organized by the Electromagnetic Wave Propagation Panel of AGARD.

It is the objective of these Introductory Notes to provide fundamental information on propagation media, as a basis for subsequent papers which review the more detailed problem areas and illustrate typical applications. These notes also include comments on the trend in relevant research and development.

2. Aspects of Propagation Media

A propagation medium may be defined as any solid, gaseous, or liquid material in which an electromagnetic wave propagates. Depending upon the state of the art, knowledge on medium characteristics is steadily increasing with some configurations of waves of certain frequency ranges in certain media, or may have reached some saturation level prior to new advances in measurement technology and methods of analysis. Nevertheless, the media of primary interest are generally those in the atmosphere of our planet, mainly the ionosphere and troposphere and adjacent regions.

Fig. 1 illustrates the atmospheric environment and its effect on all frequencies within the spectrum 10 kHz to 100 GHz. Altitude above ground and frequency are shown by logarithmic scales. On the left-hand side, generally adopted designations for regions of the atmosphere have been mentioned. Within the troposphere, a distinction has been made for typical altitude ranges of clouds in lower, medium and upper heights. The difference in average altitude of the tropopause between equator and polar regions is also shown. The change in electron density between D, E, F_1 and F_2 layer regions is responsible for different reflection characteristics. The logarithmic altitude scale permits us to find, within the illustration, the altitude for geostaionary satellites as well as that of the moon. Looking at the atmospheric behaviour within the entire frequency spectrum we have first of all the ionospheric reflection up to frequencies of about 30 MHz with some partial transparency in the lower frequency range, depending upon the relationship of operating frequency to gyro-frequency and upon the direction of propagation with respect to the magnetic field lines. The range of ionospheric scatter propagation and meteor backscatter is shown up to 100 MHz. The so-called "radio-window" for communications with spacecraft and satellites is shown to commence at about 100 MHz, its upper limit being governed by high attenuation of frequencies of the order of 10 GHz and higher. It should be emphasized that all popular tropospheric propagation links also use frequencies within this radio window. In the upper portion with respect to frequency, the effects of precipitation may represent a significant source of difficulties while the absorption line of O_2 around 60 GHz renders impossible ordinary links through such a medium.

Another parameter of interest is the noise characteristics within the entire frequency range. Again, a "window" of low noise temperature should be taken into account when selecting frequencies for certain links. Atmospheric noise, for instance, increases with decreasing frequency, while cosmic noise decreases.

The Earth's surface is an important propagation medium which may become effective in several ways. These refer to action upon antenna characteristics in the so-called near field of the antenna, to attenuation for ground-wave propagation along the surface, to the effects of vegetation and of ground parameters on the reflection properties with respect to wave polarization, intensity, and phase, and, with sub-surface propagation, to the attenuation experienced by a wave travelling through a layer. The two characteristic parameters, ground conductivity and dielectric constant, are functions of humidity and temperature, and frequency; they may be considered variable with appropriately detrimental effects upon wave propagation. The geographical distribution of ground characteristics is of importance when considering world-wide communication or navigation applications and the behaviour of the ground in reflections. Connected with the Earth surface as a propagation medium is water in its various configurations, with salinity and other parameters being responsible for changes in the electromagnetic characteristics.

3. R & D Trend in Propagation Media

Objectives of research and development in electromagnetic wave propagation comprise an optimization of links for communication and other purposes by identifying, predicting, and, where possible, mastering difficulties as well as related limitations. Work has recently commenced on a perhaps very powerful, future tool: the artificial modification of propagation media to achieve maximum efficiency and optimization. Up to a certain degree such remedies concern all portions of the electromagnetic wave spectrum and all fields of applications. Activities of the Electromagnetic Wave Propagation Panel during the last few years, and particularly in 1976, have led to indicative results.

In Ionospheric Radio Wave Propagation, the impact of satellite technology about ten years ago caused a change with regard to the application of ionospheric research results to problems connected with this new type of paths requiring an optimum atmospheric transparency. Already in the late sixties activities of the then Electromagnetic Wave Propagation Committee concerned ionospheric irregularities as one significant source of possible limitations. Research work continues with the objective of optimizing communication and other links using this path configuration.

Additional activities in the field of ionospheric propagation refer to the extension of the useful spectrum by means of artificially modifying the medium such that the maximum usable frequency for high-frequency long-distance propagation (so-called "MUF) is increased and the lowest usable frequency ("LUF") is lowered. A promising method of extending the maximum usable frequency is represented by "ionospheric heating", or the use of extremely high radio wave energy to heat a certain portion in the ionosphere which in turn assists in forming a satisfactory propagation mechanism for frequencies above the natural maximum usable one. The reduction of the lowest usable frequency may perhaps be achieved by the release of chemical substances in a certain ionospheric volume.

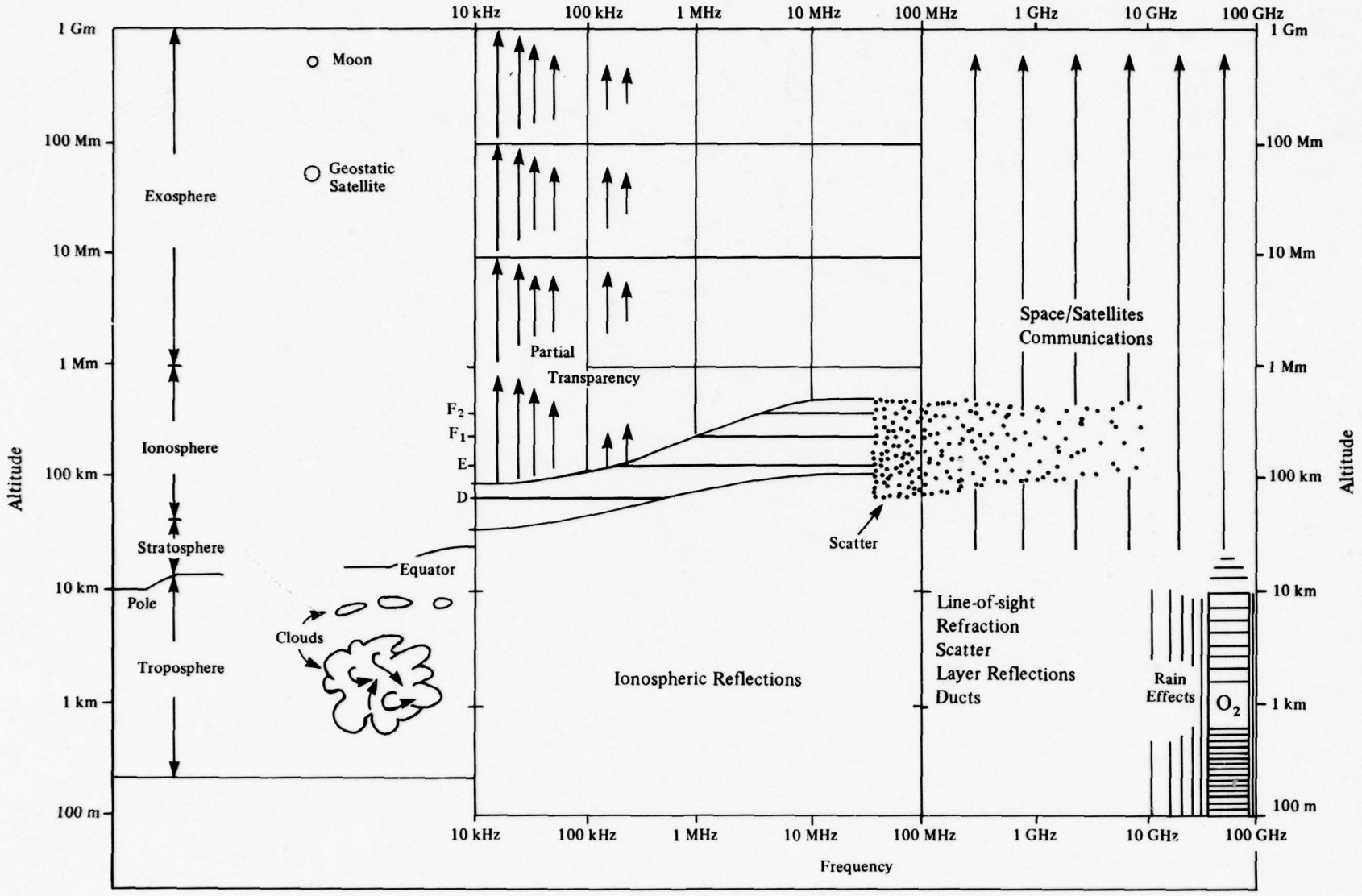
In Tropospheric Radio Wave Propagation, a steadily progressing technological development has enabled the use of increasingly higher frequencies; this process can be expected to continue with the result of reliable equipment becoming available in ranges of millimetre, sub-millimetre and optical waves. As far as research and development in propagation media is concerned, this progress leads to an increasing importance of work directed at identifying, analyzing, and predicting areas, occurrence, and intensity of natural limitations typical of those frequencies. An enormous amount of data has been collected on relevant links operating on frequencies below 10 GHz, such that only special questions remain to be solved. An example is the variable effect of inversion layers in the vicinity of line-of-sight paths; such work is just commencing and will be of particular importance in connection with the establishment of topographical data banks for automatic computerized link design. As another example measurements of attenuation and phase effects of rain and other precipitation are now being undertaken in many countries of the world; an appropriate evaluation may be expected to result in reasonable statistical data for link design. Similar to its application in ionospheric propagation, the artificial modification of tropospheric propagation media should, some time in the future, yield a higher reliability in tropospheric propagation. In this case, a predominant connection exists with experiments in weather modification.

For completeness' sake, mention is to be made of a dissemination of reflecting material within atmospheric propagation media. Release and use of such chaff represent a classical anthropogeneous method of influencing propagation media artificially. Current research may assist in optimizing its usefulness for communication purposes.

4. Propagation Criteria with Digital Communications in Avionics

The field of applications represented by the topic of this conference requires a detailed discussion of relevant propagation aspects. Following this Introduction, the particular problem areas are reviewed and criteria of modelling are dealt with. A number of typical applications are described.

The papers in this session illustrate the necessity of employing a most advanced knowledge of electromagnetic wave propagation as a vital basis of modern system design in avionics communications. It is one of the objectives of this session to assist in this special field and thus to contribute to an improvement of digital communication systems in avionics.



(H.J.Albrecht 1968/76)

Fig.1 EM-wave propagation 10 kHz – 100 GHz

PROPAGATION EFFECTS ON DIGITAL COMMUNICATION IN AVIONICS

(Review Paper)

by Ernst Lampert
Siemens-Aktiengesellschaft
Hofmannstraße 51
8000 München 70, FRG

SUMMARY

As communication is possible using different modes of EM-wave propagation their relevancy in an avionics scenario has to be looked at.

The relevant propagation modes are then discussed in terms of frequency from the HF band to EHF region. Satellite communication has been dealt with separately.

1. INTRODUCTION

Communication systems have the purpose to transfer a certain information flux over a considerable distance. If at least one of the terminals is mobile, this can only be done by means of EM-waves, radiated from one point and observed at the other. The laws to which EM-wave propagation obey cannot be altered by the system designer, therefore he has to know the relevant characteristics of the waves in real world environment. (There exist only rare cases where the box consisting of antennas and medium behaves like a variable attenuator.) As EM-wave propagation covers a large field of scientific activities, we have to establish selection criteria according to the relevancy to communications in avionics with emphasis on digital signals.

2. RELEVANCY

The task is to transmit and correctly receive data with a speed of about 50 bps to some Mbps. The terminals can be assumed to be in motion with sometimes very high speed (say 500 m/s); they may be situated on the ground, in the air or in space. In all these cases the link availability has to have a sufficiently high value, whereby the number of deep fades or sudden phase changes should be rare, as especially in encrypted or multiplexed digital links the bit integrity must be guaranteed for periods exceeding resynchronisation time by several orders of magnitude. It is therefore primarily irrelevant for the user whether the outage occurs because of the lack of signal power, or pulse distortion because of multipath, or loss of synchronization in coherent modulation systems because of sudden phase changes.

As links are to be provided mainly among mobile units, only those propagation modes should be used, which do not require bulky equipment especially onboard an aircraft. A scatter link is therefore not a means of communications with an aircraft but may only be suitable for ground support [1] . Considering furthermore that with mobile units radiation should be approximately omni-directional, we can concentrate on the following propagation modes:

(1. ground wave propagation)

2. skywave with ionospheric refraction

3. terrestrial line-of-sight and diffraction

4. satellite communication.

The modes 1 and 2 are restricted to the lower frequency region, 3 and 4 to the higher frequency bands. - Of the complete presently available radio spectrum, VLF, LF and MF can be excluded as these bands are either used for other purposes but communication or they are occupied by broadcast services. The relevant frequency range starts above 1.5 MHz, with the extended HF-band. - It can also be agreed upon that between mobile units SHF, EHF and optical frequencies can only be used for digital communication under special conditions. However, as the optimally suited bands HF, VHF and UHF are almost occupied an evasion into higher frequency bands seems to be necessary.

As EM-wave propagation is highly sensitive on the boundaries of the medium, so the following scenarios relevant in avionics will be emphasized:

Air - Air (short distance
(long distance

Air - Ground

Satellite communication

(Terrestrial support mobile/fixed)

Table I shows the relevant problems for these scenarios.

3. NOISE

As there exists not only the signal source in the propagation medium, further unwanted signals are present at the receiver. When disregarding jammers, noise signals are present because of atmospherics, man made noise and extraterrestrial sources. A spectral distribution is shown in Fig. 1 [2] . - As man made noise is propagated predominantly by ground wave, the horizontal component of the overall external noise is usually considerably smaller than the vertical component.

Assuming that the equipment has a noise figure below 10 dB, it is obvious that only in the HF-band this external noise is important and dominates in the region below 10 MHz as long as omnidirectional antennas are used.

In the HF band the part of the external noise caused by atmospherics varies according to daytime, season and geographic location [3] .

On the other hand in the region above UHF very low noise receivers are only efficient, if no external noise source (e.g. the earth) is illuminated by the receiver antenna. So only in satellite links with sufficiently high path elevation receiver noise temperatures below 300 K are possible.

4. PROPAGATION EFFECTS IN THE FREQUENCY BANDS OF INTEREST

As it has been shown already, different scenarios require the use of different frequency bands. In a survey on propagation effects it seems to be an efficient way of classification, to continue the discussion in terms of the frequency bands.

4.1 HF-band (1.5), 3 to 30 MHz; (200 m), 100 m to 10 m

Wave propagation in this band is characterized by two particular modes. The ground wave is still of sufficient strength to overcome distances up to 50 km when using frequencies below 5 MHz. No multipath is present, if no interference with other modes occurs. As the ground wave range is limited because of its 40 lg (d/λ) dB dependence of the loss (d = distance), the skywave is the dominant mode for overcoming long distances, as long as this wave has the appropriate frequency to be reflected by the ionosphere.
unique means of communications all over the world without an artificial aid.

As the electron density of the refracting ionospheric layers show diurnal, seasonal and sunspot cycle variations, the Maximum Usable Frequency (MUF) for reflection does so, too. For short distance communication the MUF varies between about 3 MHz at night and 10 MHz at noon. This however does not mean that at daytime a broader frequency band can be used for communication, as then the D-layer will absorb a considerable amount of signal power (up to 25 dB at a free-space pathloss of about 100 dB) 4 . That means that a relatively narrow available band of some MHz is only shifted in its center frequency. As MUF and the lowest usable frequency (LUF) because of D-layer absorption, are proportional to the secants of the angle of incidence longer distances require a frequency band with a higher center frequency than smaller distances. This problem can be solved by using the correct frequency range and an appropriate transmitter power. Because of the different layers in the ionosphere multipath will, however, always be a problem which results in selective fading as layer heights and electron densities vary continuously 5 . The propagation medium may involve:

a) "multiple hop propagation" even with short links, the reflection from the earth is quite often not to be neglected

b) *multiple layer propagation*

c) low and high angle ray in obligue propagation close to junction frequency 4

d) both ordinary and extraordinary ray
(the ray splits because of magneto-ionic effects)

Bayley [6] has calculated the principal time delay bounds for ionospheric propagation paths in terms of the ratio of the radio frequency and the momentary F-MUF. The resulting diagram Fig. 2 shows that only when operating very close to the actual MUF time delay values less than 0.5 ms will occur. In general up to 4 ms have to be taken into account. So when looking at the time depending transmission characteristic of the path, *fading occurs with attenuation peaks frequently up to 40 dB which are spaced quite regularly in frequency.* Because of the slow height variations of the layers the attenuation peaks move in frequency through the channel with a speed up to 200 Hz/s. In most cases only two propagation paths are of importance, showing a differential time delay of about 1.5 ms in the average (this has been found in measurement over distances from 100 km to 1500 km).

Using the ionospheric reflection mode will therefore imply that the data rate usually has to be restricted to values below 200 baud, as in most cases tracking the MUF with the radiofrequency of the equipment according to the actual ionogram is only possible in rare occasions because of the lack of personally assigned frequencies [7] . Higher data rates (up to $\sim$ 2.4 kbit/s) can be obtained using multicarrier techniques with low speed transmission on each carrier [8]. In most cases however the error probability of uncoded links will not be below 10 % for any time interval [9].

Furthermore it is difficult to use the advantages of long distance communication in systems designed to operate simultaneously over short distances as a channel frequency can either be assigned for a long distance link or a short one, but hardly for both.

4.2 Very-High-Frequency Band (30 MHz to 300 MHz, λ= 10 m to 1 m)

Higher data rates may be transmitted in the VHF region. Propagation takes place in the line-of-sight mode and is therefore restricted to relatively short ranges. So the propagation characteristics depend on the troposphere as well as the boundaries of the medium, in this case earth, source and sink.

Although the troposphere may be considered to be completely transparent for EM-waves, as long as the wavelength is above 3 cm, ray traces are different from free space, they are not straight. The refractivity n of the earth differs slightly from unity and alters with height. On the average we get [10]

$$n = 1 + 289 \cdot 10^{-6} \cdot \exp(-0.136h/km) = 1 + N(h) \cdot 10^{-6}.$$

Because of the negative value of the n-gradient the ray trace is bent towards the earth. As in link planning it is convenient to work with straight ray traces, a higher value of the earth radius is assumed (that means a smaller curvature):

$$r_{e\,eff} = k \cdot r_e \quad \text{with } k = \frac{1}{1 + r_e \cdot \frac{dn}{dh}} \quad ; \quad r_e = 6370 \text{ km}.$$

As the horizon width depends on the effective earth radius, we get

$$d = \sqrt{2 \cdot k \cdot r_e \cdot h} \qquad h = \text{antenna height}.$$

Because of varying meteorological conditions k is not fixed but is a random variable. The distribution of k is shown in Fig. 3 11 . Table II shows how the actual value of k affects the ray trace.

Besides this rather regular effect the refractivity index may show bilinear behaviour as shown in Fig. 4a. Here we use the modified refractivity index, the modulus M [12] . A jump caused by an inversion layer will lead to total reflection as long as the grazing angle ψ is below a certain limit $\psi_g \leq \sqrt{2 \cdot \Delta M}$.
The M-profile shown in Fig. 4b causes ducting of the wave, i. e. the wave is trapped either between the earth and the break or between the two breaks. This propagation mode obeys to waveguide laws and will only take place when the wavelength is below a certain limit $\lambda_g = 2.5 \cdot \Delta h \cdot \sqrt{\Delta M}$. For a difference of $\Delta M = 4 \cdot 10^{-6}$ and a height $\Delta h = 30$ m this will lead to a minimum frequency of 2 GHz. Ducts are therefore only probable for the higher end of the UHF band and in SHF. These quasi deterministic effects on wave propagation lead to shielding and strong multipath effects, so they cannot be neglected. Turbulence always causes second order signal variations compared to the direct ray and will therefore not be dealt with ($\lambda \leq 3$ cm).

In the lower VHF-region ($f < 50$ MHz) there may occur sporadic E-layer reflections. Compared to line-of-sight-signals their differential delay is of the order of 0.5 ms and its magnitude becomes comparable to this signal if distances of about 200 km are involved.

We will now consider the propagation conditions on a specific link as shown in Fig. 5. The medium is limited by source, sind and ground. As it is easily visualized, wave propagation can be spearated into three components, the direct wave, the ground reflected (and eventually the inversion layer-reflected) component and the ground wave. As the power of the ground wave decays with $(\lambda/d)^4$ and also with the height above ground of transmitter and receivers it may be neglected in the "avionic" scenario [13] . The multipath caused by ground reflection however cannot be ignored when source or sink are moving. Deep fades occur even when the reflection coefficient is still far below unity, Fig. 6. Because of the differential time delay selective fading resp. intersymbol interference is produced, so the maximum symbolrate on the path is limited as shown in Fig. 7. When looking at the dependence of the reflection coefficient on grazing angle and ground properties, it is obvious that this is the typical situation according to the table of relevancy Tab. I and is applicable to airground (or low flying aircraft), and between low flying aircraft. The calculation of the effects may be found in literature [14] .

The following major relation can be stated between isotropic radiators:

$$\frac{P_R}{P_T} = 2\left(\frac{\lambda}{4\pi d}\right)^2 \sin^2\left(2\pi \cdot \frac{h_T \cdot h_R}{\lambda d}\right) = 2\left(\frac{\lambda}{4\pi d}\right)^2 \cdot \sin^2\left(2\pi \frac{h_R}{\lambda} \cdot \sin\psi\right)$$

h_T, h_R = Transmitter, receiver height

d = Distance

ψ = grazing angle

Fading will occur within a distance $d < \frac{4 h_T h_R}{\lambda}$ or for grazing angles above $\psi > \arcsin(\lambda/4h_R)$. Beyond these limits the received power is not dependent on frequency and varies with d^4. As this usually occurs in the vicinity of the horizon, here the actual value of the k-factor is very important as shown in Fig. 8a. Ince and Williams have proofen this fact experimentally [15] .

The scenario introduced is still rather ideal. Therefore a rough surface has to be considered. Because of the rough surface, the average specular reflection coefficient is reduced, however fluctuates heavily [16] . The resulting field strength of the receiver now does show the completely regular variations but has a Rice distribution [17] . The fading pattern therefore will become lightly irregular but the order of magnitude of the fading depth will still remain. The same is true for the effective transmission bandwidth.

The effect of the rough surface in the avionic scenario has been clarified rigorously by Hortenbach [17], showing that the conditions for air - ground communication are essentially not altered. The adverse effects of multipath reduce the higher the source or sink height are located above the rough ground for a given grazing angle (specular point scattering theory [18]).

The situation is different for terrestrial ground support. (The fixed services are generally well planned with guaranteed unobstructed first Fresnel zone, see f. i. [19]). Here usually diffraction and terrain scattering are responsible for signal transmission, as in most cases the direct line-of-sight is obstructed. (Transmitter and receiver only some wavelength above ground.)
Multipath will here limit the maximum transmission band to

$$B = \frac{5 \text{ kHz}}{d/100 \text{ km}}$$

when d is the diameter of the coverage zone of the system.

4.3 Ultrahigh-Frequency-Band (300 MHz to 3 GHz, λ = 1 m to 0.1 m)

In UHF no ionospheric reflection occurs. Considering the terrestrial scenario no principal differences occur compared to VHF. However as wavelength is smaller, fading-frequency increases, when flying through the interference pattern and also the width of the doppler-spectrum of the signal increases linearly.

Obstructions in the direct line-of-sight path will result in higher values of additional attenuation, and also in the region just beyond the horizon the field strength will decay more rapidly see Fig. 8b.

As the geometric dimensions of the environment of source and sink now become comparable to be wavelength, particularly above about 1,5 GHz bodies resonate (roof gutters, aircraft structure etc.). This secondary radiation in the vicinity of the antennas deteriorates its pattern and omnidirectional radiation is hardly possible any longer (baseband diversity combining).

As this sort of antenna-pattern-lobing is strongly dependent on frequency, it is possible to improve the propagation conditions when using spread modulation [20] (frequency hopping) and simultaneously forward error correction.

In mobile terrestrial systems, the terrain variability of the received field strength, because the scatter cross-section of geometrically fixed target, increases with f^2 (as long as it is small compared to the reflection-fresnel zone at that distance), furthermore we are faced with an increase in diffraction attenuation, see Fig. 9.

4.4 Super-High-Frequencies (3 to 30 GHz, λ = 10 cm to 1 cm)

It was already stated in 4.3 that the upper part of the UHF-band is difficult to be used for mobile services. Therefore the SHF band is mostly used for fixed services or mobile services transmitting analogue information not that susceptible to fading.

The use of digital signal transmission in conjunction with spread spectrum modulation bowever could improve the situation, because the multipath degradations within acceptable limits.

Above about 10 GHz other propagation effects than the one mentioned in 4.2 become noticable. Considerable attenuation is caused by rain, Fig. 10b [21] .

4.5 Extremely-High-Frequencies (30 GHz to 300 Ghz, λ = 1 cm to 0.1 cm)

It is not only attenuation because of molecular resonance, Fig. 10a and rain which becomes appreciable now, but also path attnuation. When changing in frequency from VHF to EHF this results in about 40 dB additional attenuation, as the link margin has also to be increased, at least about 50 dB have to be gained by using directive antennas. This effect is further increased when considering that the available transmitter power decreases, too, because of technological reasons, antennas with 40 dB seem to be a general necessity. This only results in antennas of about 40 cm diameter, but the beamwidth will be only 1.5°. In radio relay links either tracking systems or very stable masts are required and are range limiting factors. Atmospheric scintillation will also be appreciable causing phase changes of some 10° [22] .

4.6 Satellite links

Satellite communication is principally possible at frequencies above the highest ionospheric reflection frequency, i. e. above 50 MHz. When asking for optimum frequency ranges, it should be borne in mind that as the satellite is to illuminate a particular area on the earth, so its antenna gain has a fixed upper bound which is frequency dependent. The same is true for terminals on the earth which when mobile can only use low gain antennas, with say $\approx$ 3 dB gain. As the sensitivity of a ground terminal is limited, this results in the rule to use, an as low as possible frequency because of the increasing free space attenuation (20 lg(f) dB).

Compared to radio relay links data transmission may occur with much higher speed when using high gain antennas, as usually the elevation of the transmission path is above 10°, which prevents the ground reflected wave from causing severe limitations of correlation bandwidth.

In the VHF band, however, quite a lot of regular ionospheric effects have to be taken into account, limiting the performance. In the UHF band they have almost died out, and one is only faced with ionospheric scintillations causing amplitude and phase fluctuations. Although, from wave propagation point ov view, frequencies between 500 and 1000 MHz seem to be optimum [23] , SHF has to be used when services using bandwidths up to 100 MHz are required. This however then requires sophisticated terminals which can only be installed on larger aircraft.

Because of the lack of unoccupied bands, frequencies above 10 GHz will be applied by future systems. The propagation medium is the critical factor as in terrestrial communications particularly precipation causes attenuation, noise and depolarization. EHF has been already used in intersatellite communication (LES 8,9) [24] where such restrictions posed f. i. by the 60 GHz oxygen band do not exist and could even preserve those links from terrestrial interference.

5. CONCLUSION

It has been shown that multipath in terrestrial environment limits the data rate of mobile avionic systems. At the lower frequencies in HF limitations exist, because of the varying structure of the ionosphere and its inherent multiple reflection characteristics causing differential delays of the order of ms and fading depths

of up to 40 dB. These restrictions are not as stringent in the VHF, UHF region. It has however been shown that as the scenario is independent on frequency, differential delay values of some µs will always be present and fading will also be of considerable depth. The use of the higher frequency bands SHF, EHF is limited not only by additional attenuation because of rain water vapour and oxygen but also because of antennas and the increasing free space loss.

LITERATURE

1. Gunther, F. A. — Tropospheric scatter communication etc. IEEE Spectrum Sept. 1966, pp. 79 - 98

2. CCIR-Rep. 258-2 — ITU CCIR XII Plenar Assembly Geneva 1974, Vol. VI

3. CCIR-Rep. 322-1 — World distribution and characteristics of atmospheric radio noise. ibid.

4. Davies, K. — Ionospheric Radio Propagation National Bureau of Standards, Boulder NBS-Monograph 81

5. AGARD Conf. Proc. 13 — Oblique Ionospheric Radiowave Propagation, Editor T. B. Jones, June 1969, Technivison Services

6. Bailey, D. K. — The effect of multipath distortion on the choice of operating frequencies for high-frequency communication circuits. Trans IRE, AP-7, p. 398

7. AGARD Conf. Proc. 173 — Radio Systems and the Ionosphere (Athens, 26-30 May, 1975)

8. CCIR Rec. 456 — Data transmission at 1200/600 bit/s over HF circuits when using multichannel voice-frequency telegraph systems and frequency shift keying. ITU XIII Plenar Assembly Geneva 1974

9. Hillam, B. and Gott, F. — An experimental evaluation of interleaved block cooling in aeronautical HF-channels, AGARD Conf. Proc. on Dig. Communications in Avionics

10. CCIR Rec. 369-1 — Definition of a basic reference atmosphere. ITU CCIR XIII Plenas Assembly, Geneva 1974, Vol. V

11. Großkopf, J. — Wellenausbreitung I, Bibliograph. Institut Mannheim/Wien/Zürich No. 141/141a

12. Fehlhaber, L. and Gilois, H. G. — Effects of Nocturnal Ground-Based Temperature Inversion Layers on Line-of-Sight Radio Links, AGARD Conf. Proc. 208 EM Propagation Characteristics of Surface Materials and Interface Aspects

13. Bullington, K. — Radio Propagation Fundamentals Bell Syst. Tech. J. Vol. 36 pp. 593-626, May 1957

14. Kerr, D. E. — Propagation of Short Radio Waves Dover Publications, 1951

15. Ince, A. and Williams, P. — Influence of Topography and Atmospheric Refraction in VHF Ground-Air Communication. AGARD Conf. Proc. 144 on Electromagnetic Wave Propagation Involving Irregular Surfaces etc.

16. Beckmann, P. and Spizziccino, A. — The Scattering of Electromagnetic Waves from Rough Surfaces, Macmillan, New York, 1963

17. Hortenbach, K. J. — Die statistischen Eigenschaften der Feldstärkeschwankungen auf Satelliten-Flugzeug-Funkstrecken in verschiedenen Höhen über unregelmäßigem Gelände. Dissertation Techn. Hochschule Aachen, 1977

18. Barrick, D. E. — Rough Surface Scattering Based on the Specular Point Theory IEEE Trans AP-16, No. 4, July 1968, pp. 449 - 454

19. Piquenard, A. — Radio Wave Propagation 1974

20. Dixon, R. C. — Spread Spectrum Systems Wiley 1975, New York

21. Weibel, G. E. and Dressel, H. O. — Propagation Studies in Millimeter-Wave Link Systems Proc. IEEE Vol. 55, pp. 497 - 513, Apr. 1967

22. Strohbehn, J. W. and Clifford, S. F. — The Theory of Microwave Line-of-Sight-Propagation Through a Turbulent Atmosphere. IEEE Trans. AP-18, pp. 264 - 274, Mar. 1970

23. Lebow, I. L. et al. Satellite Communications to Mobile Platforms
Proc. IEEE, Vol. 59, No. 2, Feb. 1971, pp. 139 - 154

24. USAF Studies Spacecraft Survivabibility, Aviation Week & Space Technology, August 4, 1975,
pp. 41 - 42

TABLE I

Relevancy of Communication Problems in Avionics

Scenarios		Propagation mode	Frequency ranges	Medium effects	Boundaries to be considered	Bandwidth limitation
Air-Air	short distance	diffraction ionospheric Ref. (line of sight)	HF, VHF UHF	Ionosp.Condition ducting	Ionospheric Layers ground	Ionospheric M.P. multiple ground scatter
	long distance	line of sight Ionosph.Refr.	HF, VHF UHF	Ionosp.Condition ducting	Ionospheric Layers ground under low grazing angle	Ionospheric M.P. direct-ground reflected w.
Air - Ground		line of sight Ionosph.Refr.	HF, VHF UHF	Ionosp.Condition ducting	Ionospheric Layers ground under low grazing angle	Ionospheric M.P. direct-ground reflected w.
Satellite		line of sight	UHF, SHF EHF	Ionosp.Cond. Trop.ray bend. absorption	ground	(Scintillation) Multipath under low elevation
Terrest.supp. mobile		ground wave Ionosph.Refr. (line of sight) diffraction	HF, VHF UHF	Ionosp.Cond. ducting	Ionospheric Layers ground	Ionosp. M.P. Multipath and Ground scatter
Terr,support fixed		line of sight Ionosph.Refr. scatter	UHF, SHF HF	ray bending turbulence absorption	trop.layers ground	direct-ground ref.w. Multiple scatter in common vol.

TABLE II

Parameters of the Path of Rays in the Atmosphere

dn/dh N-units/km	Curvature	K	Atmospheric refraction	Virtual earth	Horizontally launched ray
> 0	upwards	1	below normal	more conve than actual earth	moves away from the earth
0	nil	1		actual earth	
$0 > \frac{dn}{dh} > -39$	downwards	1		less convex than actual earth	
-39		4/3	normal		
$-39 > \frac{dn}{dh} > -157$		4/3	above normal		
-157				plane	remains parallel to earth
< -157		0	super-refraction	concave	draws closer to the earth

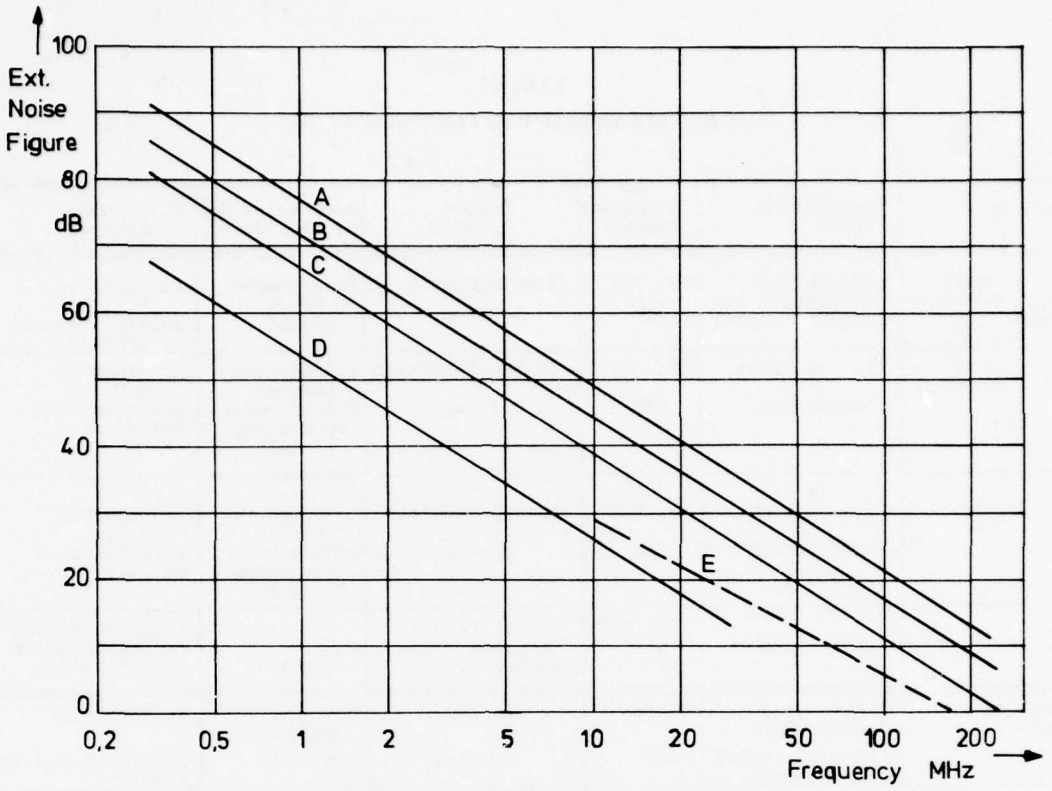


Fig.1 External noise in HF

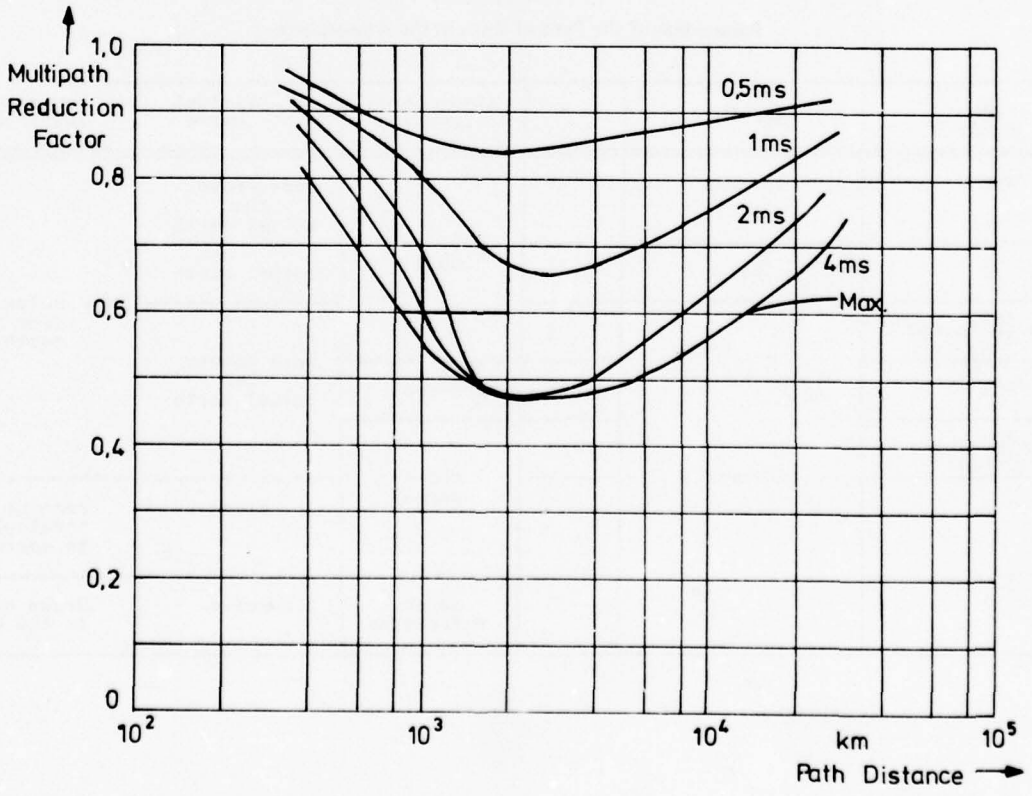


Fig.2 Multipath reduction factor



Fig.3 Statistical distribution of the k-factor in the lower troposphere

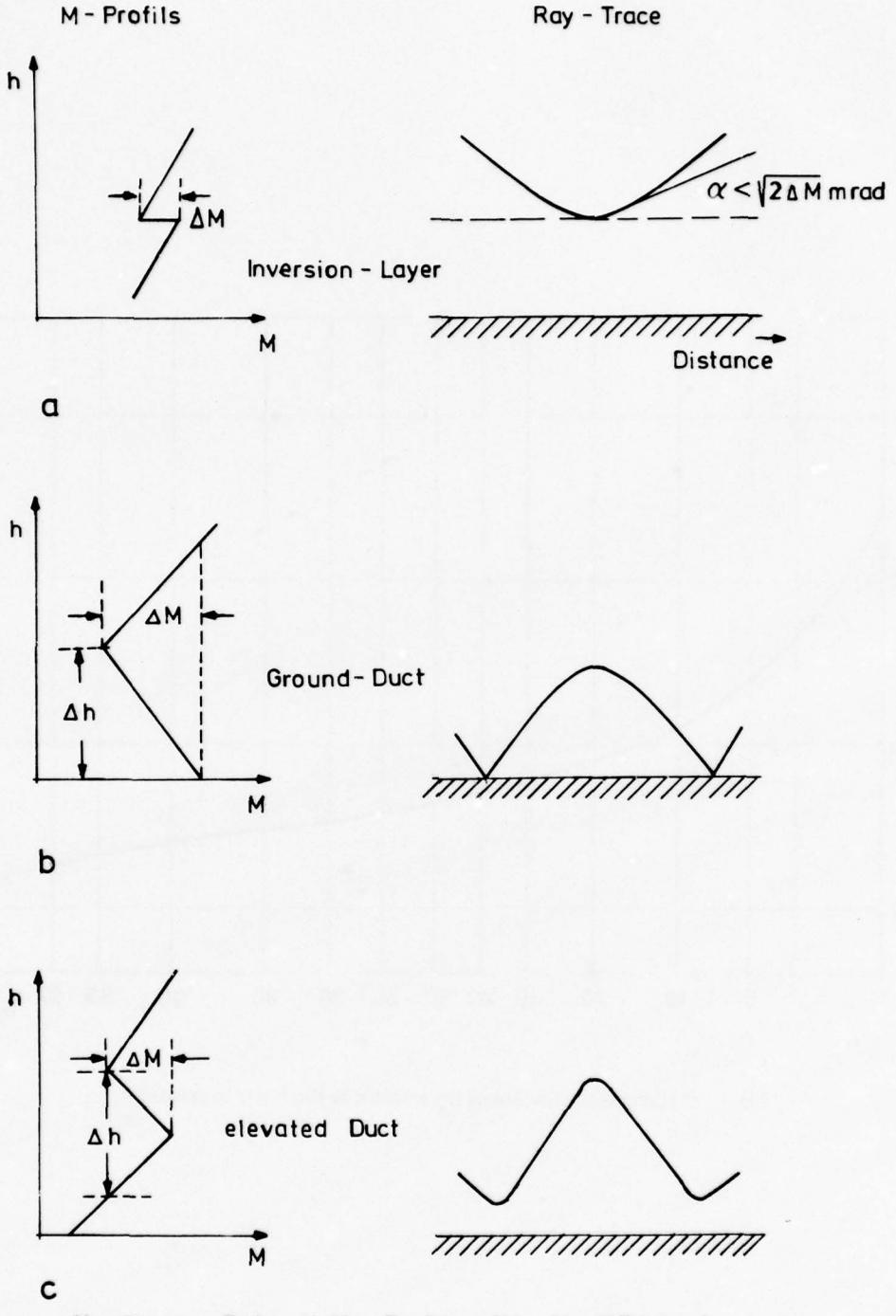


Fig.4 (a, b, c) Inversion layer and duct effects in the lower troposphere

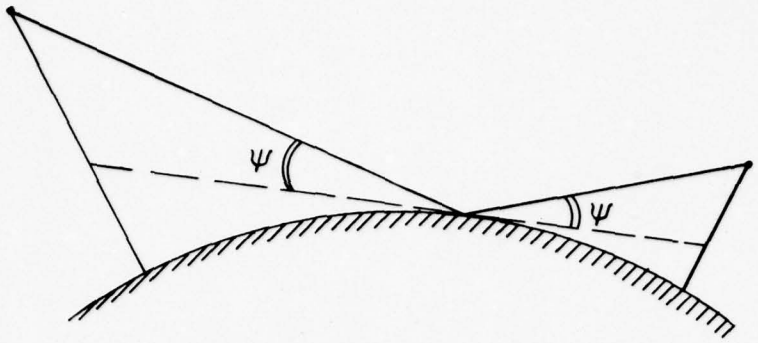


Figure 5

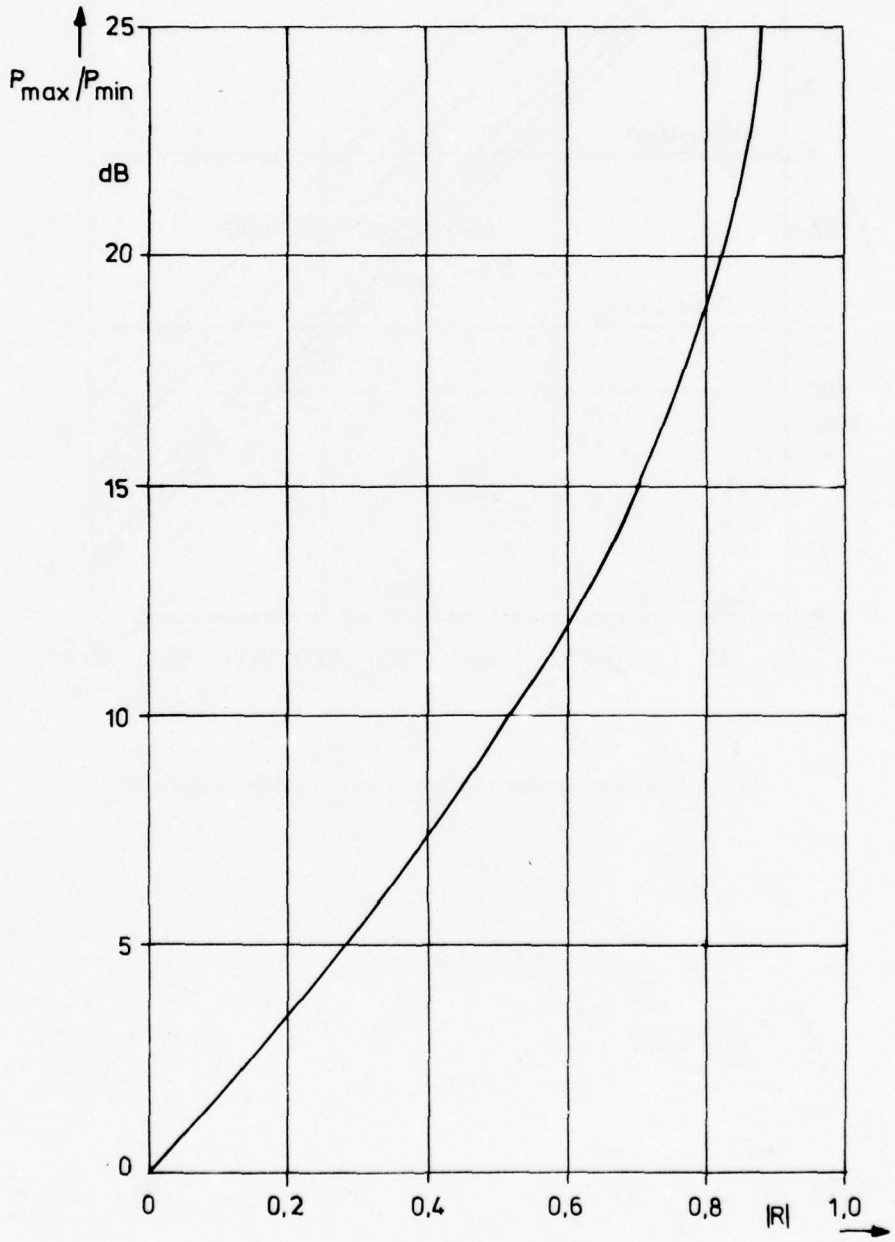


Fig.6 Multipath conditions for air-air, air-ground links

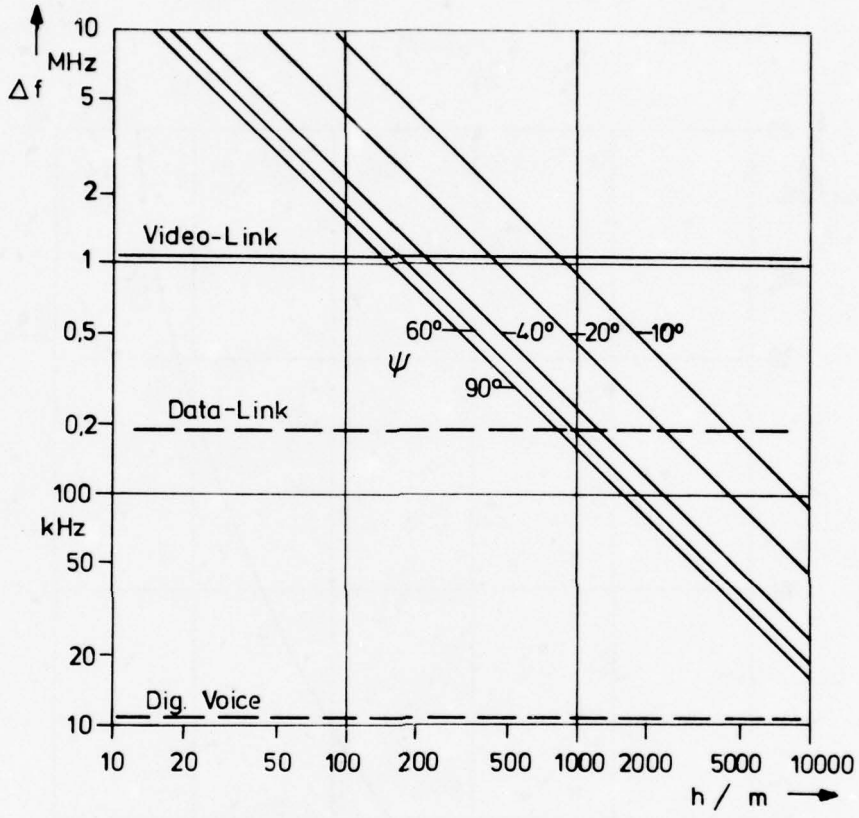
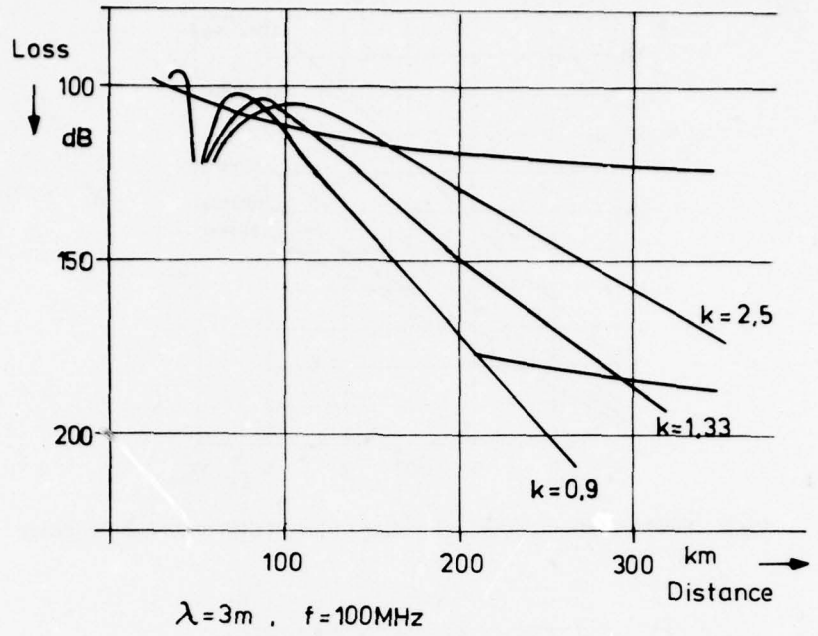
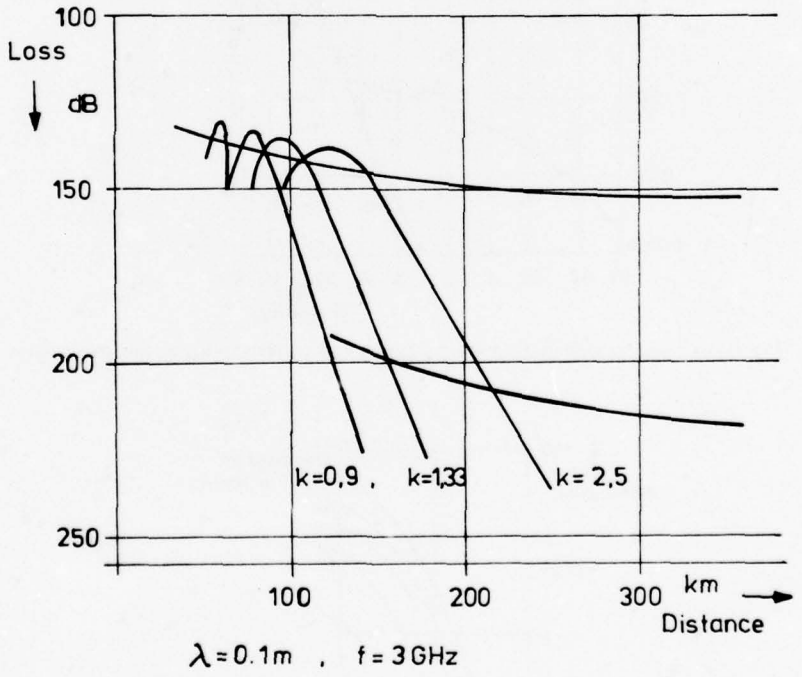


Fig.7 Max. data rate for air-ground links using classical modulation



(a)



(b)

Fig.8 Distance-dependence of overall path loss (isotropic radiators)
Ground antenna height 15 m
Airborne antenna height 600 m

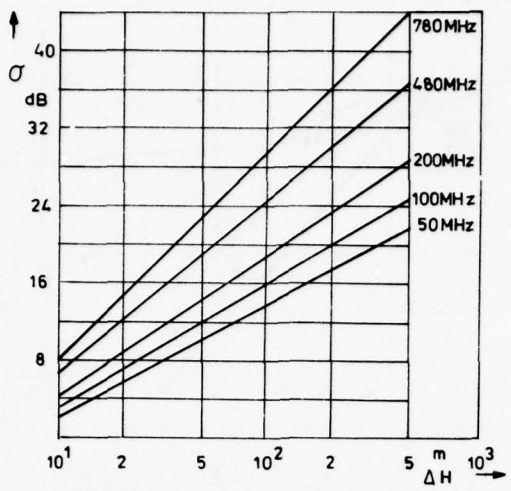


Fig.9 Fieldstrength variation as function of height difference and frequency

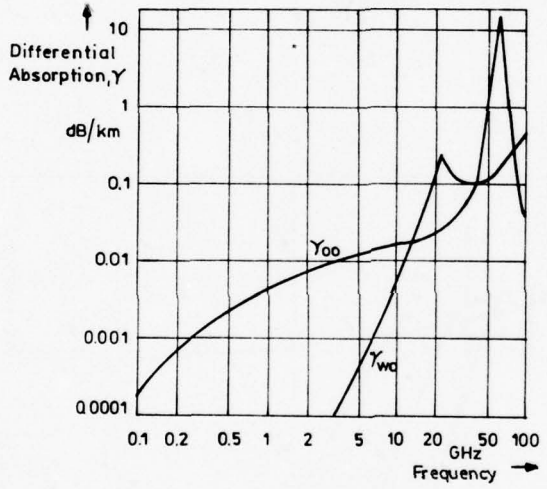


Fig.10(a) Signal attenuation from oxygen and water vapour in the atmosphere

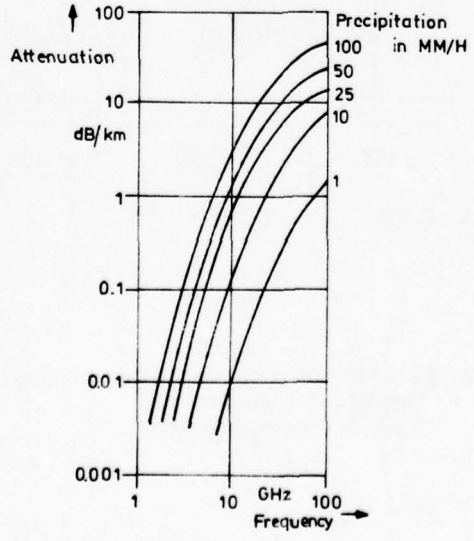


Fig.10(b) Signal attentuation for various rainfall rates

DISCUSSION

H.P.Kuhlen, Ge

What are the influences on propagation due to different antenna polarizations (in the VLF to SHF bands) like vertical, horizontal, circular etc.?

Author's Reply

As in terrestrial wave propagation, the received signal depends on the strength of the ground reflected signal, vertical polarization is to be preferred in principle (VHF and above). At low frequencies (< 5 MHz) this is true as well, as their primary propagation mode is the ground wave.

H.P.Kuhlen, Ge

Did you recognize in the investigations the correlation between

- time of year
- day/twilight/night
- sun spot activity
- antenna polarization

as parameters for propagation? (reliability!)

Author's Reply

In the HF-region, there is a well known dependence of the propagation conditions on daytime, time of year, sun spot activity, when considering skywave, the polarization of the refracted signal usually varies causing polarization fading. On the higher frequency region propagation is dependent on weather conditions which are clearly dependent on the mentioned parameters.

E.Ante, Ge

Are there any known effects in the VHF-region due to tiding during shore to shore communications?

Author's Reply

In terrestrial radio links over water, we get a strong reflected ray (R = −1). When the relative height between radiation (reception) point and sea level varies because of tiding the received field strength will vary between twice the free space value and zero. This effect occurs independent from a high value of the link clearance (f.i. up to the 10th fresnel zone).

MODELLING OF PROPAGATION ASPECTS OF DIGITAL COMMUNICATION SYSTEMS

by

H. R. Raemer
Northeastern University
Boston, Massachusetts

ABSTRACT

This paper reviews certain propagation considerations important in modelling of digital communication systems. It covers nearly the entire radio frequency spectrum. Since the coverage is so broad, the paper is limited to brief discussions of a number of topics with extensive references to the literature. It was impossible within the limited space to include in-depth mathematical derivations or accounts of system details. Section II covers some basic channel modelling theory concepts. Section III covers some propagation topics of importance in digital communication systems analysis and design, i.e., atmospheric, extraterrestrial and man-made noise, refraction, diffraction, interference, path attenuation, propagation through random media, scattering from random surfaces and multipath effects. Section IV covers propagation consideration in the various radio bands from VLF through the millimeter band. Section V covers two specific channels, troposcatter and satellite-ground.

I. INTRODUCTION. This paper reviews some considerations important in modelling propagation aspects of digital communication systems over the radio spectrum. Before proceeding with such a vast topic, one must ask whether there exist propagation effects important in digital communication that are not equally important in analog communication. A partial answer is that most analog systems can stand more signal parameter distortion than can the high data rate digital systems now in increasing use for many applications. However, most propagation effects detrimental to digital systems are also detrimental to analog systems. Hence, the analog-digital separation is not usually possible in discussing propagation effects.

In modelling the propagation environment in order to evaluate an analog communication system operating within that environment, emphasis is usually on such parameters as mean SNR, mean or median values of depth of fading, mean outage time, etc. With digital systems evaluation ultimately reduces to error probability calculations, feasible only if propagation channel statistics are known in some detail. Purely deterministic propagation theory is helpful but may be inadequate for analysis of a digital modem.

This paper emphasizes statistical aspects of propagation media and ways in which randomness along the path causes random variations of key signal parameters. In PSK, FSK and on-off keying systems, where mark-space decisions are based on phase, frequency and amplitude respectively, it is particularly important to know the degree to which propagation anomalies introduce random fluctuations in these parameters and the statistics of such fluctuations.

The territory covered is so vast that limits must be imposed on the paper's scope. The spectrum covered is that between 3kHz and 300GHz, thus excluding ELF, infrared and optical channels. Channels included are those wherein propagation occurs within the terrestrial environment but not through man-made ducts or underground. This restriction eliminates such topics as guided millimeter wave transmission, rock-strata channels, links between buried or submerged antennas and communication between ground surface and a mine. All of those channels involve interesting propagation theory and are of practical importance. Their omission does not reflect the author's priorities, but rather the need to limit the paper's length. Similar remarks apply to omission of deep-space channels, optical links, etc.

Section II covers certain considerations in theoretical modelling of digital communication channels. The discussion focusses on the work of P. A. Bello, whose contributions in this area are significant. Section III covers some propagation phenomena affecting both analog and digital communication systems, emphasizing their effects on digital systems. Section IV covers propagation considerations within various frequency regions, beginning with VLF and proceeding through centimeter and millimeter bands. Section V covers two specific channels, namely troposcatter and satellite-ground.

In a paper reviewing such a broad subject, it is important to cite key source references on each topical area covered. An IEEE Press publication containing key papers on Channel Modelling [Goldberg, 1976] covered some topics so well that the author drew heavily from that source. In citing references, some very "classical" propagation papers were included because they contain essentially everything needed for a good channel model. The more modern papers cited (except for those on certain topics of great recent interest, such as millimeter wave propagation) tend to be those dealing more directly with digital communication channels, where issues of propagation theory are inextricably interwoven with those of communication theory. Where this close interaction exists, the propagation theory is often of the very simple "first order" variety.

II. CHANNEL MODELLING THEORY APPLICABLE TO DIGITAL COMMUNICATION SYSTEMS. A radio communication channel can usually be modelled as a time-variant linear system which is a superposition of deterministic and random components. Important contributions to this type of theoretical modelling of a number of radio communication channels were made by P. A. Bello [1963, 1966, 1969, 1973], whose work stresses the idea that a complete channel description is provided by certain specified functions. The channel output waveform $w(t)$ and its complex spectrum $W(\omega)$ are given in terms of the input time waveform $z(t)$ and its complex spectrum $Z(\omega)$ by $w(t) = \int_{-\infty}^{\infty} d\omega\, Z(\omega)\, T(\omega,t)\, e^{j\omega t}$ (1) and $W(\omega) = \int_{-\infty}^{\infty} dt\, z(t)\, M(t,\omega)\, e^{-j\omega t}$ (2), where $T(\omega,t)$ is the "Time-Variant Transfer Function" of the channel and $M(t,\omega)$ is the "Frequency Dependent Modulation Function". Those names imply the roles of these functions, e.g. $M(t,\omega)$ applies amplitude and phase modulation to the input signal $z(t)$. Either $T(\omega,t)$ or $M(t,\omega)$ contains all deterministic and statistical channel information. The transfer function $T(\omega,t)$ and a function $G(\omega,\nu)$, called the "Output Doppler Spread Function", are Fourier transform pairs, as are $M(t,\omega)$ and another function $h(t,\xi)$, called the "Output Delay Spread Function". The latter is equivalent to a time-variable impulse response of the channel.

Another characterization of the channel is in terms of "delay-Doppler coordinates". In this scheme, we have $w(t) = \iint_{-\infty}^{\infty} d\xi\, d\nu\, z(t-\xi)\, e^{j\nu(t-\xi)}\, V(\nu,\xi)$ (3), $W(\omega) = \iint_{-\infty}^{\infty} d\xi\, d\nu\, Z(\omega-\nu)\, e^{j\xi(\omega-\nu)}\, U(\xi,\nu)$ (4), $U(\xi,\nu) = e^{-j\nu\xi}\, V(\nu,\xi) = \int_{-\infty}^{\infty} dt\, g(t,\xi)\, e^{-j\xi t} = \int_{-\infty}^{\infty} d\omega\, H(\omega,\nu)\, e^{j\xi(\omega-\nu)}$ (5), where $g(t,\xi)$, $U(\xi,\nu)$, $H(\omega,\nu)$, $V(\nu,\xi)$ are respectively the "Input Delay Spread Function", its spectrum, the "Input Doppler Spread Function" and its transform. The interpretation of Eq. (3) in terms of propagation of a radio wave through a medium is that the transmitted wave encounters a continuum of "scatterers", each imposing on the signal a delay ξ and a Doppler shift ν. Each scatterer with its delay-Doppler pair (ξ,ν) has a weighting $V(\nu,\xi)$.

Equation (4) yields a similar interpretation applied to the complex spectrum of the transmitted signal waveform rather than the waveform itself.

In applying these ideas to digital communication channels, it is recognized that w(t) is a random process due to the randomness of the channel, whether the input z(t) is random or deterministic. The result of analysis of such a system is an error rate. Usually, (particularly with Gaussian statistics), the result involves correlation functions of w(t), which contain complex two-dimensional correlation functions of the function characterizing the channel, e.g. the correlation function of $T(\omega,t)$, $R(t,\tau,\omega,\Omega) = \langle T^*[\omega - (\Omega/2), t - (\tau/2)]\, T[\omega + (\Omega/2), t + (\tau/2)]\rangle$ (6).

Random variations of the Time-Variant Transfer Function (or other function characterizing the medium, e.g. the Frequency Dependent Modulation Function) are of two types: slow non-Gaussian and fast Gaussian. The latter can usually be considered as wide-sense stationary over a long time interval while the medium's constitutive parameters remain nearly constant. Hence fast fluctuation statistics of many channels can be considered as wide-sense stationary (WSS). Also, the scatterers can often be considered uncorrelated (US); when both conditions prevail, the channel can be modelled as "wide-sense stationary-uncorrelated scatterers" (WSSUS). In such cases correlation functions are simplified, becoming functions of delay and Doppler differences alone, e.g. Eq. (6) becomes $R(t,\tau,\omega,\Omega) = R(\tau,\Omega)$ (7). Bello [1965, 1966, 1969, 1973] has applied his modelling concepts to a number of radio communication channels. This will be discussed later in the context of those channels (Sections IV.C, V.A, V.B).

III. PROPAGATION PHENOMENA. In Section III some key propagation phenomena affecting digital communication systems are discussed. The classification in this section is in terms of the phenomenon rather than frequency regions. The latter classification is used in Section IV.

A. Atmospheric Noise. In terrestrial radio systems operating at all frequencies there is a certain amount of noise originating in the atmosphere [Watt and Maxwell, 1957; Watt et al, 1958; Pierce, 1969]. Being additive to the signal it is indistinguishable from internal receiver noise if it has Gaussian statistics and is spectrally flat over the receiver passband as is most internal noise. The primary sources of atmospheric noise are intense current surges in cloud-to-ground and intercloud lightning flashes within the world's thunderstorms. Each surge generates a radio pulse whose energy is primarily in the kHz region but has components as high as 40MHz and beyond. Hence, lightning generated noise is primarily at VLF, LF and MF, but may still limit communication well into HF. The earth-ionosphere waveguide (Section IV.A) carries VLF waves over hundreds or thousands of kilometers. As frequency increases, attenuation of "waveguide"-propagated RF energy increases; as we proceed through LF and into MF, noise decreases because less is generated at its source and it is less efficiently propagated. The higher the frequency, the closer the receiver must be to the source in order that the noise will seriously limit performance.

It follows from central limit theorem arguments that some atmospheric noise, being a superposition of many independent random variables, has Gaussian statistics; i.e., its amplitude is Rayleigh or Rice distributed, the latter with a nonzero mean value, the former otherwise. This is observed for high atmospheric noise levels, where the high intensity of the noise indicates that it came from a large number of lightning discharges. For lower noise levels at the tail of the distribution where the number of sources is too small for validity of central limit theorem arguments, the noise is more impulsive and (approximately) lognormally distributed. Digital communication system errors are, in some cases, caused more frequently by short intense noise pulses than by steady Gaussian noise. Hence, although the latter type is more prevalent, the former is often more important in evaluating system error statistics. In accounting for atmospheric noise in error calculations it is insufficient (for reasons indicated above) to add the atmospheric noise power level to the internal noise thus increasing SNR. This procedure applies only to the Gaussian portion of the atmospheric noise. To include the non-Gaussian portion we must determine the amplitude (and possibly phase) PDF [Bello, 1965; Omura and Shaft, 1971]. The noise waveform (or envelope) ACF is also important. Hence, atmospheric noise measurements should result in ACF's for maximum utility in evaluating digital communication systems. World-wide atmospheric noise power and amplitude probability distribution measurements [Disney and Spaulding, 1970] have been made regularly for years by scientific organizations. These results can be used for error rate calculations provided corrections are made for bandlimiting effects of filters used in measurements.

In recent years there has been considerable work on construction of analytical models for atmospheric noise suitable for calculation of digital error rates [Giordano and Haber, 1972; Field and Lewinstein, 1978].

B. Extraterrestrial Noise. The ionosphere is nearly transparent to radio waves at frequencies beyond a few megahertz. Beyond 50MHz radio noise originating on the sun, on the galactic plane and in certain noisy constellations penetrates the ionosphere and significantly affects communication [Bolton and Westfold, 1950; Smith, 1960]. This is particularly true at the high end of UHF and in SHF and millimeter bands, where extraterrestrial noise may sometimes be the ultimate performance limiter. Because most extraterrestrial noise can be regarded as a superposition of many random variables, it can usually be modelled as Gaussian noise spectrally flat over the receiver passband and merely added to internal noise to increase SNR.

C. Man-Made Noise and Interference. Man-made sources of unwanted radio signals exist in most terrestrial environments, especially in urban areas [Disney and Spaulding, 1970; Esposito and Buck, 1973; Herman, 1971]. Included are impulse noise (e.g., ignition noise) and modulated CW tones (e.g., nearby radio transmissions). Power lines and vibrating machinery may be sources of radio noise. If transmissions from enough man-made sources are simultaneously active, then by central limit theorem arguments their aggregate behaves like Gaussian noise and can be treated as such in analysis. If only a few discrete impulsive sources are present, their statistics are non-Gaussian and error rate calculations are more difficult.

D. Atmospheric Refraction, Diffraction and Interference. The density and hence refractive index of the earth's atmosphere is vertically stratified. Resulting transhorizon bending of radio waves [Bullington, 1947; Ikegami, 1959] affects long distance transmission and is often modelled by assuming an earth radius of 4/3 of the true value. This convenient rule-of-thumb is not always applicable. Atmospheric refraction is highly dependent on meteorological conditions along the path. Refractive index variation with altitude can be predicted theoretically. Propagation analysis, for paths with variable refractive index, can be done numerically given index profiles (usually determined experimentally, because theory is based on idealized atmospheres which rarely exist).

The earth represents a diffracting obstacle to radio waves. Theory of propagation around the curved earth and corroborating experimental results yield a diffraction pattern with valleys and peaks respectively

reducing and enhancing reception in certain zones beyond the horizon. Protuberances, e.g., mountains, buildings, etc., are nearly opaque at higher frequencies, having dimensions of many wavelengths. At lower frequencies, where wavelength and dimensions are comparable, they become diffracting obstacles, allowing reception beyond the obstacle. Rigorous EM theory describing these effects may be intractable, necessitating empirical methods of evaluation. However, simple theory exists (e.g., theory of knife-edge diffraction) by which rough orders of magnitude can be determined [Bullington, 1947; Dickson et al, 1953; Reudink and Wazowicz, 1973].

Interference between direct and ground-reflected waves is important in propagation between two points with height comparable to or greater than a wavelength. This condition exists in air-to-air and air-ground links from MF into UHF and in ground communication between elevated antennas from HF into UHF. One wave path from a transmitted antenna is directly to the receiver ("direct wave"), the other indirectly to the receiver through ground reflection ("reflected wave"). The field at the receiver is a superposition of direct and reflected waves whose phase difference is $(2\pi\delta)/\lambda$ (δ = path length difference, λ = wavelength). If these two waves have comparable magnitudes constructive interference produces a peak when $\delta = n\lambda$ where n is an integer and destructive interference results in a valley when $\delta = [n + (1/2)\lambda]$. Simple "flat earth" theory yields the result $\delta \simeq (2h_T h_R)/\lambda d$ (8), where d is separation distance and h_T and h_R (both assumed small compared to d) are transmitter and receiver heights respectively.

Theory and observation show that fluctuations in signal amplitude or phase can occur due to motion of transmitting and/or receiving platforms (as in air-air or air-ground transmission), or fluctuations in reflecting point position (as with a rough sea reflecting surface), resulting in errors in digital links operating in such environments.

E. Scattering from Random Surfaces. The theory of random surface scattering has received significant attention in the literature. Emphasis has been placed on backscatter from a rough sea surface in connection with radar applications. In communication systems in ocean environments, emphasis is on low grazing angles, i.e., near-forward scattering.

S. O. Rice [1951] modelled surface roughness in terms of small deviations from flatness treated as small perturbations. The validity of this approach requires height fluctuations small compared with wavelength. Rice's first order results, assuming Gaussian statistics for height deviation, provided a basis for much work by subsequent researchers [Beckmann and Spizzichino, 1963; Fuks, 1966].

An approach to modelling of large scale surface roughness is the "Kirchoff approximation", where the field at a surface point is assumed to be that on an infinite plane tangent to the surface at that point. This model requires a local radius of curvature large compared to a distance parameter proportional to wavelength.

A true sea surface consists of fluctuations of a continuum of scale sizes, ranging from those small compared to wavelength ("capillary waves") to those very large compared to wavelength ("sea swell" or "gravity waves"). The scale distribution depends on sea state. No tractable theory covers all possibilities; hence existing theories consider large scale deviations from the smooth sea condition on which are superposed small scale perturbations ("two scale model"). This is about the best that current theory can do. Field experiments to determine sea scattering parameters are so dependent on widely variable sea conditions that communication system analysts must rely on results of simple theory or simulation techniques [Zornig and McDonald, 1975]. P. A. Bello [1973] has considered sea surface scattering in modelling of aeronautical channels (Section V.B).

F. Propagation Through Random Media. In the high UHF, SHF and millimeter regions, propagation is primarily line-of-sight. In ground-to-ground, air-to-air, air-ground and space-ground links at those frequencies the wave propagates through atmospheric regions with random irregularities ("turbulence") in refractive index with scale comparable to wavelength or larger. For example, suppose typical irregularity dimensions are roughly 50cm. A 3GHz wave sees this as 5λ long whereas a 3MHz wave sees it as $.005\lambda$. In turbulent atmospheres many irregularities appear within a wavelength. Their effect "averages out", producing an "equivalent constant refractive index". Thus, a 3MHz field doesn't fluctuate due to irregularities on that scale. However, a 30GHz wave experiences fluctuations of scales including many times λ. This results in temporal and spatial fluctuations in amplitude, phase, polarization and wave normal direction. These effects become more pronounced at millimeter and optical wavelengths.

The theory of propagation through random or turbulent media has large coverage in the optics literature. A well-known source reference on the subject is a book written by Tatarskii [1961]. Most of the work is in optics and acoustics contexts, but there have been applications to centimeter and millimeter radio waves [Clifford and Strohbehn, 1970; Hong et al, 1977; Ott, 1972; Radio Science, 1975]. The theoretical model involves assignment of a small spatially variable perturbation component of refractive index about a value of unity. The key simplifying assumption is $\lambda << \ell_o$ where ℓ_o is the scale of turbulence. This reduces vector wave equations to scalar form removing effects of turbulence on polarization and reducing the analysis to that used with acoustic or non-coherent light waves. Clifford and Strohbehn [1970] show that the analysis is valid when $\lambda \gtrsim \ell_o$, implying validity in the microwave region. Most calculations based on this theory culminate in the two-dimensional spectra for the log-amplitude and phase, the two-dimensional covariance function or the "structure functions", of log-amplitude and phase, the latter being mean-square differences of two spatially separated log-amplitude or phase values.

Current theories of propagation through turbulent media do not lead to analytically tractable probability density functions of the random signal seen at the receiver. (Temporal spectra can be inferred from spatial spectra predicted by the theory.) It is difficult to adapt the results to communication channel modelling theory or to relate them directly to error probability calculations.

G. Multiplicative Noise - Multipath. In many channels (HF, Tropo, Microwave, line-of-sight, etc.), there is multiplicative noise arising from path irregularities. One can usually model the medium as a set of scattering regions, seen by the receiver as point scatterers s_1, s_2, ... s_n. The mechanism for generation of multiplicative disturbances on the transmitted signal can be viewed as arising from single scatterings from s_1, ..., s_n. (More accurate theory is discussed in Section III.G; this section treats a much simpler viewpoint.)

To generate this ultra-simple model we assume a sinusoidal transmitted signal of frequency ω. The field phasor for the transmitted wave incident on s_ℓ is $E_{i\ell}(\omega) = \left\{Z(\omega)\, F(\omega, \underline{\Omega}_{T\ell})\, e^{-j\omega[t-(r_{T\ell}/c)}\right\}/r_{T\ell}$ (9)

where $Z(\omega)$ is the Fourier transform of the transmitted signal waveform, $r_{T\ell}$ is distance from transmitter to s_ℓ, $\underline{\Omega}_\ell$ is a vector indicating s_ℓ's angular position relative to transmitter, and $F(\omega, \underline{\Omega}_{T\ell})$ contains transmitter and medium parameters. The scattered signal phasor at the receiver for a transmitted sinusoidal

signal at frequency ω is $W_\ell(t,\omega) = \left\{[Z(\omega)\ F(\omega,\underline{\Omega}_{T\ell})]\ G(\omega,\underline{\Omega}_{\ell R})\ e^{-j\omega\{t-[(r_{T\ell}+r_{\ell R})/c]\}}\right\}/r_{T\ell}\ r_{\ell R}$ (10). where $r_{\ell R}$ is distance from s_ℓ to receiver, $\underline{\Omega}_{\ell R}$ represents receiver angle relative to s_ℓ and $G(\omega,\underline{\Omega}_{\ell R})$ contains the scattering pattern of s_ℓ, the receiver's antenna pattern and other receiver and medium parameters.

Assuming that $F(\omega,\underline{\Omega}_{T\ell})\ G(\omega,\underline{\Omega}_{\ell R})$ is frequency-independent over the transmission bandwidth (a reasonable approximation except for extremely wideband transmission), then superposing phasors at all transmitted frequencies, we obtain $w_\ell(t) = \int_{-\infty}^{\infty} d\omega\ w_\ell(t,\omega) = C\ z(t - \xi_\ell)$ (11), where $\xi_\ell = (r_{T\ell} + r_{R\ell})/c$ = delay for s_ℓ, $C = FG/r_{T\ell}R_{\ell R}$.

If s_ℓ moves relative to transmitter and/or receiver, or if transmitter and/or receiver move relative to a stationary s_ℓ then $r_{T\ell}$, $r_{\ell R}$, $\underline{\Omega}_{T\ell}$ and $\underline{\Omega}_{\ell R}$ are time-varying. This implies dependence of C on both t and ξ_ℓ. Summing contributions from all scatterers and accounting for the indicated functional dependence on t and ξ_ℓ for each scatterer, we obtain $w(t) = \sum_{\ell=1}^{N} z(t - \xi_\ell)\ g(t,\xi_\ell)$ (12). Modelling the medium as a continuum rather than a collection of discrete scatterers, we express Eq. (12) in the form $w(t) = \int_{-\infty}^{\infty} d\omega\ g(t,\xi)\ z(t - \xi)$ (13). Equation (13) is used by Bello [1963] in his basic paper on channel modelling (Section II). He calls $g(t,\xi)$ the Input Delay Spread Function.

The component of scatterer motion along the line-of-sight induces a Doppler shift in the received signal. This is modelled by expressing $g(t,\xi)$ as the Fourier transform of the "Delay-Doppler Spread Function" $U(\xi,\nu)$, $g(t,\xi) = \int_{-\infty}^{\infty} d\nu\ U(\xi,\nu)\ e^{j\nu t}$ (14). Equation (3) follows from (14) with the aid of (5). Equation (3) expresses the received signal waveform as a superposition of delayed and Doppler-shifted transmitted signal waveforms weighted with a function of both delay and Doppler shift, the latter function being determined by electromagnetic properties of the medium.

If the transmitted signal is purely sinusoidal and the medium consists of a small number of discrete scatterers (ionospheric scattering points in HF channels, scatterings from a few terrain protuberances small compared to path length but large in terms of wavelength, etc.), then Eq. (3) becomes $w(t) = \mathrm{Re}\left[\sum_{\ell=1}^{N} |g(t,\xi_\ell)|e^{-j\omega[t-\psi(t,\xi_\ell)]}\right]$ (15), where $g(t,\xi_\ell)$, a complex function, is expressed in terms of its amplitude $|g|$ and its phase ψ. Each of these has a deterministic part and a random part. Random fluctuations of amplitude and phase of the signals from the scatterers produce fading. For example, given two paths with comparable signal amplitude, constructive or destructive interference occurs during different intervals causing the composite signal to experience peaks and valleys. If the amplitude and phase fluctuations of each of a large number of scatterers are random and mutually statistically independent, then the central limit theorem shows that the received signal statistics are complex Gaussian, resulting in Rayleigh or Rice-distributed fading. If the scatterers are so widely separated that the propagation path lengths are widely different, then short-duration signals, e.g., pulses as in digital signalling transmissions, are completely resolved, resulting in intersymbol interference.

Multipath and its resulting fading and intersymbol interference problems [Cox and Leck, 1975; Liu and Yeh, 1975; Nesenbergs, 1967; Painter et al, 1973; Ruthroff, 1971] will be further discussed in Sections IV and V in the context of the topics treated therein. Fading arising from causes other than multipath (e.g., ducting, etc.) will also be discussed.

IV. PROPAGATION CONSIDERATIONS IN RADIO FREQUENCY BANDS. Section IV contains discussions of propagation phenomena affecting digital communications systems operating in various frequency regions from 3kHz to 300GHz.

A. Very Low Frequency (VLF, 3-30kHz, λ = 10-100km). Terrestrial radio propagation at VLF can be effectively modelled with "waveguide mode theory" [Wait, 1957], based on the idea that earth and ionosphere being very good conductors at VLF, form boundaries of a spherical waveguide. Propagation occurs much as it does within metal-bounded waveguide. One can deduce many results of waveguide mode theory to good approximation from a simple model using planar waveguide theory (i.e., neglecting earth curvature) and image theory, in which the field at a receiving point is a superposition of waves originating from image points within the earth and ionosphere. Locations of these image points are derived from the law of reflection for the multiple bounces between the ground and ionosphere. Superposing single, double, triple reflections, etc., one obtains the same result as those of the formal boundary value problem involving earth, air and ionosphere.

Mode theory shows that VLF waves can propagate over enormous distances, i.e., hundreds or thousands of kilometers. Excluding the attenuation factor $e^{-2\alpha d}$, variation with distance d in the case $d/Re \ll 1$, where Re is earth radius (roughly valid for d less than 1000km and not too far off out to 2000km) is roughly as $1/\sqrt{d}$, as opposed to the usual 1/d variation in free-space. This makes for a slow dropoff in received power out to very long distances.

VLF radio transmission is very stable and fading is not usually a major problem. Some fading may occur due to ionospheric fluctuations, but because ionospheric distances are not large compared to wavelength, these effects tend to average out over a wavelength. Hence, fading is a minor effect, and atmospheric noise (Section III.A) is a major performance limiter [Omura and Shaft, 1971].

B. Low Frequency (LF, 30-300kHz, λ = 1-10km) and Medium Frequency (MF, 300kHz-3MHz, λ = 0.1-1km). At LF and MF, the "waveguide" model can be replaced by a view of the received wave as a superposition of a ground wave and one or more sky waves resulting from ground and ionosphere reflections. The ground-wave [Goldberg, 1966; Wait and Walters, 1963] is essentially vertically polarized but with a slight forward tilt in its E-vector due to continuous flow of energy into the earth along the propagation path. The dependence of loss on the earth's constitutive parameters is complicated and different functional dependence on these parameters is exhibited in different frequency regions. The ground wave is the principle mode of propagation for MF communication. The ionosphere reflects only a small fraction of the MF wave energy incident upon it and usually cannot support significant skywave transmission. However, ionospheric reflections produce interference and hence cannot be neglected. At high MF (1.5-3MHz), skywave is sometimes the primary transmission mode. The field-strength for ground wave propagation follows roughly an inverse path length law with modifications due to dependence on polarization, frequency and earth constitutive paramete

This transmission mode is very stable and exhibits little fading. A rough sea path, with its parameter fluctuations, exhibits some fading but it is a minor effect compared to that occurring in skywave transmission. In studies of digital communication systems at LF and MF emphasis is often on atmospheric noise (Section III.A) rather than fading.

C. High Frequency (HF, 3-30MHz, λ = 10-100m). HF communication systems [Goldberg, 1966] operate on the sky-wave transmission mode. The ground wave above 3MHz is highly attenuated and increases with frequency. An ionospheric reflection at long range produces a signal far more intense than the ground wave signal. An ionospheric "reflection" is really a refraction of the upgoing wave as it traverses a gradually increasing refractive index at the lower end of an ionospheric layer. The wave vector's upward component gradually decreases with altitude until the wave "turns around" and propagates downward. This is interpretable as a reflection from an abrupt discontinuity at an "equivalent height". If the RF exceeds a lower layer's critical frequency, it penetrates the layer, but is "reflected" from the next highest layer. If its frequency exceeds the MUF for all ionospheric layers at the particular range of incidence angles corresponding to a given transmitter-receiver separation, then it penetrates the ionosphere and never appears at the receiver (not usual at HF).

The D-layer (50-90km) exhibits high attenuation at HF. Since the D-layer usually exists only in daytime, nighttime reflections occur in E- and F-layers. The E-layer (90-160km) supports some HF skywave transmission for distances up to 2000km. It is present in daytime and partially at night, but the critical frequency is much lower because nighttime electron densities are well below daytime levels. The F-layers (F_1 (160-250km) and F_2(250-450km)) support most long-range HF transmission (i.e., beyond 2000km). For example, a single hop F-layer transmission can exist over a path 4000km or longer at frequencies well above 30MHz (i.e., into VHF). Ionospheric propagation parameters are temporally and spatially variable, depending on sunspot cycles and being different in equatorial, polar, and temperate regions. Realistic HF design for maximum reliability must provide for a wide range of ionospheric conditions, sometimes not easily predictable [Rawer, 1975].

The possibility of reflection from more than one ionospheric altitude creates a major problem in digital communication systems. For example, with a transmitter-receiver separation of 300km, simple theory shows that a single reflection at 100km (E-layer) and a single reflection at 300km (F-layer) will both produce a signal at the receiver. Simple geometric arguments show that the delay difference between those waves is about 1msec. If the path-length is doubled, then the same two "single-hop" signals appear in addition to "double-hop" signals from both layers (ionosphere-to-earth-to-ionosphere-to-earth). Delays for these signals are: 2.1msec for single hop at 100km, 2.8msec for single hop at 300km, 2.4msec for double hop at 100km and 4.4msec for double hop at 300km. The delay separation ranges from 0.3 to 2.3msec.

In addition to this easily predictable source of multipath, other phenomena, less predictable because they are dependent on complicated ionospheric processes, generate additional radio paths, e.g. sporadic E, where clouds of dense ionization traverse the ionosphere at E-layer heights, and Spread F, where F-layer reflections are spread over a continuum. Solar flares, magnetic storms and SID (sudden ionospheric disturbances) generate new random scattering regions introducing additional paths responsible for multipath fading and intersymbol interference. Another agent increasing the number of possible ionospheric paths is magnetoionic splitting due to the terrestrial magnetic field. Ordinary and extraordinary waves travel with different phase velocities and different polarizations; hence even with a single "hop" the received signal field is a superposition of two differently-polarized fields with different path delays. Whatever the multipath mechanism, there are fluctuations in each of the path characteristics. Modelling of the received signal as a superposition of a number of "random scatterer" contributions follows the lines of the discussion in Section III.G.

The net result of these considerations is that HF digital links encounter a high level of multipath with its attendant fading and intersymbol interference. Fading bandwidths range from about 0.05 to 15Hz. Fading amplitude statistics are usually Rayleigh or Rician because of the large number of statistically independent paths contributing to the signal. The time-delay spread of received signals varies from less than 100μsec to about 4msec. Good design can usually keep the spread below 1msec. With highly turbulent ionospheric paths fading rates may exceed 25Hz and correlation bandwidths may be as small as 50Hz [Boys, 1968; Pickering, 1975; Shepherd and Lomax, 1967].

Atmospheric noise and some man-made noise may also be important at HF (Sections III.A, III.C).

To model the HF channel for analysis of digital communication systems, Bello [1965] has considered the sources of error to be fading and atmospheric noise. Fading is assumed slow and non-frequency selective, characterized by a complex Gaussian process. The noise (Section III.A) is assumed lognormally distributed. The basis for this is that errors are caused primarily by occasional large noise spikes, which have lognormal statistics, rather than the more frequent Gaussian "background" noise disturbances. Using this model and parameter values characteristic of HF ionospheric transmission, Bello calculates FSK and PSK error probabilities. Considering both diversity and nondiversity operation, he concludes that with diversity errors due to lognormal atmospheric noise are more frequent than are those due to Gaussian noise at the same SNR and the situation is reversed in the nondiversity case.

The comparison of Bello's error rate calculations with experimental results on real HF systems is good. Other HF channel models [Shaver et al, 1967] have been studied extensively with the aid of channel simulators [Walker, 1966; Watterson et al, 1970]. In a typical channel simulator the input signal is fed to a delay line and is taken off at a number of taps each with adjustable delay. Each tap signal is modulated in amplitude and phase and the sum forms the output. By suitable adjustments of tap delays and modulation functions, propagation behavior of the type following the discussion in Section III.G is simulated. This is an efficient way (much less costly than field testing) to test the many theoretical channel models generated during the past several years. These methods are particularly effective when playbacks of signals used in field measurements can be used in combination with simulated signals.

D. Very High Frequency (VHF, 30-300MHz, λ = 1-10m). At VHF [Bullington, 1947; Vergera et al, 1962], primarily used for short ranges, most propagation occurs along or near the ground and includes line-of-sight and surface wave transmission. There is some sky wave to frequencies as high as 50MHz, but ionospheric propagation plays a much-reduced role at VHF (compared with HF).

Theory and observation both show ground wave transmission above 30MHz to be roughly independent of polarization and of ground constitutive parameters. Elevated antennas are usually used and elevations are within wavelength range, hence interference between direct and reflected wave is important (Section III.D). The field at the receiver is a superposition of four components: (1) The direct wave (as if propagated between two points in free space); (2) The earth-reflected wave; (3) The ground (surface) wave, dependent on ground constitutive parameters; and (4) Fields due to secondary effects of ground, induction effects

and ground irregularities.

From simple analysis the relationship between the free space wave field E_0 and the actual field seen at the receiver can be shown to be approximated by $E/E_0 \simeq 2 \sin (2 h_1 h_2/\lambda d)$ (16), where h_1 and h_2 are antenna heights and d the propagation distance. Equation (16) shows a lobe structure which becomes more sensitive to antenna height and RF as d decreases. For large propagation distances and small antenna heights the ratio is roughly proportional to RF and inverse distance and the lobing is negligible. In many practical VHF systems, it is significant. This is particularly true for air-ground systems or vehicular links where antenna heights may change during transmission.

At VHF, many features discussed in Section III in addition to wave interference are also important, e.g., diffraction around the earth and terrain obstacles, both natural and man-made (Section III.D), many of which are of the order of a few wavelengths at VHF, atmospheric refraction (Section III.D), etc. In particular, ground terrain and buildings may significantly affect VHF communication [Egli, 1957], e.g., large buildings or hills, of the order of tens to hundreds of wavelengths, may be nearly opaque at VHF. At the low end ($\lambda \approx$ 10m), small hills or protrusions, of the order of a few wavelengths at most, are diffracting obstacles and there is a "shadow loss" behind these obstacles which is very sensitive to the geometry. In vehicular links [Clarke, 1968] on paths which involve significant numbers of terrain obstacles, transmitter and/or receiver travel through peaks and nulls causing signal fluctuations. Most of this is relatively small and rarely qualifies as "deep" fading. Another observation is that (usually) these effects are only weakly dependent on polarization.

There have been studies of VHF propagation in forest terrain [Dence and Tamir, 1969; Tamir, 1967], effects of mountains [Dickson et al, 1953] and effects of urban environments, particularly with respect to short-range vehicular communication systems. These studies were motivated by the fact that these are typical environments for VHF links.

From the digital communication viewpoint [Bello, 1973; Painter et al, 1973; Tucker, 1972], we are primarily interested in the error sources in VHF systems operating over their typically short ranges. Some important sources of errors are: (1) ignition noise, largely impulsive, from vehicles passing near the propagation path; (2) atmospheric noise due to local or nearby thunderstorms (since the spectrum of typical lightning strokes has some energy at frequencies as high as 50-60MHz, there is some storm-generated noise; however, unlike such noise at lower frequencies, it propagates along the ground, is attenuated rapidly and is subject to the same kind of vagaries as VHF signal transmissions. Thus such noise tends to be impulsive rather than Gaussian because there are typically only a few contributing lightning strokes at any given time); (3) multipath and its attendant fading due to terrain irregularities and low-flying aircraft (and sometimes ground-based vehicles) that produce a reflected signal superposed on and interfering with the ground-propagated signal. It has been observed [Tucker, 1972] that these effects lead to error bursts in a digital data transmission system.

Channel simulators using tapped delay lines to model medium behavior were discussed in Section IV.C in connection with HF channel simulation. These simulators have also been applied to VHF systems [Arredondo et al, 1973; Bussgang et al, 1974; Walker, 1966]. However, because the VHF channel, involving so many sources of disturbance, is so complex and difficult to model theoretically, this simulation is best accomplished through the "stored channel" concept, wherein signals received on actual field channels are played back into the simulator. This obviates the need for (usually unreliable) assumptions about the channel's amplitude and phase statistics.

E. Ultra High Frequency (UHF, 300MHz-3GHz, λ = 10cm-1m), Super High Frequency (SHF, 3-30GHz, λ = 1-10cm); (Terrestrial Microwave Propagation). Propagation problems important to line-of-sight (LOS) microwave communication are discussed below [Abraham, 1966; Dougherty and Hartman, 1977; Hearson, 1967]. Most considerations discussed in Section IV.D (VHF) apply to UHF, particularly the low end (300MHz-1GHz). Referring to the discussion of various transmission components, as frequency increases into UHF, sky wave becomes negligible (since the ionosphere becomes more nearly transparent), atmospheric noise reduces to negligible importance but extraterrestrial noise penetrates the ionosphere and assumes greater importance. The surface wave is highly attenuated and less important than at VHF. The ground reflected component is still present but with wavelengths below 10cm, smaller antenna structures may be made more directive and angular exclusion of this wave is feasible. Thus most of the propagation is direct LOS. Many terrain obstacles, usually many wavelengths in dimension, are more nearly opaque at UHF than at VHF [Kozono and Watanabe, 1977]. There are some obstacle scattering and diffraction effects and effects of atmospheric refraction and ducting due to pronounced large scale refractive index variations [Ikegami, 1959; Früchtenicht, 1974]. There is also multipath due to random fluctuations in atmospheric refractive index. This turbulence, with scale of the order of a few wavelengths at UHF and SHF, causes random amplitude and phase scintillations [Clifford and Strohbehn, 1970; Ruthroff, 1971; Thompson et al, 1975]. All of these effects assume more importance as RF increases from VHF through UHF because wavelengths become progressively smaller compared with objects and refractive index nonuniformities. These same effects assume even greater importance as RF enters the SHF region, where propagation is entirely LOS. Antennas (horn or paraboloidal dish variety) can be made very small and provide enormous directivity.

An important limitation of systems operating at cm and mm wavelengths is inability to penetrate heavy rainfall. Absorption and scattering by raindrops attenuates waves significantly beyond 30GHz because raindrops, with diameters of a few mm, have dimension of the order of the wavelength. There has been a great deal of theoretical work and field experimentation on rain attenuation of microwaves [Barsis and Samson, 1976; Crane, 1967, 1975; Godard, 1970; Goldhirsh, 1975; Rogers, 1976]. Analysis is based on the Mie theory of absorption and scattering of an electromagnetic wave by a lossy sphere. To calculate these effects, one assumes single scattering, calculates absorption and scattering cross-sections for a single sphere, then invokes available data on drop size distribution vs. rainfall rate and system parameters such as antenna beamwidth and propagation path-length to complete the calculation of attenuation in dB/km. These calculations involve so many meteorological variables that cannot be accurately modelled that they furnish only rough order-of-magnitude results that have limited value for prediction of actual performance in rain. However, there has been enough experimental corroboration of the main features of the theory to provide some confidence that the theory can at least provide rough guidelines for design. A satisfactory theory including multiple scattering (known to be important with dense raindrop distributions) is unavailable and a single scattering model must suffice for the present. Another limitation is inability to correlate attenuation with a specific number characterizing the level of rainfall. The best that can be attained along these lines is a rough statistical average.

Still another important source of SHF attenuation is water vapor absorption, which peaks at about 22GHz and is present to some extent throughout much of the SHF and mm bands. This effect is sensitive to atmospheric temperature and humidity.

Rain scattering, in addition to its role in attenuation, also produces phase shift, crosspolarization [Thomas, 1971] and interference between systems operating at the same frequency. All of those effects may be important in digital communication systems.

F. Millimeter Band (30-300GHz, λ = 1-10mm) - Terrestrial Propagation. The advantage of millimeter waves for digital transmission is the enormous bandwidth and hence high attainable data rates. The principal disadvantage, aside from equipment difficulties (e.g., small tolerances, inefficient antennas, etc.) is attenuation due to rain and molecular gases. This is the principal reason for development of guided transmission for this band. The latter will be omitted, discussion being limited to atmospheric propagation [Altshuler et al, 1968].

Communication system geometry is essentially equivalent to that for SHF (Section IV.E) except that structures are scaled down due to wavelength reduction.

Atmospheric turbulence and multipath problems are aggravated at millimeters because irregularities of mm scale appear as major scattering obstacles, being comparable to wavelength. Some terrain obstacles become larger relative to wavelength and hence totally opaque. In general, the transmission must be direct LOS. Analytical modelling of mm wave propagation invokes "quasi-optics" concepts, i.e., many effects follow standard physical optics arguments. The same holds to some degree at cm, but it becomes increasingly true as wavelengths decrease into and through the mm region.

The discussion of rain effects in Section IV.E applies to mm waves. As indicated there, attenuation, phase shift and interference due to rain scattering may be important at the high end of SHF. It is even more important at mm because raindrop dimensions are comparable to wavelength [Hogg, 1968; Sander, 1975; Wulfsberg and Altshuler, 1972; Ulabi and Straiton, 1970]. Atmospheric attenuation is serious at millimeters, even under dry conditions. There are a number of molecular absorption bands, the most important being oxygen absorption, which peaks at about 60GHz and depends on atmospheric temperature and pressure.

V. SPECIAL CHANNELS. Discussions in this section concern propagation aspects of two specific classes of digital communication systems, scatter and satellite-ground.

A. Troposcatter Systems. Propagation analysts have done considerable work on theories to explain "scatter propagation", the mechanism of troposcatter and ionoscatter communication systems [Barzilai, 1975; Wheelon, 1959]. Over-the-horizon transmissions at VHF, UHF and SHF believed attributable to "blobs", i.e., tropospheric or ionospheric regions of refractive index inhomogeneity, were first modelled analytically by Booker and Gordon [1950]. Subsequently others developed more elaborate scatter propagation theories. For most "engineering solutions", e.g., error rate calculations with digital modems, simple theory usually suffices. It was used by Bello [1969] in developing a model in which the ideas discussed in Section II are applied to tropo channels. Further work was done subsequently on tropo channel models and their use in evaluating tropo link performance [Daniel and Reinman, 1976; Engel, 1968; Gjessing and McCormick, 1974].

The classical model as used by Bello regards the atmosphere as a collection of "blobs" each scattering some transmitted signal energy into the receiver. Single scattering is assumed and the use of highly directive antennas, feasible at UHF and SHF, limits the scattering volume to the small region covered by both transmitter and receiver antenna beams. The assumed received signal is a superposition of signals scattered by each "blob". Individual blob signals are assumed mutually statistically independent and scattered power is proportional to the blob's volume and scattering cross-section. The latter depends on the mean square of the dielectric constant fluctuation and is roughly proportional to the mth power of the angle between incident and scattered waves, where m is a positive number near 5.

In Bello's model, central limit theorem arguments are used to show that the signal from the scatter volume is a complex Gaussian random process and hence its statistics can be determined from a knowledge of the "frequency correlation function" and the "delay power spectrum". The ACF (with respect to both time and frequency) of the "Time Variant Transfer Function" (Section II) is $R(\tau,\Omega) = \langle T^*(\omega,t)\, T(\omega + \Omega,\ t + \tau)\rangle$ (17). If the signal is a sample function of a complex Gaussian random process, its complete statistics can be determined from $R(\tau,\Omega)$. $R(0,\Omega)$ [Branham, 1970; Kennedy, 1972; Manders, 1970] is called the frequency correlation function. The latter's Fourier transform, called the delay power spectrum, is a measure of signal power distribution along the delay continuum.

For high data rate transmission on a tropochannel, fading is slow compared with data rate; hence Bello's channel model is assumed time-invariant during long message blocks and stationary complex Gaussian statistics are assumed. Consequently, either the delay power spectrum or the frequency correlation function provides all information needed to calculate error rates [Arnstein, 1971]. In addition to multipath and fading, additive atmospheric and extraterrestrial noise also cause errors in tropo systems (Sections III.A, III.B).

Simulators, mostly based on the tapped delay line model (Sections IV.C, IV.D) have been developed to model fading and multipath in tropo channels [Fitting, 1966] and have been used to evaluate error performance of actual tropo links.

Troposcatter systems are operated at frequencies between (about) 40MHz and 10GHz, covering much of VHF and UHF and part of SHF. Ionoscatter systems use lower frequencies (i.e. low VHF) where significant scattering from electron density inhomogeneities can occur. Theoretical considerations relating to channel statistics are roughly the same for troposcatter and ionoscatter systems. Although the latter are not discussed explicitly here, the general ideas discussed above carry over into ionoscatter channel models (with different parameter regimes, of course).

B. Satellite-Ground Communication. Theoretically, communication between satellites and points on or near the earth (e.g., ground station or aircraft) is feasible at any frequency that penetrates the ionosphere [Lawrence et al, 1964]. Ideally these transmissions should be at SHF or mm, where the RF is so high relative to the highest ionospheric plasma frequency that the ionosphere is nearly transparent. However, the higher the frequency within those bands, the greater the difficulties due to molecular absorption and rain in the lower atmosphere part of the propagation path [Gusler and Hogg, 1970; McDonald and Reber, 1973] (Sections IV.E, IV.F).

Satellite-ground transmissions at lower frequencies, e.g., UHF and VHF, although well above plasma frequencies, are affected by the ionosphere in various ways [Berger, 1976; Lawrence et al, 1964; Shepherd and Lomax, 1967; Ulaszcek et al, 1976]. The ionospheric electron density exhibits both smooth large-scale variations (with dimension of many wavelengths at VHF or above) and small-scale random fluctuations. A radio wave traversing the ionosphere undergoes amplitude scintillations and phase fluctuations due to these random inhomogeneities [Aarons, 1970; Deckelman and Ziemer, 1975; Pope, 1974; Rino, 1976; Umeki et al, 1977; Whitney and Basu, 1977]. Multipath effects due to widely separated scattering regions are also present and cause fading and intersymbol interference.

A change in electrical path length of a wave traversing the ionosphere occurs because the refractive

index differs from that of free space. This change is proportional to the integrated electron density along the path. Refractive bending occurs due to smooth gradients in electron density. Such gradients give rise to wave front tilting, proportional to the transverse (to the propagation path) gradient of integrated electron density. The ionosphere's dispersive nature introduces a group path delay, because the group velocity of a pulse of radio energy at frequency ω differs from the phase velocity at ω. This delay is proportional to integrated electron density. Collision-induced absorption also occurs. All of the above effects vary inversely as frequency-squared.

As the wave encounters random electron density variations, scattering occurs. This generates fluctuations in both received signal amplitude and apparent source position due to random phase shift. These effects, known respectively as amplitude scintillation and angular scintillation, both decrease with increasing frequency. Because phase shift varies with time, there is also a random fluctuation in observed frequency. These random phase and frequency fluctuations cause error-producing noise in PSK or FSK systems. In addition to the above effects there are the effects of magnetic-field induced ionospheric birefringence, namely the difference in radio path length between ordinary (O) and extraordinary (E) wave (due to a difference in refractive index and hence phase velocity) and the rotation of the polarization vector of the wave emerging from the ionosphere due to different polarizations of O and E waves. These two effects are inversely proportional to the cube and square of RF respectively.

All of the above-mentioned effects decrease with frequency and hence are most pronounced at VHF, diminish considerably at UHF and SHF and are negligible in the mm band [Brookner, 1969]. However, as remarked above, these higher frequencies present other problems such as rain attenuation (Sections IV.E, IV.F) and increases in extraterrestrial noise (Section III.B).

Bello [1973] has applied his channel modelling theories (Section II) to the "aeronautical channel", i.e., the link between an aircraft and a satellite. He develops his "Gaussian Wide-Sense Stationary-Uncorrelated Scatterers" (WSSUS) channel models based on propagation statistics of the "surface scatter channel", defined as the "collection of radio paths between a transmitter and receiver (one a satellite, the other an aircraft) existing due to the intervention of the earth's surface plus a distortion-free 'direct path' between transmitter and receiver". He does not discuss the direct path but limits the analysis to the transmitted wave reflected toward the receiver from the earth's surface (specifically the sea surface). He uses the results of random surface scattering theory (Section III.E) to derive the time-frequency correlation function of the channel $R(\Omega,\tau)$ (Section II), from which, given the assumed complex Gaussian WSSUS channel statistics, one could evaluate error rates for a digital modem operating on this channel.

REFERENCES

J. Aarons (January 1970), "A survey of scintillation data and its relationship to satellite communications", AFCRL, Hanscom Field, Massachusetts, Report AFCRL 70-0053.

L. G. Abraham (December 1966), "Reliability of microwave radio relay systems", IEEE Trans. Commun., COM-14, pp. 805-823.

Ya L. Al'pert (1963), "Radio wave propagation and ionosphere", (Translation from Russian), Consultants Bureau, New York.

E. E. Altshuler, V. J. Falcone and K. N. Wulfsberg (July 1968), "Atmospheric effects on propagation at millimeter wavelengths", IEEE Spectrum, 5, pp. 83-90.

D. Arnstein (April 1971), "Correlated error statistics on troposcatter channels", IEEE Trans. Commun., COM-19, 2, pp. 225-228.

G. A. Arredondo, W. H. Chriss and E. H. Walker (November 1973), "A multipath fading simulator for mobile radio", IEEE Trans. Commun., COM-21, 11, pp. 1325-1328.

A. P. Barsis and C. A. Samson (April 1976), "Performance estimation for 15GHz microwave links as a function of rain attenuation", IEEE Trans. Commun., COM-24, 2, pp. 462-470.

G. Barzilai (July 1975), "Research on statistical aspects of tropospheric propagation", Radio Science, 10, 7, pp. 745-752.

P. Beckmann and A. Spizzichino (1963), "The scattering of electromagnetic waves from rough surfaces", Pergamon, New York.

P. A. Bello (December 1963), "Characterization of time-variant linear channels", IEEE Trans. Comm. Sys., CS-11, pp. 360-393.

P. A. Bello (September 1965), "Error probabilities due to atmospheric noise and flat fading in HF ionospheric communication systems", IEEE Trans. Commun., COM-13, pp. 266-279.

P. A. Bello (August 1966), "Binary error probabilities over selectively fading channels containing specular components", IEEE Trans. Commun., COM-14, pp. 400-406.

P. A. Bello (April 1969), "A troposcatter channel model", IEEE Trans. Commun., COM-17, pp. 130-137.

P. A. Bello (May 1973), "Aeronautical channel characterization", IEEE Trans. Commun., COM-21, pp. 548-563.

L. C. Berger (September 1976), "Formulas for signal time delay through the ionosphere", IEEE Trans. Commun., COM-24, 9, pp. 1052-1054.

J. G. Bolton and K. C. Westfold (March 1950), "Galactic radiation at radio frequencies: 1-100mc/s survey", Aust. J. Sci. Res. A, 3, pp. 19-33.

H. G. Booker and W. E. Gordon (April 1950), "A theory of radio scattering in the troposphere", Proc. IRE, 38, pp. 401-412.

J. T. Boys (October 1968), "Statistical variations in the apparent spectral component of ionospherically reflected radio waves", Radio Science, 3, pp. 984-990.

R. A. Branham (September 1970), "Correlation bandwidth measurements over troposcatter paths", Proc. AGARD Conf. Paper 38.

E. E. Brookner (1969), "Characterization of millimeter wave earth-space links communication channels", Proc. 1969 IEEE Int. Conf. Commun., pp. 7.7-7.14.

L. Bullington (October 1947), "Radio propagation at frequencies above 30 megacycles", Proc. IRE, 35, pp. 1122-1136.

J. J. Bussgang, E. H. Getchell and B. Goldberg (October 1974), "VHF channel simulation", IEEE EASCON '74, pp. 562-564.

R. H. Clarke (July/August 1968), "A statistical theory of mobile-radio reception", Bell Sys. Tech. J., 47, pp. 957-1000.

S. F. Clifford and J. W. Strohbehn (March 1970), "The theory of microwave line-of-sight propagation through a turbulent atmosphere", IEEE Trans. Ant. Prop., AP-18, pp. 264-274.

D. C. Cox and R. P. Leck (September 1975), "Correlation bandwidth and delay spread multipath propagation statistics for 910MHz urban mobile radio channels", IEEE Trans. Commun., COM-23, 11, pp. 1271-1280.

R. K. Crane (March 1967), "Coherent pulse transmission through rain", IEEE Trans. Ant. Prop., AP-15, pp. 252-256.
R. K. Crane (September 1975), "Attenuation due to rain - mini-review", IEEE Trans. Ant. Prop., AP-23, 5, pp. 750-752.
L. D. Daniel and R. A. Reinman (June 1976), "Performance prediction for short-range troposcatter links", IEEE Trans. Commun., COM-24, 5, pp. 670-672.
W. F. Deckelman and R. E. Ziemer (April 1973), "Computer modeling of the statistical properties of trans-ionospheric scintillation channels", IEEE Trans. Commun., COM-23, 4, pp. 462-467.
D. Dence and T. Tamir (April 1969), "Radio loss of lateral waves in forest environments", Radio Science, 4, pp. 307-318.
F. H. Dickson, J. J. Egli, J. W. Herbstreit and G. S. Wickizer (August 1953), "Large reductions of VHF transmission loss and fading by the presence of a mountain obstacle in beyond-line-of-sight paths", Proc. IRE, 41, pp. 967-969.
R. T. Disney and A. O. Spaulding (February 1970), "Amplitude and time statistics of atmospheric and man-made radio noise", ESSA, Tech. Rep. ERLI50-ITS 98.
H. T. Dougherty and W. J. Hartman (April 1977), "Performance of a 400Mbit/s system over a line-of-sight path", IEEE Trans. Commun., COM-25, 4, pp. 427-432.
J. J. Egli (October 1957), "Radio propagation above 40mc over irregular terrain", Proc. IRE, 45, pp. 1383-1391.
J. S. Engel (June 1968), "The mathematical equivalence of digital troposcatter models", IEEE Trans. Commun., COM-16, 3, pp. 464-467.
R. Esposito and R. E. Buck (November 1973), "A mobile wide-band measurement system for urban man-made noise", IEEE Trans. Commun., COM-21, 11, pp. 1224-1232.
E. C. Field and M. Lewinstein (January 1978), "Amplitude probability distribution model for VLF/ELF atmospheric noise", IEEE Trans. Commun., COM-26, 1, pp. 83-87.
R. C. Fitting (1966), "Wideband troposcatter radio channel simulator", IEEE 1966 Int. Commun., Conf. Dig.
H. W. Früchtenicht (March 1974), "Notes on duct influences on line-of-sight propagation", IEEE Trans. Ant. Prop., AP-22, 2, pp. 295-302.
I. M. Fuks (1966), "Toward a theory of radio wave scattering at a rough sea surface", Izvestiya VUZ Radiofizika (USSR), 9, 5, pp. 876-887.
A. A. Giordano and F. Haber (November 1972), "Modeling of atmospheric noise", Radio Science, 7, 11, pp. 1011-1023.
D. T. Gjessing and K. S. McCormick (September 1974), "On the prediction of the characteristic parameters of long-distance tropospheric communication links", IEEE Trans. Commun., COM-22, 9, pp. 1325-1331.
S. L. Godard (July 1970), "Propagation of centimeter and millimeter wavelengths through precipitation", IEEE Trans. Ant. Prop., AP-18, pp. 530-534.
B. Goldberg (December 1966), "300KHz-30MHz MF/HF", IEEE Trans. Commun., COM-14, pp. 767-784.
B. Goldberg, Editor (1976), "Communications channels: Characterization and behavior", IEEE Press.
J. Goldhirsch (November 1975), "Prediction methods for rain attenuation statistics at variable path angles and carrier frequencies between 13 and 100GHz", IEEE Trans. Ant. Prop., AP-23, 6, pp. 786-791.
L. T. Gusler and D. C. Hogg (January 1970), "Some calculations on coupling between satellite-communications and terrestrial radio relay systems due to scattering by rain", Bell Sys. Tech. J., 49, pp. 1491-1511.
L. T. Hearson (August 1967), "Unusual propagation factors in point-to-point microwave system performance", IEEE Trans. Commun., COM-15, pp. 615-625.
J. W. Herbstreit and M. C. Thompson (July 1956), "Measurements of the phase of signals received over trans-mission paths with electrical lengths varying as a result of atmospheric turbulence", IRE Trans. Ant. Prop., AP-4, pp. 352-358.
J. R. Herman (1971), "Survey of man-made radio noise), in "Progress in Radio Science 1966-1969", 1, C. M. Ninnis, Ed. URSI.
D. G. Hogg (January 5, 1968), "Millimeter wave communication through the atmosphere", Science, 159, pp. 39-46.
S. T. Hong, I. Sreenivasiah and A. Ishimaru (November 1977), "Plane wave pulse propagation through random media", IEEE Trans. Ant. Prop., AP-25, 6, pp. 822-828.
F. Ikegami (July 1959), "Influence of an atmospheric duct on microwave fading", IRE Trans. Ant. Prop., AP-7, pp. 252-257.
D. J. Kennedy (April 1972), "A comparison of measured and calculated frequency correlation functions over 4.6 and 7.6GHz troposcatter paths", IEEE Trans. Commun., COM-20, 2, pp. 173-178.
S. Kozono and Kunio Watanabe (October 1977), "Influence of environmental buildings on UHF land mobile radio propagation", IEEE Trans. Commun., COM-25, 10, pp. 1133-1143.
R. S. Lawrence, C. G. Little and H. J. A. Chivers (January 1964), "A survey of ionospheric effects upon earth-space radio propagation", Proc. IEEE, 52, pp. 4-27.
C. H. Liu and K. C. Yeh (December 1975), "Frequency and spatial correlation functions in a fading communi-cation channel through the ionosphere", Radio Science, 10, 12, pp. 1055-1061.
A. M. Manders (September 1970), "Frequency correlation function for troposcatter circuits", Proc. AGARD Conf. Paper 37.
O. V. McDonald and E. E. Reber (June 1973), "Rainfall and space diversity for millimeter wave earth-satellite communications systems", J. Appl. Meteorology, 12, pp. 709-715.
M. Nesenbergs (December 1967), "Error probability for multipath fading - the 'slow' and 'flat' idealiz-ation", IEEE Trans. Commun., COM-15, 6, pp. 797-805.
J. K. Omura and P. D. Shaft (October 1971), "Modem performance in VLF atmospheric noise", IEEE Trans. Commun., COM-19, 5, pp. 659-668.
R. H. Ott (March 1977), "Temporal radio frequency spectra of multifrequency waves in a turbulent atmosphere characterized by a complex refractive index", IEEE Trans. Ant. Prop., AP-25, 2, pp. 254-260.
J. H. Painter, S. C. Gupta and L. R. Wilson (May 1973), "Multipath modeling for aeronautical communi-cations", IEEE Trans. Commun., COM-21, pp. 658-662.
L. W. Pickering (May 1975), "The calculation of ionospheric Doppler spread on HF communication channels", IEEE Trans. Commun., COM-23, 5, pp. 526-537.
J. H. Pope (July 1974), "High latitude ionospheric irregularity model", Radio Science, 9, 7, pp. 675-682.
Radio Science (January 1975), Special Issue, Waves in Random Media.
K. Rawer (July 1975), "The historical development of forecasting methods for ionospheric propagation of HF waves", (Review), Radio Science, 10, 7, pp. 669-679.
D. Reudink and M. F. Wazowicz (November 1973), "Some propagation experiments relating foliage loss and dif-fraction loss at X-band and UHF frequencies", IEEE Trans. Commun., COM-21, pp. 1198-1206.

S. O. Rice (1951), "Reflection of electromagnetic waves from slightly rough surfaces", Commun. Pure and Appl. Math., No. 2/3, pp. 351-378.
C. L. Rino (November 1976), "Ionospheric scintillation theory - a mini review", IEEE Trans. Ant. Prop., AP-24, 6, pp. 912-915.
R. R. Rogers (July 1976), "Statistical rainstorm models: their theoretical and physical foundations", (Tutorial) IEEE Trans. Ant. Prop., AP-24, 4, pp. 547-566.
C. L. Ruthroff (September 1971), "Multiple-path fading on line-of-sight microwave radio systems as a function of path length and frequency", Bell Sys. Tech. J., 50, pp. 2375-2397.
J. Sander (March 1975), "Rain attenuation of millimeter waves at λ = 5.77, 3.3 and 2mm", IEEE Trans. Ant. Prop., AP-23, 2, pp. 213-220.
H. N. Shaver, B. C. Tupper and J. B. Lomax (February 1967), "Evaluation of a Gaussian HF channel model", IEEE Trans. Commun., COM-15, pp. 74-88.
R. A. Shepherd and J. B. Lomax (April 1967), "Frequency spread in ionospheric radio propagation", IEEE Trans. Commun., COM-15, 2, pp. 268-275.
A. G. Smith (April 1960), "Extraterrestrial noise as a factor in space communications", Proc. IRE, 48, pp. 593-599.
T. Tamir (November 1967), "On radio wave propagation in forest environments", IEEE Trans. Ant. Prop., AP-15, pp. 806-817.
I. Tatarskii (1961), "Wave propagation in a turbulent medium", New York, McGraw-Hill.
D. T. Thomas (October 1971), "Cross-polarization distortion in microwave radio transmission due to rain", Radio Science, 6, pp. 833-839.
M. C. Thompson, L. E. Wood, H. B. Janes and O. Smith (November 1975), "Phase and amplitude scintillations in the 10 to 40GHz band", IEEE Trans. Ant. Prop., AP-23, 6, pp. 792-797.
J. R. Tucker (1972), "Error behavior of VHF channels", IEEE Int. Conf. Commun., pp. 15-25 to 15-30.
F. T. Ulabi and A. W. Straiton (July 1970), "Atmospheric absorption of radio waves between 150 and 350GHz", IEEE Trans. Ant. Prop., AP-18, pp. 479-485.
S. J. Ulaszek, C. H. Liu and K. C. Yeh (October 1976), "A study of signal decorrelation through the ionosphere", IEEE Trans. Commun., COM-2, pp. 1191-1195.
R. Umeki, C. H. Liu and K. C. Yeh (March-April 1977), "Multifrequency studies of ionospheric scintillations", Radio Science, 12, 2, pp. 311-317.
W. C. Vergera, J. L. Levatich and T. J. Carroll (September 1962), "VHF air-ground propagation far beyond the horizon and tropospheric stability", IRE Trans. Ant. Prop., AP-10, pp. 608-621.
J. R. Wait (June 1957), "The mode theory of VLF ionospheric propagation for finite ground conductivity", Proc. IRE, 45, pp. 760-767.
J. R. Wait and L. C. Walters (January 1963), "Curves for ground wave propagation over mixed land and sea paths", IEEE Trans. Ant. Prop., AP-11, pp. 38-45.
W. F. Walker (1966), "A simple baseband fading multipath channel simulator", Radio Science, 1, 7, pp. 763-767.
A. D. Watt, R. M. Cook, E. L. Maxwell and R. W. Plush (December 1958), "Performance of some radio systems in the presence of thermal and atmospheric noise", Proc. IRE, 46, pp. 1914-1923.
A. D. Watt and E. L. Maxwell (June 1957), "Characteristics of atmospheric noise from 1 to 100kc", Proc. IRE, 45, pp. 787-794.
C. C. Watterson, J. P. Juroshek and W. D. Bensema (December 1970), "Experimental confirmation of an HF channel model", IEEE Trans. Commun., COM-18, pp. 792-803.
A. D. Wheelon (1959), "Radio wave scattering by tropospheric irregularities", J. Res. NBS (Radio Prop.), 63D, pp. 205-233.
R. F. White (February 1968), "Space diversity on line-of-sight microwave systems", IEEE Trans. Commun., COM-16, 1, pp. 119-133.
H. E. Whitney and S. Basu (January/February 1977), "The effect of ionospheric scintillation on VHF/UHF satellite communications", Radio Science, 12, 1, pp. 123-133.
K. N. Wulfsburg and E. E. Altshuler (March 1972), "Rain attenuation at 15 and 35GHz", IEEE Trans. Ant. Prop., AP-20, pp. 181-187.
J. G. Zornig and J. F. McDonald (March 1975), "Experimental measurement of the second-order interfrequency correlation function of the random surface scatter channel", IEEE Trans. Commun., COM-23, 3, pp. 341-347.

DISCUSSION

R.M.Harris, UK

In the course of the literature survey has the author uncovered any documented work on the interaction between fast fading and radio receivers' automatic gain control characteristics?

Author's Reply

I have no recollection of this issue from my literature search.

R.M.Harris, UK

The author referred to analytical models for the atmospheric noise; the CCIR have published (Report 322) statistics of atmospheric noise related to sea level observations. Has the author come across any theoretical or experimental work relating sea level values of atmospheric noise field strength to values which obtain at greater altitudes?

Author's Reply

You may possibly find a clue in one of the references I cited on atmospheric noise models, e.g., Giordano and Haber, 1972; Field and Lewinstein, 1978, *or* in one or more of my references cited on atmospheric noise in general, e.g. Watt et al. 1958, Watt and Maxwell June 1957, Disney and Spaulding 1970. A crude idea may be obtained by thinking of a lightning stroke as a vertical dipole antenna and using the vertical pattern function for a vertical dipole to determine the angular variation of field strength and thereby the vertical distance variation.

E.Ante, Ge

Concerning atmospheric noise:

Can you give any values (or reference papers) about field-strength, duration and frequency distribution (up to 100 MHz) of lightning flashes during thunderstorms versus distance from origin (up to e.g. 20 to 50 km).

Author's Reply

From my written paper reference list: Watt, et al., December 1958; Disney and Spaulding, February 1970; Watt and Maxwell, June 1957.

Also: Watt and Maxwell, Proc. IRE, Vol.45, pp.55-62, Jaunary 1957; Crichlow, Smith, Morton and Corlis (August 1955), "World wide radio noise levels expected in the frequency band from 10 kc to 100 mc". NBS Circular 557; E.T.Pierce, Radio Science, Vol.4, No.7, pp.661-666, July 1969.

Very crude analysis: d = Distance from origin, λ = wavelength; at VLF (3–30 KHz) varies roughly as $\sqrt{\lambda/d}$; at LF (30–300 KHz) or MF (300 KHz–3 MHz) varies roughly as (λ/d).

PERFORMANCE PREDICTIONS AND TRIALS OF A HELICOPTER UHF DATA LINK

by

R. M. Harris

Procurement Executive, Ministry of Defence
Royal Aircraft Establishment, Radio and Navigation Department
Farnborough, Hampshire GU14 6TD, Great Britain

SUMMARY

A 16 kilo bauds digital communication system has been proposed for use on helicopter to helicopter, and on helicopter to surface radio links over the sea. Operating in the UHF band it is expected that line-of-sight ranges should be achieved and studies have been undertaken to estimate the bit error rate performance.

Initial predictions of system performance were based on known characteristics of some helicopter-installed systems and classical treatment of multipath radio wave propagation over a smooth sea. Certain parameters have been measured by means of airborne experiments in order to improve the predictions. Two helicopters and a shore station were specially instrumented and an experimental 16 kilo bauds data link was also operated.

The probability of achieving a particular grade of service is considered as the arithmetical product of several independent statistical parameters. Some of these are related to the temporal and spatial variability of radio wave propagation; others reflect the almost random orientation of the aircraft under operational conditions affecting the signal received via the aircraft's radiation pattern.

A method of statistical analysis based on approximate Gaussian frequency distribution functions has been applied to the experimental data. Performance predictions have been presented in terms of Time Availability for selected bit error rates.

1 INTRODUCTION

This paper reviews a combination of theoretical and experimental studies conducted in order to estimate the bit error rate performance of a digital communication system. The system is proposed for the transmission of binary digital data over dedicated simplex UHF radio links between helicopters, and between helicopters and ships. The signalling rate is 16 kilo bauds and conventional voice communications equipment is to be used, with 50 kHz channel spacing. For the purposes of this paper amplitude shift keying is considered. The data link is to be used at sea for approximately line-of-sight UHF radio communications ranges, the helicopters operating between 300 and 1200 metres above the sea.

For the purposes of either theoretical modelling or flying trials it has been convenient to rationalize the system as follows. The UHF band (225-400 MHz) has been represented by two (trials) frequencies 237.3 MHz and 386.1 MHz. Four cases have been considered: air to surface, at altitudes of 1000 ft (304.8 m) and 4000 ft (1219.2 m), and air to air, at 1000 ft and 4000 ft (both aircraft).

A project of this kind can be pursued along three or four different lines.

(1) A purely theoretical approach would comprise classical radio wave propagation modelling supported by documented characteristics of aircraft and ship antenna systems. Laboratory measurements on radio equipment would establish relationships between bit error rate (BER) and signal strength, leading to complete system BER predictions.

(2) In contrast to (1) an experimental system can be built and tried using many hours of reasonably unstructured flying to simulate operational conditions. Statistical analysis of such results may prove to be of general application if a sufficient amount of trials data can be collected.

Falling between these two extreme approaches come two hybrid procedures which employ planned airborne experiments.

(3) One procedure uses airborne measurements to establish the propagation characteristics of the system and also provides quantitative information about received signal strengths under controlled conditions. If the BER dependence on signal strength is known then the equivalent BER performance under the same conditions may be determined.

(4) The other procedure is similar except that an experimental data link would be flown under controlled conditions in addition to the propagation measurements. In both (3) and (4) a mathematical system model is constructed and used as the basis for statistical predictions of performance.

In this case, initial predictions were based on the theoretical approach (1) but a new approach was found necessary owing to the lack of certain information about practical systems. Two RAE helicopters, a Wessex and a Sea King, were specially instrumented for the purpose of measuring and recording radio signal strengths. A cliff-top shore station was also equipped so as to simulate a typical ship terminal. The plan was to fly a programme of propagation measurements to be followed by experimental data link trials under similar flying conditions. However, the second phase was not a complete success and so procedure (3), rather than (4) was adopted.

Copyright ©, Controller HMSO London, 1978

The essential objectives of the experimental flying were as follows:

(i) Verification of the theoretical propagation predictions, including the magnitude of multi-path interference fading.

(ii) Measurement of the practical values of certain parameters leading to quantitatively accurate predictions of received signal strength.

(iii) Characterization of the real channel noise which governs the relationship between BER and signal strength.

(iv) Comparison of 'predicted' BER performance with the experimental data link results.

The usefulness of the experimental results was greatly enhanced by the close agreement with theoretical propagation predictions. The subsequent statistical treatment was thereby solidly based on this foundation.

The designer of any data link system requires to know as much as possible about the frequency of occurrence of different short term bit error rates. From this it is possible to obtain the probability distribution of various levels of performance which are determined by the probability of bit errors. The percentage of time for which a given level of performance is exceeded, known as the Time Availability, Q%, is closely related to the probability distribution of short term bit error rates (CCIR Report 322). In this paper it is assumed that the time scale of any fluctuations in the level of performance is less important than the aggregate time spent at a stipulated level. The predictions of Time Availability will be related to the contingent variations in the strength of received signals with the possible exclusion of very fast fading (scintillation).

This requires the mathematical convolution of several, independent frequency distribution functions belonging to respective transmission factors of the system. A method of simplifying this step has been developed using Gaussian functions to approximate the individual frequency distributions. This method, although applied here to experimental results, may be alternatively applied to theoretical system modelling.

2 TRANSMISSION FACTORS

For a properly designed communication system the bit error rate (BER) or instantaneous probability of error, is determined by the signal to noise ratio at the instant of detection. Different forms of digital encoding and carrier modulation have different quantitative relationships between BER and signal to noise ratio. Analytical functions have been derived for the particular case of Gaussian (random) noise but for other types of noise only empirical relationships exist. For the case of binary digital signalling in the presence of Gaussian noise, the relationship has been plotted at Fig 1 (Carlson 1968).

For any communication channel there is a theoretical minimum to the inevitable thermal electrical noise power p, given by

$$p = kTB \quad \text{watts} \tag{1}$$

where T is the absolute temperature, kelvin
B is the channel bandwidth, Hz
k is Boltzmann's constant, 1.38×10^{-23} joules per kelvin.

An ideal radio receiver amplifies the aerial signal and the thermal noise equally so that the same signal to noise ratio present at the input is also present at the receiver's demodulator. Practical radio receivers introduce extra noise which originates in the active devices especially near the front end and the frequency translator circuits. This added noise may be equated to a ficticious noise source acting at the input to an ideal receiver (Bleaney 1965). The ratio of the equivalent input noise to the thermal noise is called the Noise Factor of the receiver. If, after demodulation, the noise has a single-sided spectral power density of N_0' then the best signal to noise ratio for any system is $2E_b/N_0'$ where E_b is the signal energy per 'bit'

$$E_b = S/K \text{ joules} \tag{2}$$

where S is the demodulated signal (average) power
and K is the signalling rate, bauds.

Knowledge of the Noise Factor, demodulator characteristics and statistical behaviour of the noise enables the BER dependence on signal strength to be estimated. Alternatively, BER may be measured as a function of signal strength for a particular receiver, using digital test transmissions. Usually, the noise power produced at the receiver output is substantially constant, so for given BER specification there exists a corresponding signal strength threshold.

The radio transmission system may be considered as a chain of elements each of which contributes to the attenuation of the signal on its passage from sender to receiver. The transmitter is assumed to produce a constant signal power to its antenna. The gain of the receiver is immaterial; the signal to noise ratio at the receiving antenna terminals and Noise Factor alone govern the BER performance. The strength of the electromagnetic wave radiated from the sending antenna depends on its power efficiency and gain in any particular direction in space. For most applications the azimuthal radiation pattern (ARP) is of interest. Similarly at the receiver the strength of the signal produced at the antenna terminals depends on the efficiency and gain in any particular direction. The ratio of the electromagnetic wave

field energy at the receiving antenna to the wave field energy near the transmitter is referred to as the propagation factor (<1) whilst the end to end received signal to transmitted signal power ratio is referred to as the transmission factor.

The propagation factor, usually the most important single factor in a system, is usually considered in isolation from antenna gains and efficiencies etc. It is convenient to compute electromagnetic wave field strengths set up by a standard source of unit power radiated from an ideal half-wave dipole. The computations are usually complex and real systems are usually related to standard propagation curves by applying a single scaling factor. This factor includes transmitter power, feeder losses, antenna losses, polarization mismatches and multiplexing losses.

The propagation characteristic may be a complex (*eg* oscillatory) function of distance or height, as in the case of multipath propagation between aircraft. In this case the characteristic can be resolved into a slow variation with distance (or height) on which is superimposed a rapidly varying, oscillatory function. Changing the distance between sender and receiver results in a fluctuating receiver signal strength. The signal can be regarded, statistically, as a median level with a quasi-random fluctuation for which there exists a particular frequency distribution function. The median level can be considered as a simpler function, changing relatively slowly with distance or height. The sender or receiver or both may be mobile units and so in addition to relative motion between them, their antenna radiation patterns may be scanned as rotational motion takes place.

Expressed in decibels, the steady part of the received signal strength is given by the algebraic sum of the standard propagation factor, the scaling factor and the median gain of each of the antennas used. The fluctuating part is also given by the algebraic sum of the instantaneous departures from the median of the propagation factor and each of the respective antenna gains. The fluctuating part of the signal is thus a very complex function of up to three independent variables; time dependent, or temporal, fading may add yet a fourth variable. Thus the only feasible representation of received signal strengths and hence of BER predictions, is a statistical one. The reasonable assumption is made that the instantaneous orientation of each helicopter and the position in the complex radio field strength structure are independent and randomly chosen variables. Within defined limits, any parameter may take on any value with equal probability.

3 GAUSSIAN FREQUENCY DISTRIBUTION FUNCTIONS

Let the overall power gain be represented by U dB, where

$$U = R(x_1) + S(x_2) + T(x_3) \quad \text{dB} \tag{3}$$

$R(x_1)$ represents the transmitter antenna gain as a function of its orientation, x_1, relative to the receiver.

$S(x_2)$ represents the receiver antenna gain as a function of its orientation, x_2, relative to the transmitter.

$T(x_3)$ represents the radio wave propagation factor as a function of the relative position (usually separation distance), x_3, of the sender and receiver.

The variables x_1 and x_2 are usually azimuth angles and their range is 0-360 degrees.
The variable x_3 is usually constrained to lie within a relatively small distance increment, just wide enough to scan the short-range variability of the propagation characteristic. Thus for a statistical analysis of the system at an operating distance, d,

$$d - \delta < x_3 < d + \delta \tag{4}$$

where $\delta \ll d$. This constraint may be relaxed if the slowly varying part of the propagation characteristic is relatively small, in comparison to the fluctuating part, over a wider range of distances.

Each of the functions, R, S and T has a median value denoted by R_m, S_m and $T_m(d)$ respectively. $T_m(d)$ is the median value of the propagation factor in the incremental region from $(d - \delta)$ to $(d + \delta)$. The departures from the median values, y_1, y_2 and y_3, are such that

$$R(x_1) = R_m + y_1(x_1) \tag{5}$$

$$S(x_2) = S_m + y_2(x_2) \tag{6}$$

$$T(x_3) = T_m(d) + y_3(x_3) \quad . \tag{7}$$

The y's are assumed to possess arbitrary frequency distributions, so that the probability of a random choice of x_i resulting in the functional value, y_i, is given by:

$$p(y_i) = f_i(y_i) \quad . \tag{8}$$

The probability of y_i having any value in its range, y_{min} to y_{max}, must be identically equal to unity.

Therefore
$$\int_{y_i(min)}^{y_i(max)} p(y_i)dy_i = 1 \tag{9}$$

therefore
$$\int_{y_i(min)}^{y_i(max)} f_i(y_i)dy_i = 1 \ . \tag{10}$$

The overall transmission factor U, can also be written as

$$U(d) = R_m + S_m + T_m(d) + z(x_1,x_2,x_3) \tag{11}$$

where
$$z(x_1,x_2,x_3) = y_1(x_1) + y_2(x_2) + y_3(x_3) \ . \tag{12}$$

The frequency distribution function for z is given by the double convolution integral,

$$f(z) = \int_{-\infty}^{\infty} dv \int_{-\infty}^{\infty} f_1(x_1)f_2(v - x_1)dx_1 \left[f_3(z - v)\right] \tag{13}$$

(where v is an intermediate parameter only). If, as invariably in practice, the individual y's are bounded, then their frequency distribution functions will vanish at

$$f_i(y_i) = 0 \quad \text{for} \quad y_i > y_i(max) \tag{14}$$

$$f_i(y_i) = 0 \quad \text{for} \quad y_i < y_i(min) \ . \tag{15}$$

For simplicity, let

$$y_i(min) = -(y_i(max)) = q_i \ . \tag{16}$$

If the frequency distribution functions in equation (13) are each truncated at $\pm q_1$, $\pm q_2$ and $\pm q_3$ respectively, as defined by equations (14) to (16), then the resulting frequency distribution function, $f(z)$ is also bounded.

Thus,

$$f(z) = 0 \quad \text{for} \quad z > z(max) \tag{17}$$

$$f(z) = 0 \quad \text{for} \quad z < z(min) \ . \tag{18}$$

It can be shown that

$$z(max) = q_1 + q_2 + q_3 \tag{19}$$

and

$$z(min) = -(q_1 + q_2 + q_3) \ . \tag{20}$$

If the original truncation results in a discontinuity in the function f_i at $\pm q_i$ then the effect is smoothed out by the convolution process; the effect on $f(z)$ is a smooth, tapering approach to zero at $\pm z(max)$, and moreover

$$\frac{df(z)}{dz} = 0 \quad \text{at} \quad \pm z(max) \ . \tag{21}$$

At $z(min)$, all possible values of z lie above it and hence there is a 100% probability that

$$U > R_m + S_m + T_m(d) + z(min) = U_{min} \ . \tag{22}$$

A system performance having a threshold signal strength, V_{100}, determined by U_{min} would therefore possess a Time Availability of 100%. A system performance having a threshold signal strength of V_0, determined by U_{max} would have a zero Time Availability, where

$$U < U_{max} = R_m + S_m + T_m(d) + z(max) . \tag{23}$$

The Time Availability, Q% , of a system with a corresponding signal strength threshold of V_Q, where $V_{100} < V_a < V_0$, is given by:

$$Q = 100 \int_{z(Q)}^{z(max)} f(z)dz \quad \% \tag{24}$$

where

$$z(Q) = z(min) + V_Q - V_{100} . \tag{25}$$

If V_m is the median signal strength (Q = 50%) then equations (24) and (25) can be written as:

$$Q = 50 - \left(100 \int_{0}^{z(Q)} f(z)dz\right) \quad \% \tag{26}$$

where

$$z(Q) = V_Q - V_m . \tag{27}$$

In general the convolution integral, equation (13), is intractable since the constituent frequency distributions are not usually analytical functions. A special result may be noted in the case where two or more Gaussian functions are convoluted together. If

$$f_1(x_1) = \sigma_1^{-2} \exp\left(-\frac{x_1^2}{\sigma_1^2}\right) \tag{28}$$

and

$$f_2(x_2) = \sigma_2^{-2} \exp\left(-\frac{x_2^2}{\sigma_2^2}\right) \tag{29}$$

then the convolution integral,

$$f(z') = \int_{-\infty}^{\infty} f_1(x_1) f_2(z' - x_1)dx \tag{30}$$

becomes

$$f(z') = \sigma_0^{-2} \exp -\left(\frac{z'}{\sigma_0}\right)^2 \tag{31}$$

where

$$\sigma_0^2 = \sigma_1^2 + \sigma_2^2 . \tag{32}$$

If the naturally occurring frequency distribution functions can be approximated to suitably chosen Gaussian functions the solution to equation (13) can be written down straight away. If variances, σ_1, σ_2 and σ_3, characterize the equivalent Gaussian frequency distribution functions in equation (13) then

$$f(z) = \sigma_T^{-2} \exp\left(-\frac{z^2}{\sigma_T^2}\right) \tag{33}$$

where

$$\sigma_T^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 . \tag{34}$$

In general

$$\sigma_T^2 = \sum_i^n \sigma_i^2 \tag{35}$$

where there are n independent variables.

The naturally occurring frequency distribution functions may be turned into cumulative distribution functions, or they may be obtained directly as such, and are plotted on arithmetic probability graph paper. The best straight line through the points is the cumulative plot of the corresponding Gaussian approximation. From the slope of the line the variance of the corresponding Gaussian frequency distribution function may be calculated. The truncation points of the Gaussian cumulative function may be determined by the most judicious 'fit' to the natural curve.

The truncation of the final (convoluted) Gaussian function is performed, as a separate operation, at $\pm z(\max)$

where

$$z(\max) = \sum_i^n q_i \quad . \tag{36}$$

The value of the Gaussian function (equation (28)) is determined by the parameter $r_i = \left(x_i^2/\sigma_i^2\right)$ and at the truncation points, for each contributory function:

$$r_{iT} = (q_i/\sigma_i)^2 \quad . \tag{37}$$

For the final function,

$$r_{TT} = \left(\sum_i^n q_i\right)^2 \Big/ \sum \sigma_i^2 \quad . \tag{38}$$

Clearly $|r_{TT}| > |r_{iT}|$ always. This means that the value of $f(z)$ at the truncation points is always smaller than any of the $f_i(x_i)$ at the corresponding truncation points, $\pm q_i$, since the Gaussian is a decreasing function of $|r_i|$. Therefore, the convolution process helps to reduce the magnitude of the errors of approximation associated with truncation.

The application of the Gaussian approximation might entail a slight adjustment in the median values taken for $R(x_1)$, $S(x_2)$ and $T(x_3)$. These can be added together to give the median value for the overall transmission factor, as in equation (3). The variance, σ_T, for the overall frequency distribution function is obtained from equation (34) or equation (35). Thereby, a new cumulative probability distribution function can be drawn on arithmetic probability graph paper and the truncation points inserted, as given by equation (36). Time Availability can then be read directly from the scales as a function of either z, or U.

If the median value ($U_m(d)$) is, itself, a function of some other parameter such as distance (see equation (7)) then an alternative method of presenting the results is called for. For a given specified threshold signal strength the corresponding value of U will comprise a median value, U_m, and a departure, z dB. As the median value, U_m, falls with increasing distance, equation (11) shows that z must increase by the same amount. As z increases (equation (24)) the Time Availability Q, decreases until z equals $z(\max)$, when Q becomes zero. Thus for a particular threshold signal level, a graph may be plotted showing the dependence of Q on distance. This distance at which Q has fallen to the lowest acceptable value for the particular system is defined to be the range of the system.

If Q is plotted against distance it is advantageous to use an arithmetic probability scale for Q. Linear displacement parrallel to the Q-axis then corresponds to z directly and hence decibels. Thus the Q-axis is also linear in decibels which facilitates the mapping of the propagation characteristic onto this form of presentation (see Figs 16 and 17). At the distances where z approaches $z(\min)$ and $z(\max)$ the curve runs parallel to the Q-axis, as Q goes immediately to 100% and 0% respectively.

4 THEORETICAL AND EXPERIMENTAL MODELS

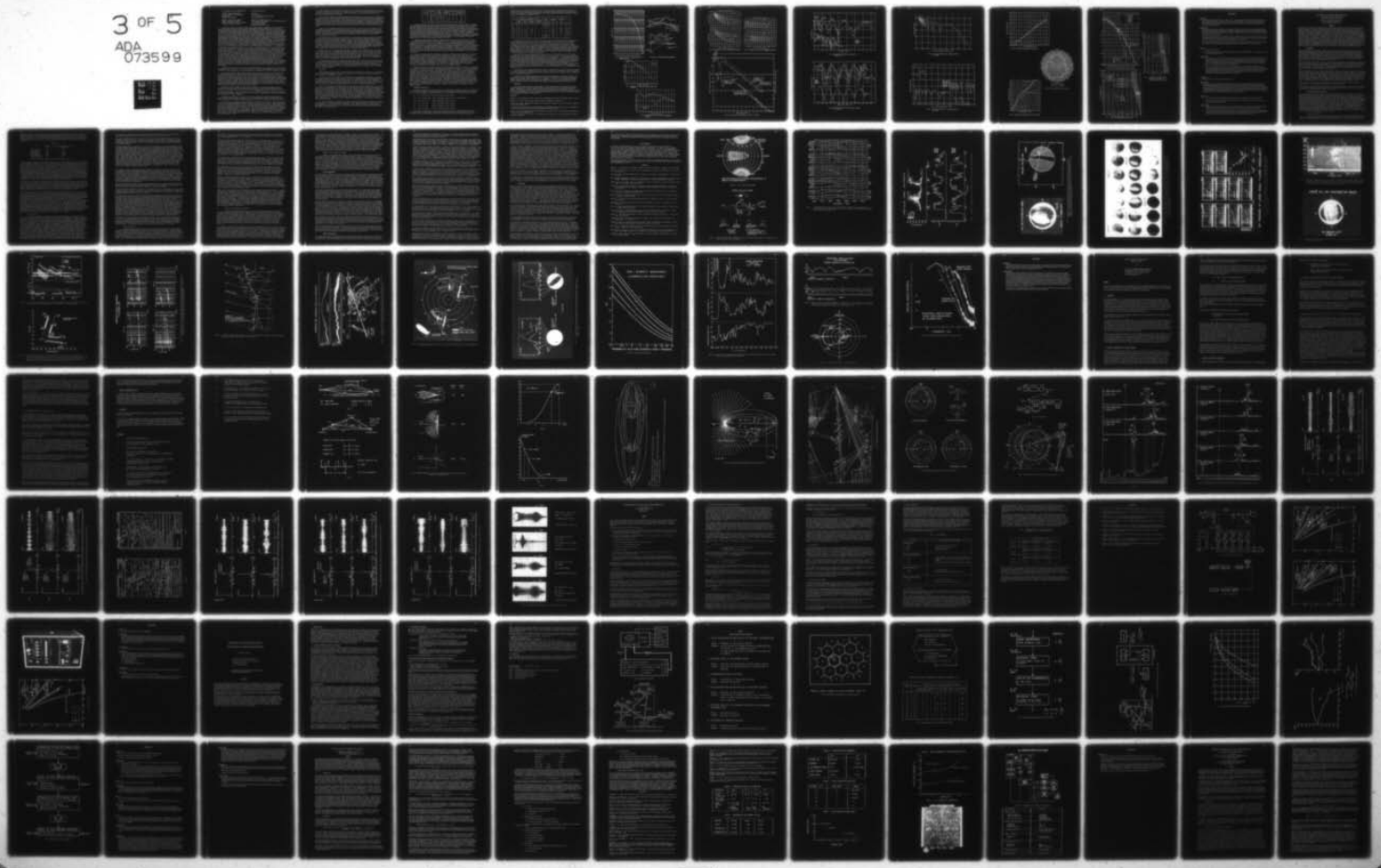
The theoretical predictions of radio wave propagation were based on a model system with the following specification across the UHF band:

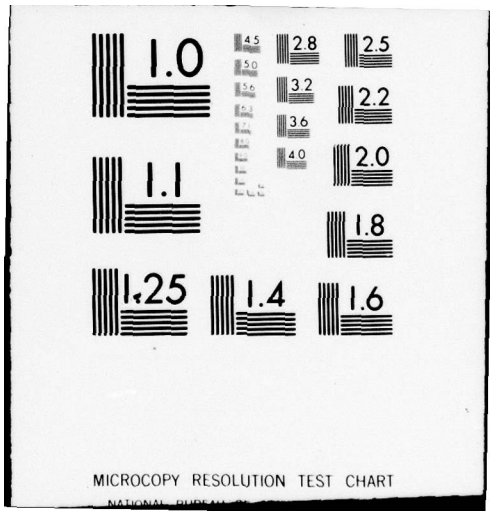
Transmitter power (CW)	10 W (all units)
Antenna gain	1.64 (2.15 dB) for half-wave dipole
Antenna feeder loss and efficiency	0.5 (6 dB loss)
Antenna polarization	vertical.

AD-A073 599 ADVISORY GROUP FOR AEROSPACE RESEARCH AND DEVELOPMENT--ETC F/G 17/2
DIGITAL COMMUNICATIONS IN AVIONICS.(U)
JUN 79 H LUEG

UNCLASSIFIED AGARD-CP-239 NL

3 OF 5
ADA 073599




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART

Electromagnetic wave propagation mode	vertically polarized.
Receiver antenna polarization	vertical
Antenna feeder loss and efficiency	0.5 (6 dB loss)
Antenna gain	1.64 (2.15 dB) for half-wave dipole
Matched line impedance	73 ohms (resistive).
Height of surface unit antenna	90 ft (27.4 m) above sea level
Height of aircraft antennas	1000 ft (304.8 m) and 4000 ft (1219.2 m) above sea level
Carrier frequencies (for calculations)	237.3 MHz and 386.1 MHz
Sea water electrical constants	conductivity = 4 $S\ m^{-1}$ relative permittivity = 80.

For both the air to air and air to surface radio links multipath propagation obtains by means of a direct wave and an indirect wave which is specularly reflected from the sea, see Fig 2. The phase relationship between the direct and reflected waves depends on the phase coefficient of reflection and on the geometrical path length difference between the two propagation paths. At very low angles of grazing (ψ) the reflection coefficient approaches unity with a 180° phase lag; the two waves then almost cancel out. For a wave whose electric field vector is horizontal (horizontally polarized) the reflection phase coefficient, $\phi^o(\psi)$, remains close to 180° and the magnitude coefficient, $R(\psi)$, falls slightly below unity for higher grazing angles (ψ^o) over the sea. For a vertically polarized wave the phase and magnitude coefficients (R, ϕ^o) both change over wider limits with increasing grazing angle. The geometrical path length difference also depends on grazing angle; as it increases so the phase delay, θ^o, of the reflected wave increases. The size of the effect is proportional to the product of the heights of the source and receiver above the sea. The electric field resulting from the interference of these two waves is determined by the vector sum and is a sensitive function of the relative phase delay, given by $\Phi = \theta + \phi$. As ψ increases from zero, θ increases whilst ϕ decreases slowly, becoming 90° at the pseudo-Brewster angle (ψ_B). For UHF propagation over sea water, the pseudo-Brewster angle is about 3°. The increase in θ nearly always outstrips the decrease in ϕ so that Φ increases positively from its initial value close to 180°. The vector resultant for the multipath interference increases as Φ increases above 180° and hence with increasing ψ. Thus at constant distance, field strength increases with altitude, almost in direct proportion to ψ, as long as Φ is less than 360°. This is referred to as the height gain effect. For a constant height, motion towards the source also increases ψ and hence Φ but now in a non-linear manner. The field strength increase is now governed by two factors, the inverse law of distance for each individual wave, and the interference factor. When the value of Φ reaches 360° the two waves interfere constructively and the resulting field strength is nearly double that of either wave (6 dB increase). This is designated the first interference maximum; further motion towards the source, or increase in altitude, results in successive destructive and constructive interference cycles. Fig 2 depicts the loci of the interference maxima and minima; they appear to diverge radially from a point midway between the source and its image in the plane of reflection. The angular frequency of the interference pattern is proportional to the height of the source. For horizontal motion through the pattern, for a fixed height of the source, the horizontal spatial frequency increases both towards the source and also with altitude.

The depth of the interference minima (or fades) depends entirely on the magnitude of the reflection coefficient, $R(\psi)$, the nearer R is to unity the deeper the fades. Thus horizontal polarization is more susceptible to deep fades than vertical polarization and is consequently avoided, if possible, in practice.

When working beyond the radio horizon, the only radio waves that can be received at UHF are those that are diffracted round the obstacle presented by the earth's curvature (Norton 1941). The field strength falls rapidly from its value in the plane of the radio horizon towards the surface of the sea. For horizontal polarization the surface value is zero but there is a residual field strength, known as the ground wave, for vertical polarization. For vertical polarization, the height gain factor (in this region) represents the ratio of the field strength at a given height to its value at the surface. The height gain factor, as a function of height, does not vary much with distance once below the radio horizon. Thus, using published curves (Bremmer 1949) for ground wave propagation, and the height gain function the field strength at any point below the radio horizon can be computed. The field strengths in this region are also known to be smooth, monotonic functions of distance and height.

Theoretical results presented in this paper are confined to the Interference Region where all combinations of distance and height are above the radio horizon (*ie* within radio horizon range of the source) as illustrated in Fig 2. Such radio links are commonly described as line-of-sight; altitude h ft and radio horizon range r are related by the formula $r = \sqrt{1.5h}$ nautical miles. This formula and all the forthcoming theoretical results are based on a 4/3 earth radius system which anticipates the refraction of radio waves by a 'normal' atmosphere.

The reflection coefficients, $R(\psi)$, $\phi(\psi)$, for sea water have been calculated using the Fresnel formula (Reed and Russell 1966). The surface of the sea was assumed to be smooth and the effect of the slight attenuation due to divergence from the spherical surface of the earth has been taken into account.

The surface to air propagation predictions have been presented in Figs 3 and 4. The received signal strength is plotted against distance for each frequency, in Fig 3 for an aircraft altitude of 1000 ft, and in Fig 4 for an aircraft altitude of 4000 ft. The first interference max. and min. has been shown for each case. Over most of the region the inverse law of distance has been masked by the modulating effect of the multipath interference factor. This part of the curve, known as the roll-off region, has, in Fig 3 an almost constant attenuation rate of 1 dB per n mile. The air to air propagation predictions are presented in Fig 5 and 6. Signal strength is plotted against separation distance in Fig 5 with both aircraft at 1000 ft and in Fig 6 with both aircraft at 4000 ft. Apart from the forshortened roll-off region, the curves exhibit a basic (median level) inverse law of distance with an oscillatory modulation due to the multipath interference factor.

These predictions lack quantitative accuracy because certain practical system parameters were only estimated. The whole curve may have to be shifted up or down a number of decibels to match a real system, the shift being termed the scaling factor. Airborne experiments on a real system have enabled the appropriate scaling factor to be produced.

Fig 7 shows an example of many experimentally obtained propagation characteristics. The qualitative agreement with theory (Fig 3) is good and the suggestion of a log-linear decrease of signal strength with distance is borne out. The positions of the interference minima and the slope of the roll-off region both correspond closely to the theoretical predictions. Experimental propagation characteristics were obtained for each of the two helicopters (Wessex and Sea King) on both frequencies and quantitative analysis enabled the mean scaling factors to be produced for each aircraft on each frequency. A similar exercise was conducted for the experimental air to air results and the usefulness of these scaling factors will appear later.

The air to air propagation experiments were more difficult to control and it was expected that the propagation results should reflect the inevitable errors or fluctuations in navigation, altitude, pitch and yaw of both aircraft. As an intermediate step some propagation measurements were made in the Wessex flying at 1200 ft over the sea against a television broadcast transmitting mast of the same height above the sea. The signal strength of the vertically polarized sound channel radio wave on 201.25 MHz was accurately measured and recorded.

The radial distance from the mast was measured on a Decca navigator and the horizonatal propagation profiles were plotted. One of these is shown at Fig 8 and although direct comparison with Fig 5 for 237.3 MHz is not strictly admissible, the general features are in striking agreement with theory. Fig 9 shows the results of several vertical ascents (on different days) through the same multipath fading structure at a distance of 20 n miles from the transmitter. The interference minima in Fig 9 correspond fairly well with the same minima in Fig 8, the lowest orders having the greatest degree of cancellation, as expected. Taken as a whole, the experimental results indicate somewhat deeper fades than expected from theory. The Fresnel reflection coefficient leads to an expected maximum fading depth of 13 dB at the first interference minimum, whereas several fades over 15 dB have been measured.

Generally, the air to air propagation measurements produced rather irregular profiles, distorted reflections of Figs 5 and 6. The best profile has been presented at Fig 10 being the result of a 1000 ft sortie on 237.3 MHz. Both qualitative and quantitative agreement with Fig 5 is very good. The average, scaling factor taken over all the 1000 ft results turned out to be 0 dB. Fig 11 (for 237.3 MHz) is typical of the results for the 4000 ft flying. Over fairly wide regions the median signal strength was found to vary less than the amount of the multipath interference fading. The air to air experimental results also indicated slightly deeper fades than expected from theory.

Scaling factors were obtained for each of the three radio links: Wessex to shore station; Sea King to shore station; Sea King to Wessex. Expressing each scaling factor as the decibel sum of the installation losses at each end of the link leads to three simultaneous equations with three unknowns, and a unique solution is therefore possible. The installation losses came out to between 3.5 dB and 6 dB per unit, the Wessex having the lowest losses. The original assumption of 6 dB losses per unit in the model system is somewhat justified.

5 STATISTICS

It has already been stated that the method reported in this paper was to analyse, not the theoretical, but the experimental propagation curves. The following method could equally well have been applied to purely theoretical propagation predictions.

The air to surface results (1000 ft only) were found to be substantially independent of frequency and it was convenient to treat them all as one set of data. A representative 1000 ft propagation profile was constructed for the roll-off region having the mean slope and absolute value of the experimental results between 12 n miles and the radio horizon range. The slope was made -1.0 dB per n mile and the log-linear profile passed through the value of 24 dB (μV) at 32 n miles. For distances less than 12 n miles, the oscillatory signal profile was treated similarly to the air to air profiles (see Block D, *qv*).

The air to air propagation profiles were treated as random variations of signal about a well-defined median value. The dependence of the median level on distance was preserved by dividing each profile into blocks for statistical analysis. For the 1000 ft profiles, Block A covered the range 0 to 20 n miles and Block B covered the range 20 n miles to between 40 and 50 n miles where the first interference maximum was observed. The fairly broad roll-off between that and the radio horizon was treated as a non-fluctuating profile just as for the case of the air to surface results. The 4000 ft profiles showed relatively little variation in the median signal level in the ranges over which it was measured so all the data up to the first interference maximum was designated Block C. The roll-off regions for both altitudes were found to have the same slope of -1.2 dB per n mile, the signal being standardized at 12 dB (μV) at 70 n miles for 1000 ft, and 148 n miles for 4000 ft. Block D covered the fading region, 0 to 12 n miles, for the air to surface 1000 ft results.

For each of the statistical Blocks A, B, C and D the cumulative distribution functions were obtained and were plotted on arithmetic probability graph paper. Fig 12 shows an example, for Block B, and the good approximation to the Gaussian 'line' was repeated for the other Blocks. From the best Gaussian approximation in each case, the median, variance and truncation points were found, as described earlier. The results are summarized below.

Block	Altitude ft	Region n miles	Median dB(μV)	Variance dB	Truncation points dB
A	1000	Up to 20	31	7.05	±14
B	1000	20 to 40/50	23	7.0	±15
C	4000	Up to 148	19	4.7	±9.7
D	1000	0 to 12	40.7	4.7	±12.5

Although, during the air borne trials the Wessex and Sea King antenna azimuthal radiation patterns (ARP) were measured there were special reasons for confining attention to the Sea King ARP. A broadband UHF blade antenna had been specially fitted to the central, underside of the fuselage. The ARP were measured carefully on three frequencies taking the opportunity to repeat and cross check the measurements. Fig 13 shows the ARP for the Sea King on 237.3 MHz. The cumulative probability distribution functions for the received signal strength were compiled as a function of azimuth angle and plotted on arithmetic probability graph paper, an example being presented at Fig 14. As before, the median, variance and truncation points for the Gaussian approximation were determined. In this case the median has no essential significance; the antenna is ascribed a mean gain equal to that of a half-wave dipole and any discrepancies are borne by the scaling factor. For both the model frequencies, the Sea King antenna ARP yielded a common variance of 3.0 dB and the truncation points were ±6 dB.

A secondary propagation effect known as temporal fading was observed during the airborne trials. It was characterized by a roughly oscillatory excursion of the received signal strength in the frequency range from about 0.1 Hz to 0.3 Hz. It is thought to be due to atmospheric fading of the radiowave propagation, in which case the associated transmission factor would be independent of either multipath fading or ARP effects. In regard to equation (3), the propagation transmission factor T must be considered as the convolution of multipath fading and atmospheric, or temporal, fading. The frequency distribution was not analysed but reasonable estimates of the rms fading magnitude were made. These turned out to be 1 dB for the air to surface link and 2 dB for the air to air link. On the assumption that the fading was approximately random, its frequency distribution was taken as Gaussian with variances of 1 dB and 2 dB respectively.

Having the necessary frequency distribution statistics to determine the statistics of the received signal strength it remains only to convert signal strength into bit error rate. Here again *in situ* observations of bit error rate correlated with the instantaneous signal strength were preferred to laboratory measurements. The simulated data link consisted essentially of a continuously recycled 2047 bit m-sequence with synchronous (crystal controlled) detection. The modulation was amplitude shift keying with a modulation index of approximately 0.9 and demodulation was by envelope (non-coherent) detection. Only by means of an airborne experiment can the full effects of environmental electrical noise be evaluated. Taking special care to avoid error counts due to irregular causes, the recordings were scrutinized for simultaneous measurement of bit error rate and signal strength.

The plotted bit error rate versus signal curves were compared with the form of Fig 1. For reception at the relatively 'quiet' shore station the shape of the theory curve was fairly closely followed despite a wide scatter in the plotted points. The corresponding curves for the reception in the Sea King followed the shape of the theoretical curves but there was a range of about 6 or 7 dB in the threshold signal strength from curve to curve. This might have been due to signal measurement errors or to variations in the ambient noise levels. The curves for reception in the Wessex showed less variation in threshold signal strength but they all had a noticeably different slope from theoretical curve, see Fig 15. The tendency for relatively higher error rates at the higher signal end is interpreted as the effect of non-Gaussian noise in the channel. Such results would be produced by impulsive or 'spiky' noise characteristics, typical of manmade electrical interference. The Wessex bit error rate characteristic was chosen because it seems to typify the non-ideal situation. The effect of the threshold signal uncertainty, being about ±2 dB, could be incorporated into the overall statistical performance predictions by transferring it to a fictitious, independent signal fluctuation. This extra fluctuation would be assumed to have a Gaussian frequency distribution with a variance of 2 dB and could be handled in the same way as for the temporal fading.

6 PERFORMANCE PREDICTIONS

The derived frequency distribution functions can be combined in a number of different ways depending on which particular system it is desired to model. Different performance predictions will also result for the different propagation ranges considered above. In this paper, six situations arise as follows:

(i) Air to surface link: 1000 ft altitude, roll-off region;

(ii) Air to surface link: 1000 ft interference region (0 to 12 n miles) Block D;

(iii) Air to air link: 1000 ft propagation region, Block A;

(iv) Air to air link: 1000 ft propagation region, Block B;

(v) Air to air link: 4000 ft propagation region, Block C;

(vi) Air to air link: 1000 ft and 4000 ft roll-off region.

For the air to surface links, (i) and (ii), the surface station is assumed to possess an omni-directional aerial so the effects of only one ARP frequency distribution are considered. For the air to air links the Sea King ARP frequency distribution is included twice over to simulate a Sea King to

Sea King system. The two roll-off region cases, (i) and (vi), have no multipath fading contribution to the overall frequency distributions but the effects of temporal fading are included. Situations (iii), (iv) and (v) include the frequency distributions of two ARPs, multipath interference fading and temporal fading. The operations are summarized below.

Model	Variances of component F/Ds dB				Signal median	Overall variance	Truncation points	ΔQ
	First ARP	Second ARP	Multipath fading	Temporal fading	dB (μV)	dB	dB	%
(i)	3.0	0	0	1	Function of d	3.14	±6	3
(ii)	3.0	0	4.7	1	40.7	5.66	±18.5	0.05
(iii)	3.0	3.0	7.05	2	31	8.45	±26	0.1
(iv)	3.0	3.0	7.0	2	23	8.42	±27	0.075
(v)	3.0	3.0	4.7	2	19	6.62	±21.7	0.05
(vi)	3.0	3.0	0	2	Function of d	4.66	±12	0.5

For situations (ii), (iii), (iv) and (v) the statistics were stationary within the designated propagation ranges. For situations (i) and (vi) the median signal is a well behaved function (linear) of distance and therefore it is worth plotting the statistical parameters as a function of distance. The effect of the truncation is to cause the approximated Gaussian frequency distribution function to be somewhat greater than it ought to be just short of the truncation point. This error is transferred, by convolution, to the overall frequency distribution function and is worst nearest the overall truncation point. For practical purposes this error is important only at very low and at very high values of the Time Availability, Q . At the truncation points, Q goes to 0% and 100% respectively whereas the untruncated Gaussian function would give finite values of ΔQ% and 100-ΔQ%. Thus ΔQ% has been taken to indicate the possible error incurred by this method.

Fig 16 shows the predicted Time Availability plotted as a function distance for the 1000 ft air to surface data link. The specified grade of service is a bit error rate of 1 in 10^4. Fig 17 shows Time Availability as a function of distance for the 1000 ft air to air data link. The two regions can be seen where the statistics have been treated as stationary and the discontinuity at the boundary is of no intrinsic significance having been arbitrarily determined for ease of analysis. In Figs 16 and 17 additional curves have been plotted to indicate the Time Availability penalty paid by introducing a constant attenuation of Y dB into the system. Predictions of system performance may be guaranteed against a Y dB short fall in practice. Alternatively, basic doubts in the accuracy of source information used for these predictions may be covered by a suitable choice of Y . The experimental accuracy of the propagation results is not guaranteed better than ±6 dB; the rms scatter in the measured air to air propagation profiles was about 3.5 dB.

For the air to surface simulated data link trials, the useful range was found to be about 50 n miles, just a little beyond the radio-horizon range of 47 n miles. This result is compatible with the predictions but a direct comparison is meaningless since the trials aircraft did not fly in such a way as to manifest the full effects of the ARP, etc.

The theoretical propagation modelling is of particular use when seeking to extend the scope of predictions such as these. The detailed performance at different altitudes or for different system power budgets can be obtained easily by means of the Y parameter. In the air to air multipath interference region, changing altitude will hardly affect median signal levels but in the roll-off regions, a large effect will result. For the air to surface link, especially, the height gain factor will be of use.

7 CONCLUSION

Theoretical system modelling has been applied to the feasibility study of an airborne data link. Supporting experimental flying has confirmed expected general characteristics of multipath interference fading. Means have been found to improve the accuracy of the theoretical modelling and it has been possible to circumvent the need for full scale data link simulation trials. A simple but powerful technique has been used successfully to derive system performance predictions of bit error rate from either theoretical or experimental propagation characteristics.

REFERENCES

Bleaney, B.I. and Bleaney, B. (1965), "Electricity and Magnetism", Oxford at the Clarendon Press, 461-462.

Bremmer, H. (1949), "Terrestrial Radio Waves: Theory of Propagation", Elsevier, 118.

Carlson, A.B. (1968), "Communication Systems: An Introduction to Signals and Noise in Electrical Communication", McGraw-Hill, 385.

CCIR (1963), "World Distribution and Characteristics of Atmospheric Radio Noise", Report 322, Documents of the Xth Plenary Assembly, International Telecommunications Union, Geneva.

Norton, K.A. (1941), "The Calculation of Ground Wave Field Intensity over a Finitely Conducting Spherical Earth", Proc. IRE (December 1941)

Reed, H.R. and Russell, C.M. (1966), "Ultra High Frequency Propagation", Science Paperbacks and Chapman and Hall Ltd, 82-93

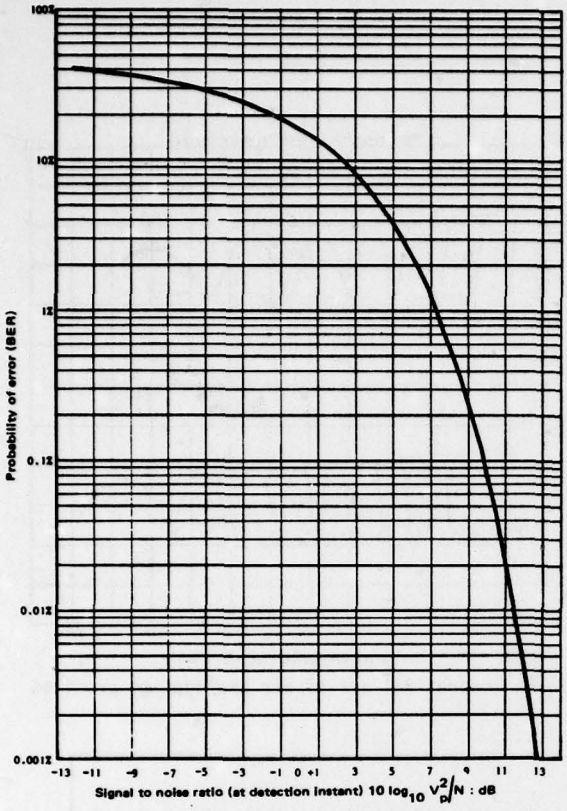
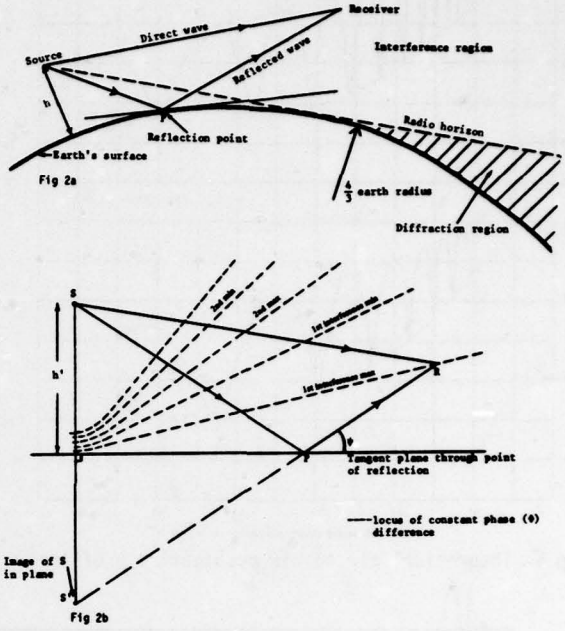


Fig 1 Theoretical error probability law, for Gaussian noise



Figs 2a&b Interference region geometry

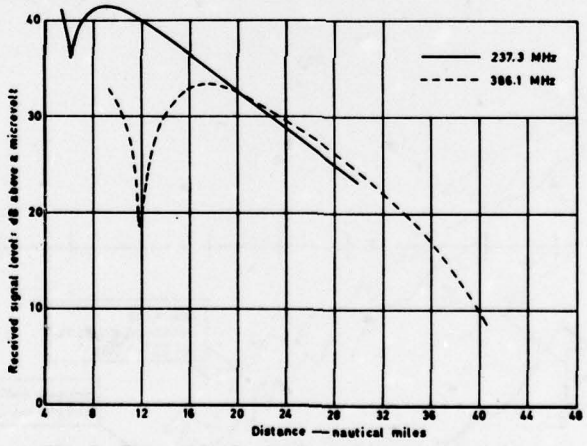


Fig 3 Theoretical single helicopter propagation profiles

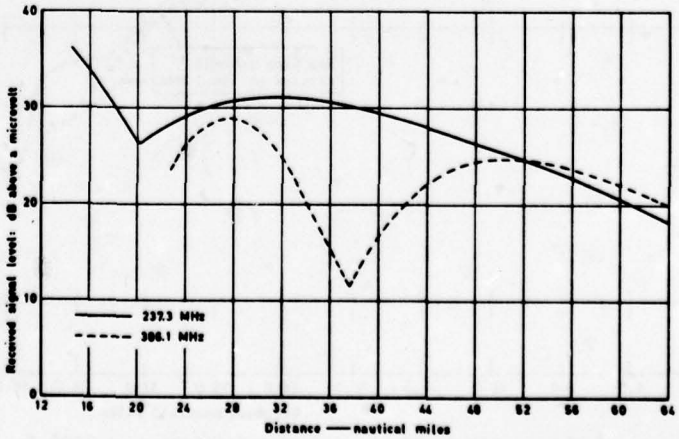


Fig 4 Theoretical single helicopter propagation profiles

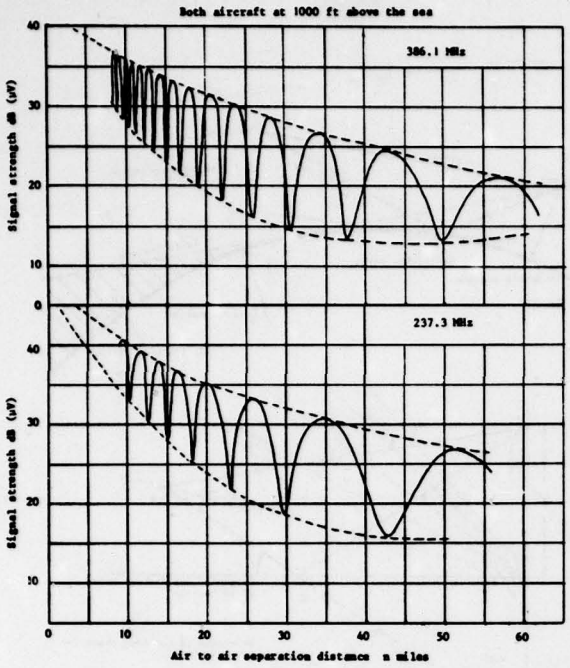


Fig 5 Theoretical air to air propagation profiles

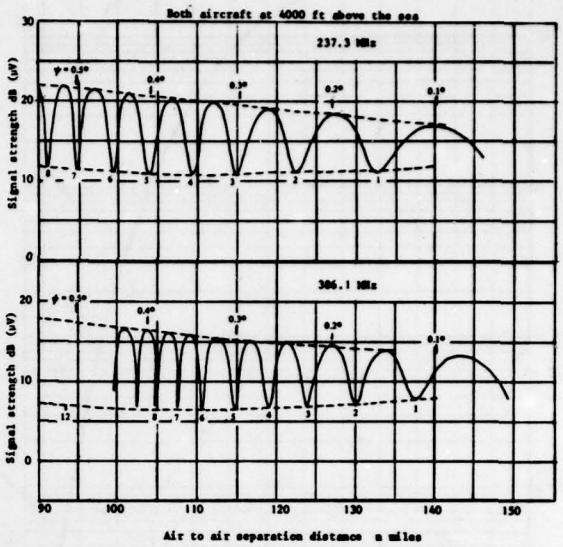


Fig 6 Theoretical air to air propagation profiles

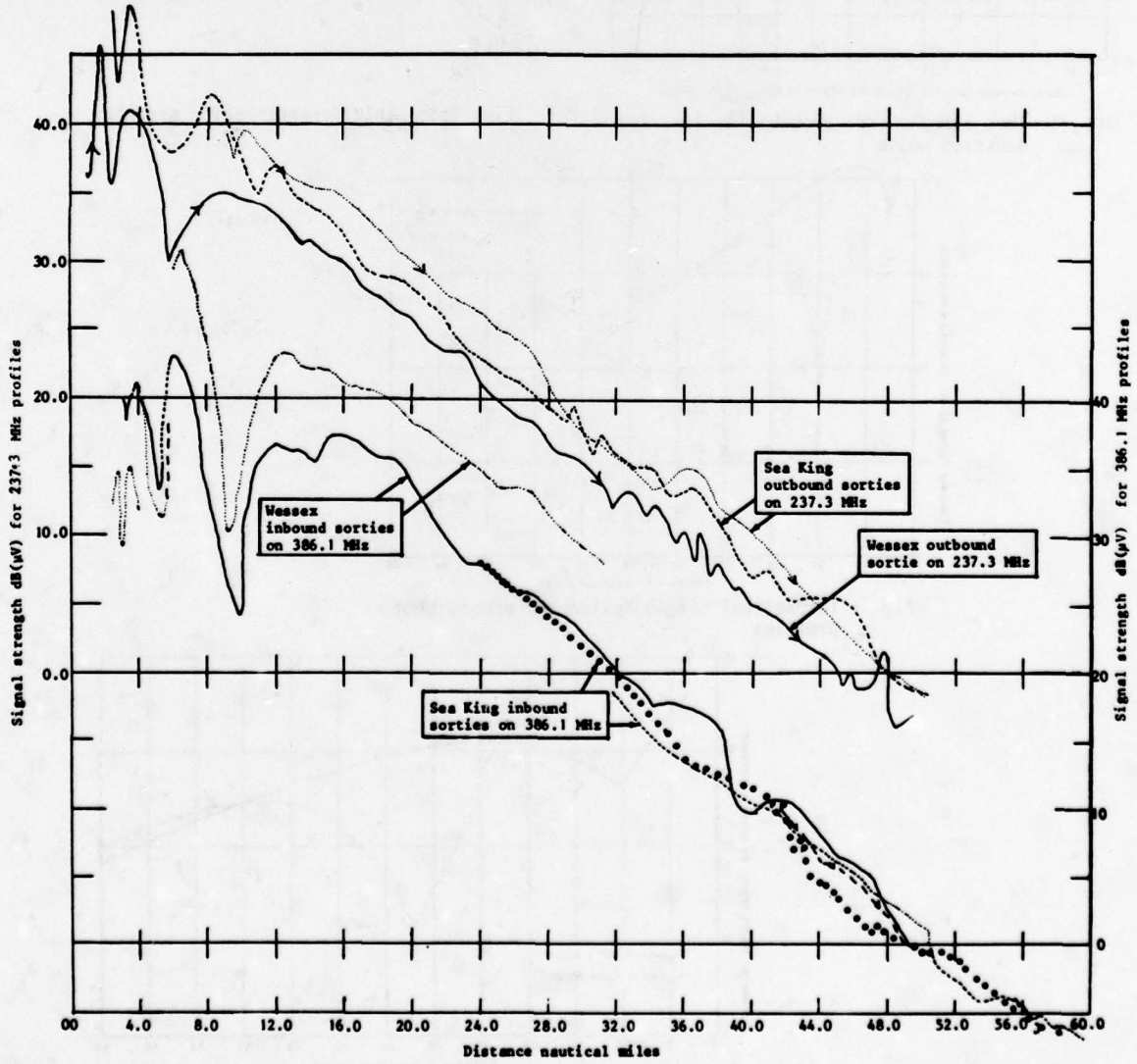


Fig 7 Air to ground transmission at 1000 ft altitude

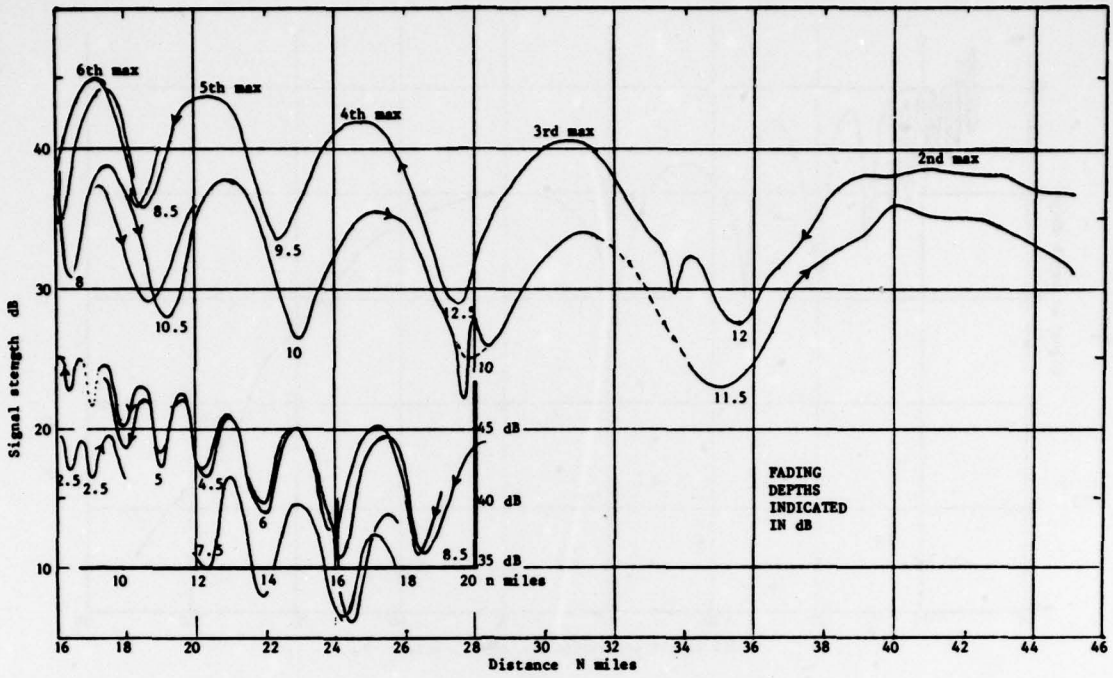


Fig 8 Horizontal profiles for sortie No 2
(1200 ft asl)

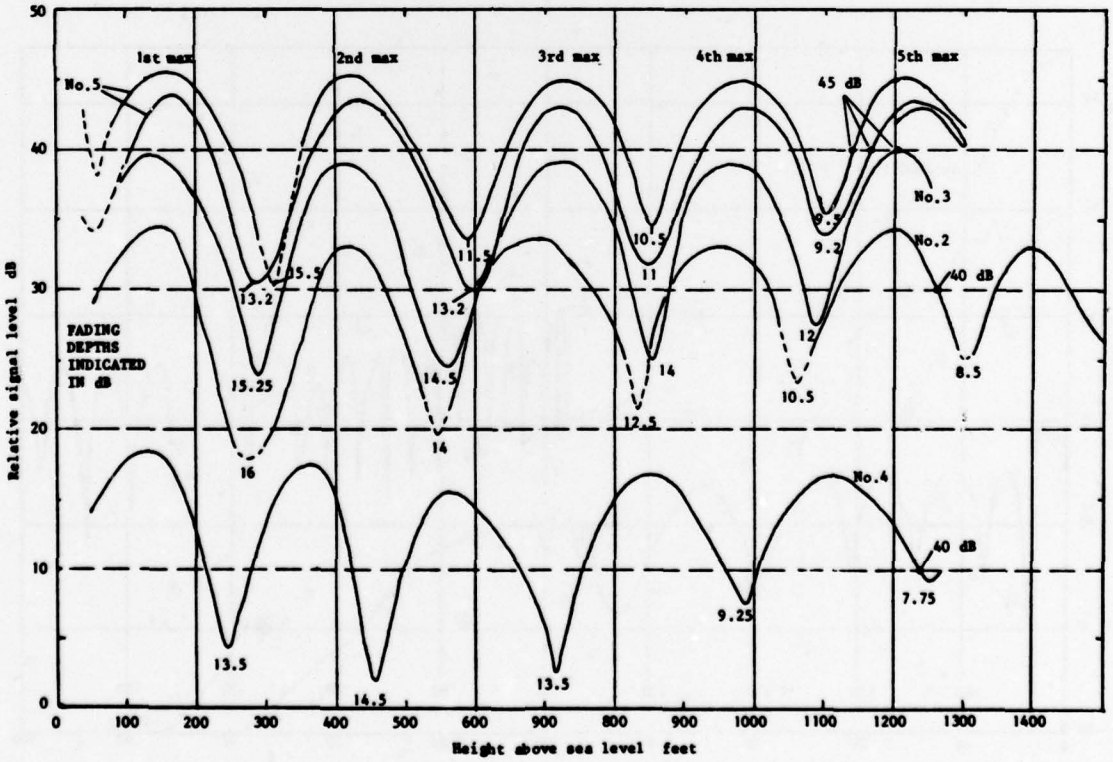


Fig 9 Vertical profiles at 20 n miles for sorties
2, 3, 4 and 5

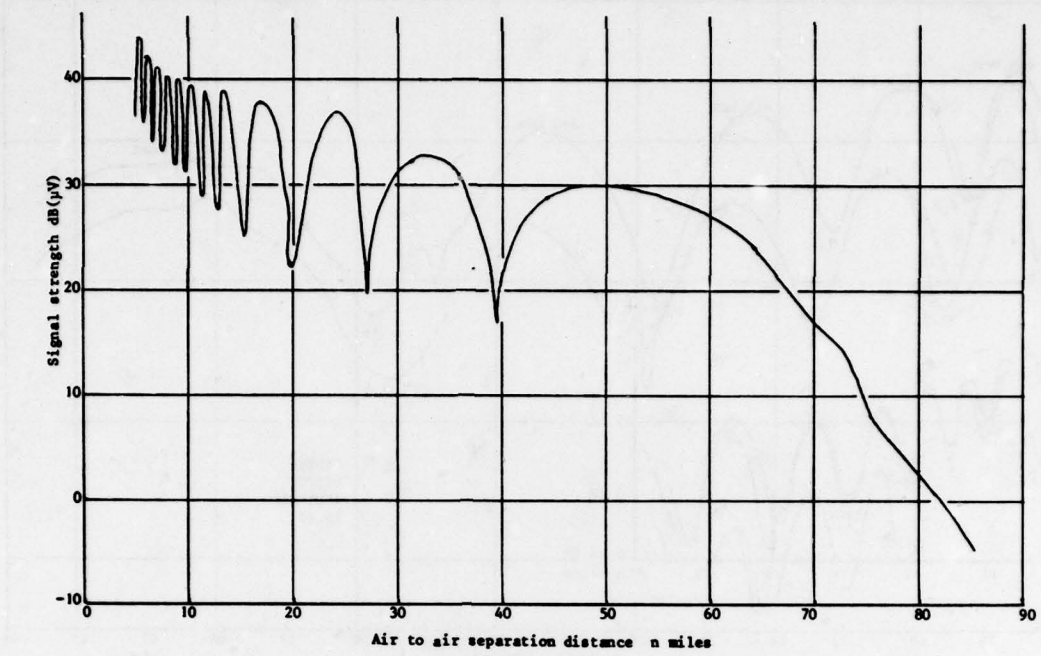


Fig 10 Air to air sortie No 4: inbound: 1000 ft: 237.3 MHz

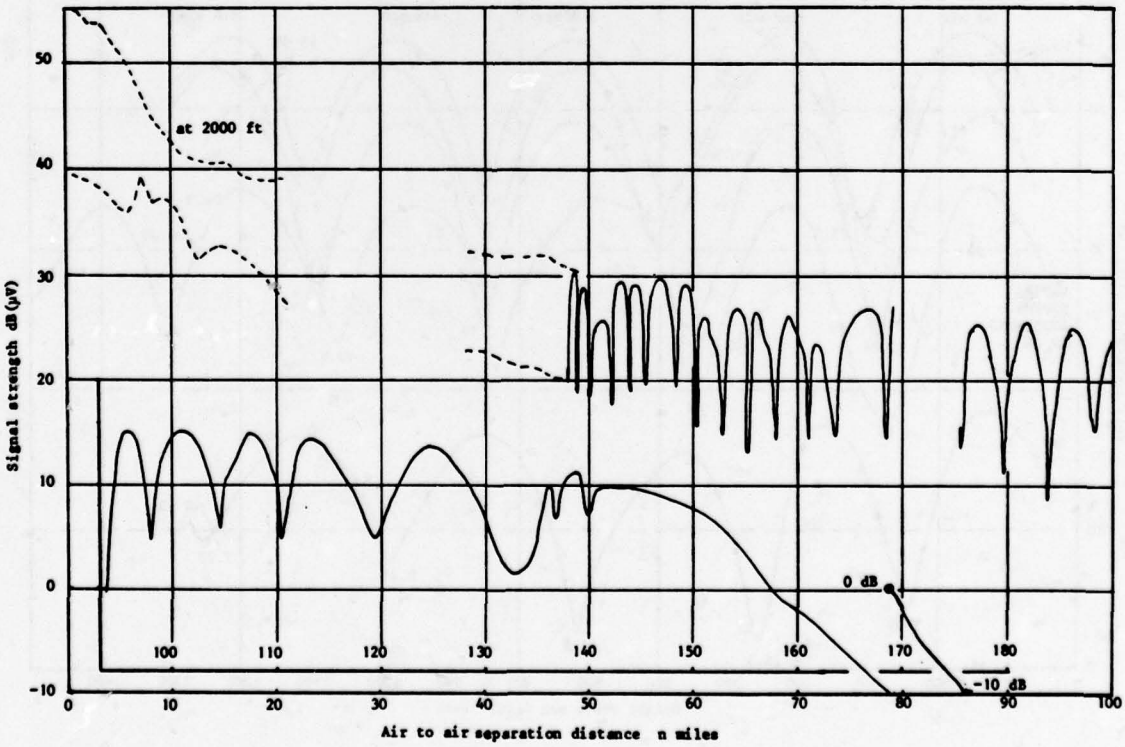


Fig 11 Air to air sortie No 1: inbound: 4000 ft: 237.3 MHz

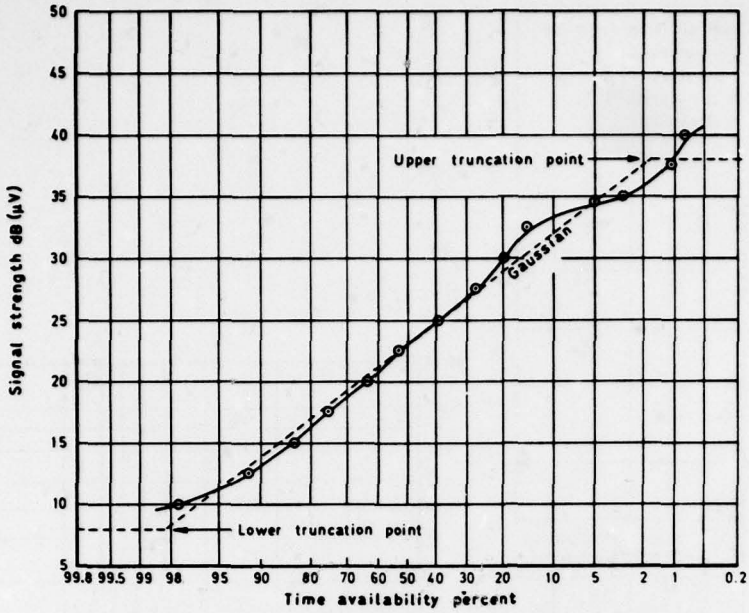


Fig 12 Cumulative distribution for block B

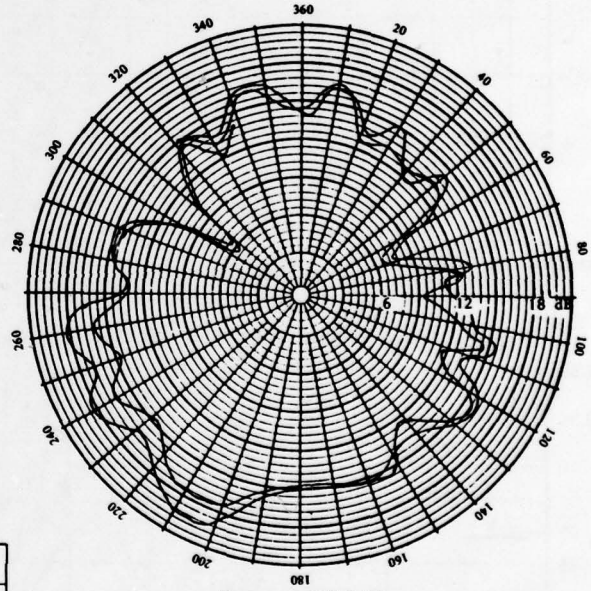


Fig 13 Sea King ARP's: trials aerial

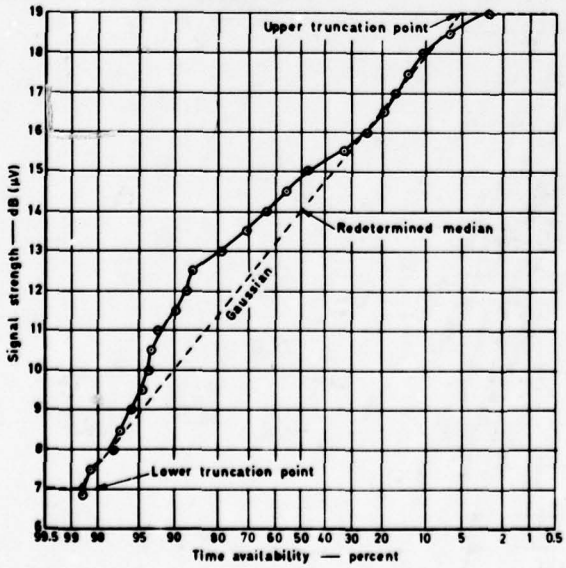


Fig 14 Cumulative distribution for Sea King ARP

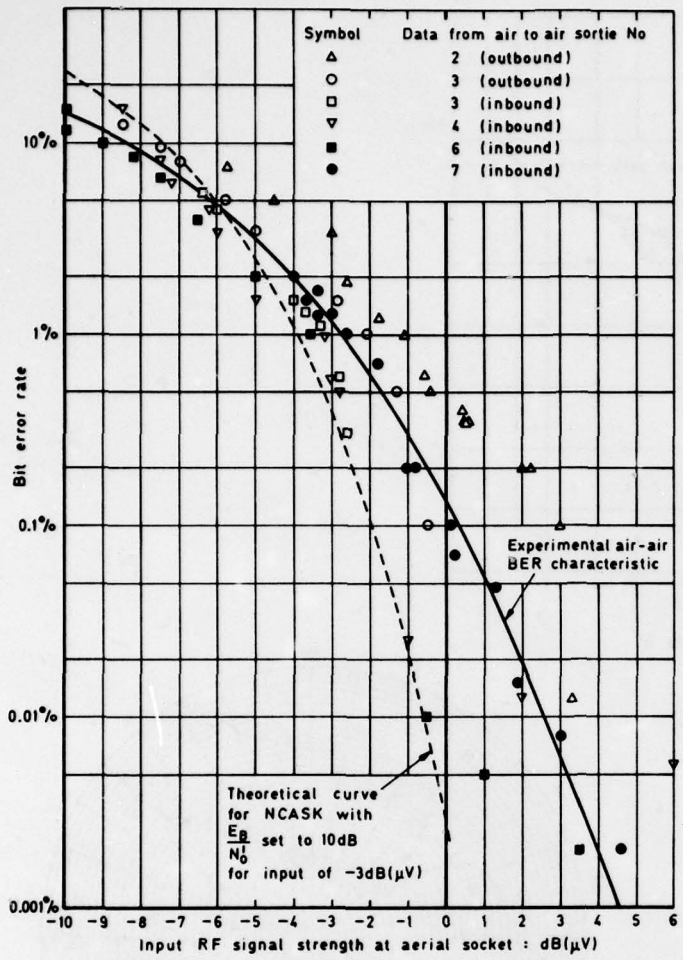


Fig 15 BER vs signal strength: Wessex (air to air)

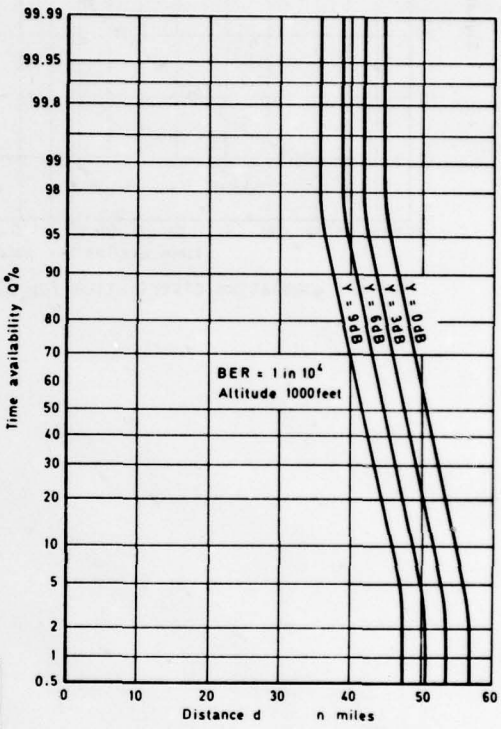


Fig 16 Estimated performance of the surface to air data link: time availability vs distance

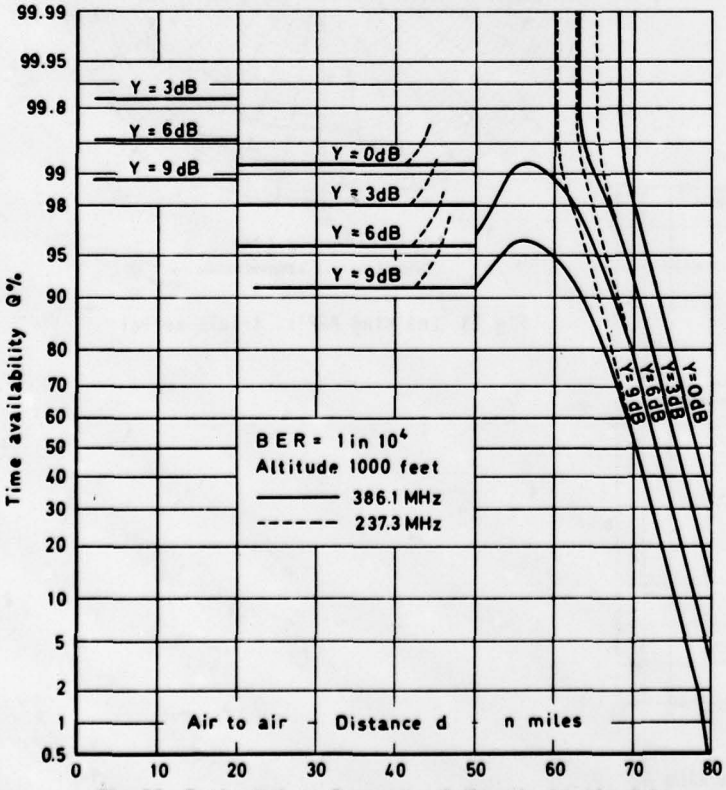


Fig 17 Estimated performance of the air to air data link: time availability vs distance

DISCUSSION

A.Sewards, Ca

With reference to Figure 15 of the paper – BER vs S/N – this figure appears to show evidence of the same type of effect of multipath on BER vs S/N as has been measured at L-band (Paper 9). Is the observed flattening of the curve believed to be due to this same effect, suitably modelled for the frequency difference between L-band and UHF, or to some other cause?

Author's Reply

The reduced slope of the BER vs SNR curve was originally thought to be due to the amplitude probability distribution of the electrical noise in the airborne environment. The flattening of the curve was observed only on Sea King to Wessex (i.e. air to air) transmissions. The normal shape of BER vs SNR characteristic was obtained for helicopter to ground station transmissions.

Practical departures from the theoretical BER vs SNR curve might be due to such phenomena as: impulsive types of electrical noise present at the receiver; rotor modulation due to either or both aircraft; the dynamic characteristics of the receiver AGC circuits (in the presence of rotor modulation or fast fading).

The bit error rates plotted in Figure 15 related to short periods of nearly constant signal strength. The only fading that could have been applicable to these measured signal levels would have been 16 Hz rotor modulation.

E.Ante, Ge

(1) Location of the antenna?

(2) Did you find any influence of the rotation of the rotor blades on the field strength – and if so can you describe the influence on the error rate?

Author's Reply

(1) Two antennae were used. The better of the two was a specially installed Chelton blade antenna (broad band) mounted on the nose, just in front of the cockpit. This antenna had an almost perfect radiation pattern in the forward view hemisphere and was used for many of the propagation measurements. The other antenna was also a Chelton blade but this time mounted in the dorsal position, on top of the tail boom to the rear of the gear box. It possessed several nulls in the radiation pattern.

(2) We have not made direct measurements of rotor modulation of the r.f. field at UHF but have noted such frequencies (typically 16 Hz) on the receiver's AGC line. Serious bit error rates were observed in bursts at the rotor crossing frequency, 16 Hz, and these were in synchronism with the oscillations observed on the AGC line. Subsequent investigation has shown variations in error rate performance with rotor modulation between different receive models according to their AGC characteristics.

H.Ecklundt, Ge

Did you observe any modulation by moving sea waves?

Author's Reply

We observed an effect, thought to be due to sea waves, at the minima of the multipath (spatial) fading pattern. It is thought that the specular reflection coefficient of the sea surface is modulated by the surface motion due to waves. The destructive interference of the direct wave and the once-reflected wave is very sensitive to small changes in the amplitude of either wave. This is expected to give rise to random, time dependent fluctuations of the signal strength observed at the multipath fading minima. The fluctuations had a strength of several dB's and characteristic frequency of 3 to 10 Hz. The sea state was less than 3, that is, fairly calm.

A.Becker, Ge

I. Did you consider diversity techniques in order to eliminate the fadings caused by ground reflections?

II. In response of the first answer of the author, I think of the use of frequency diversity. In this case only one antenna at each aircraft is required.

Author's Reply

I. Spatial diversity is impracticable on small aircraft, and although ships are capable of providing widely spaced antennae, competition for antenna sites means that only one antenna can be used for each radio.

II. All the antennae used could support frequency diversity transmissions but Service operational considerations forbid the use of more than one radio channel and associated transmitter/receiver for the system under study. Therefore we did not consider diversity techniques in this particular programme of work.

NEW INSIGHT INTO IONOSPHERIC IRREGULARITIES AND ASSOCIATED VHF/UHF SCINTILLATIONS

J. Buchau, E.J. Weber and H.E. Whitney
Air Force Geophysics Laboratory
Hanscom AFB, MA 01731, USA

SUMMARY

A program to study the physics, morphology and communications effects of ionospheric scintillations has been developed at the Air Force Geophysics Laboratory. Under this program recent airborne scintillation studies coordinated with several ground experiments have provided new insight into the structure of equatorial and auroral irregularities. Using a new all-sky photometer and airborne/ground based sounding, the large scale structure of equatorial irregularity regions responsible for VHF/UHF scintillations was determined. Field aligned electron density depletions of >1200 Km north-south extent develop after sunset in the bottomside of the equatorial ionosphere and move towards the east at ~100 m sec^{-1}. The lifetime of these depletions is several hours. They are the seat of irregularities (scale sizes meters to kilometers) which extend from the bottomside to >800 Km height. They occur single or in groups and result in scintillations of traversing signals. At arctic latitudes, the irregular F-region ionization resulting from soft particle precipitation is often associated with auroral forms. These forms have been made visible with the all-sky photometer and scintillation events observed on VHF/UHF were related to the observed features. Airborne scintillation measurements at the equator show strong dependence of the fading rate on aircraft heading, a result of the eastward drift of the irregularities. Ground based measurements show, that space diversity and time diversity are effective means to operate under strongly scintillated conditions.

1. INTRODUCTION

World-wide measurements of satellite signals, which started soon after the availability of the first satellites and which continue to the present day have established the existence of three major scintillation regions, where amplitude and phase fluctuations of the signals are observed (Aarons et al. 1971). These scintillation regions are found over both polar caps, and in a region centered on the magnetic equator (Figure 1). While high latitude scintillations are found at all local times in a region north of the scintillation boundary (Aarons et al. 1969), equatorial scintillations are a nighttime phenomenon, in general starting after local sunset, peaking before midnight and dying out in the early morning hours. They also exhibit a strong seasonal dependence (Mullen, 1973).

Satellite communications systems in use or in their test phases have made systems engineers and users well aware of the problems posed to the communications links by scintillations. These problems have increased interest in a thorough definition of the phenomenon and the geophysical processes which lead to and control scintillations. Of special interest are signal characteristics during scintillation events, geographic and temporal occurrence patterns, the predictability of the occurrence and means to mitigate or overcome the problems posed to the systems by scintillations.

Since the greatest occurrence of intense scintillations is observed in the equatorial scintillation region, the Air Force Geophysics Laboratory (AFGL) organized and conducted two equatorial campaigns in October 1976 and March 1977 to improve the understanding of the processes leading to the developmnnt of irregularities in the ionosphere and thus to scintillations. As Figure 2 shows, the campaigns involved satellite ground stations at Ancon and Huancayo, Peru and Natal, Brazil; the 50 MHz radar at Jicamarca, Peru and two jet aircraft, AFGL's Airborne Ionospheric Observatory and a communications test aircraft of the Air Force Avionics Laboratory. Section 2. of this paper deals with results from these campaigns which have contributed to a better description of the geophysical conditions observed during scintillation events.

Similar efforts have been devoted to scintillations in high latitude regions. Results pertinent to the insight into the arctic irregularity structure are discussed in Section 3.

Ground based and airborne satellite field strength data have been analyzed as to their spectral characteristics, amplitude probability distributions and cross and auto correlation functions. The results and application to diversity techniques are discussed in Section 4.

2. EQUATORIAL SCINTILLATION STUDIES

The often patchy nature of the occurrence of scintillations (Aarons, 1976) and evidence of individual, eastward travelling irregularity regions observed by HF propagation experiments (Rottger, 1973), suggested that disturbed regions of the equatorial ionosphere might be interspersed with undisturbed sections. Equatorial airglow studies had shown the existence of localized enhancements of 6300 Å airglow intensity (Steiger, 1967) and detailed airglow maps showed the presence of narrow north-south ridges of alternatingly enhanced and diminished intensity (Van Zandt and Peterson, 1968). The Jicamarca 50 MHz radar had shown the development of large irregularity structures, called plumes, starting after sunset in the bottomside F-layer and rapidly rising up to 1000 Km height. Woodman and LaHoz (1976) suggested the upwelling of a bottomside electron density (Ne) depletion, a bubble, to be responsible for the development of the plumes. McClure et al. (1977) showed the presence of Ne biteouts in the F-layer using Atmospheric Explorer satellite AE-C data and related them to the Jicamarca observations. It was considered likely that all of the observations were different manifestations of the same phenomenon. The equatorial campaigns were designed to measure simultaneously as many of the previously discussed features as possible in an attempt to tie the observations together into a unified picture.

2.1 Ground Based Observations

Scintillation data from a typical sequence of days from the March 1977 campaign are shown in Figure 3. Plotted are the scintillation indices measured on a total of six VHF links from 3 satellites received at Ancon and Huancayo, Peru for the period 24-30 March 1977. The data are organized so as to

present the ionospheric intersection points in a west-to-east sequence. Table I gives the satellites and frequencies used and the longitudes of the ionospheric intersection points of the respective signal ray paths for a height of 300 Km. The data show clearly the existence of isolated regions of irregularities and their eastward drift on 25, 26 and 28 March. Because of the smaller number of satellites observed on 29 and 30 March the pattern is not as clear on these days, while continuous scintillations were observed on 24 March.

TABLE I

Link	Frequency MHz	Ionospheric Intersection (300 Km) 0°W Longitude
ATS-3 (Huancayo)	136	77.6
GOES (Ancon)	136	77.0
GOES (Huancayo)	136	75.3
LES 9 (Ancon)	250	73.8*
LES 9 (Huancayo)	250	72.2*
MARISAT (Huancayo)	258	67.7

* Midpoints of slowly varying location of intersection points

The figure shows the tendency for the onset of scintillations on the westernmost ray path and progression of the disturbance to the east on 25 March (starting 0120 UT), 26 March (starting 0040 UT), and the second event on 28 March (starting 0300 UT). The passage of localized areas of kilometer size irregularities which are bounded to the east and to the west by undisturbed ionization, and which drift to the east at speeds of 100 to 300 m/sec^{-1} explains these observations. If the events are observed at later local times, 25 March (starting 0505 UT) and 26 March (starting 0420 UT), they often don't reach the more easterly ray paths, indicating dissipation of the irregularities in the early morning hours. On 28 March 1977 the scintillations disturbed first the more easterly ray paths (starting at 0000 UT on the MARISAT-Huancayo, the LES 9-Ancon and the GOES-Huancayo ray paths), followed by onset of scintillations on the more westerly ray paths (GOES-Ancon and ATS-3-Huancayo). The scintillation events cease on the different ray paths in the typical west-to-east manner. The development of irregularities in the vicinity of the three easterly ray paths, the growth of the developing region towards west following the direction of the motion of the terminator and the subsequent drift to the east of the fully developed irregularity region explains these observations. A similar event is shown in Figure 4 and will be discussed later.

The general relation between scintillation producing irregularity patches and Jicamarca radar "plumes" is shown in Figure 4. Jicamarca is located 340 km to the west of the Ancon-LES 9 ray path and 516 km to the west of the Huancayo-LES 9 ray path (300 km). The 50 MHz incoherent backscatter radar was operated in the digital power mapping mode such that the resulting range-time-intensity plots display received backscatter power levels above the incoherent scateer level. This technique allows mapping of the intensity of ionospheric irregularities. These irregularities form initially in the bottomside F-layer and are often observed extending as "plumes" from below 200 km to above 800 km (Woodman and LaHosz, 1976).

The top panel of Figure 4 shows the range-time-intensity plot from the Jicamarca radar on 19 October 1976. A plume formed at 2020 UT from a thin layer of bottomside irregularities, followed at 2135 UT by a second plume. The two lower panels show the scintillations observed during this night on the LES 9 signals received at Ancon and Huancayo. Scintillation events develop directly behind the terminator, first affecting the Huancayo-LES 9 link and then the more westerly Ancon-LES 9 link. These disturbances were not observed by the Jicmmarca radar, limiting the westerly extent of the disturbed region to a location east of Jicamarca. The two plumes, subsequently observed at Jicamarca can be seen in clearly isolated scintillation events, as they drift across the Ancon and Huancayo ray path The final scintillation event starting around 2315 UT on the Ancon link is not seen by the radar. It has been suggested (Rastogi and Woodman, 1977) that small scale (3 meter) scale size irregularities co-exist with kilometer size irregularities in the initial phase of equatorial F-layer disturbances. During the later phase however, the large scale irregularities persist much longer than those with smaller scale size. This results in a good correlation between Jicamarca backscatter plumes (reguling from irregularities of 3 meter scale size) and scintillations (requiring km size irregularities) in the early development phase of irregularity structures and poor correlation later, when the 3m irregularities have dissipated. (A detailed discussion of the relation between scintillations and radar backscatter is given by Basu et al., 1978).

2.2 Airglow Studies

To permit the mapping of airglow structures, which were expected to accompany the F-layer disturbances, the Airborne Ionospheric Observatory, a jet aircraft instrumented for ionospheric and propagation research, was equipped for the March 1977 campaign, with an all-sky photometer. This instrument was specifically developed for this purpose patterned after a similar ground-based system (Mende and Eather, 1976). The instrument is a wide field of view (155°), narrow spectral bandwidth TV system designed to operate in a time exposure mode. All-sky images of the equatorial airglow were made through 6300 Å and 5577 Å narrow band (30 Å) interference filters, using alternate 2.5 sec exposures to produce an image at each wavelength every 30 seconds. The resulting TV frames were then recorded on video tape and on 16 mm film by photographing a TV monitor. An example of an all-sky photometer 6300 Å image is shown in the left half of Figure 5. The grid lines are magnetic meridians in 1° increments, assuming a 250 km emission height. The 6300 Å airglow emission originates in the F-region primarily as a result of dissociative recombination of molecular ionized oxygen. This emission feature is a sensitive indicator of F-region height and electron density-changes; decreased intensity is associated with regions of low density or increased height of F-layer ionization. The bright airglow filling the portion of the sky from overhead to the western horizon and a second bright region extending from 2° east of the aircraft to the eastern horizon are indicative of substantial F-layer ionization below 300 km. The dark field aligned band or airglow depletion between these two bright regions is a phenomenon which was routinely observed during these equatorial flights and is the result of decreased ionization below 300 km. The ground projection on

the right side of Figure 2 gives an indication of the size of the field of view, and the dimensions of this airglow depletion, about 250 km in east-west direction and 1200 km north-south.

This specific depletion was observed during the 17 March 1977 flight, moving from western to eastern observation horizon within 3 hours. Figure 6 shows a series of 6300 Å images (photographs of the tape recorded video frames) selected at 15 minute intervals between 0100 and 0545 UT. All images have been reoriented with magnetic north to the top. For the purpose of this discussion the aircraft position can be considered stationary.

The images between 0100 UT and 0200 UT show a low level, unstructured glow with some enhancement towards the south, probably enhanced emission from the maximum of the Appleton anomaly. The 0215 UT image shows a prominent depletion in the 6300 Å airglow in the form of a dark band which extends from south to north along much of the western horizon. The formation of this dark band can be seen as early as 0200 UT. Within the next 2.5 hours, this band travels across the sky, leaving the instrumental field of view on the eastern horizon by approximately 0445 UT. Generally the images show that the eastern or leading edge of the airglow depletion is closely aligned in the magnetic north/south direction, (best seen in the 0330 UT image). The leading edge displays a sharp intensity gradient in the east-west direction while the western edge of the depletion region shows a somewhat more gradual, structured transition to the adjacent bright airglow region. The width of the depletion when directly overhead at 0330 UT is approximately 150-200 km. In the north-south direction, these regions extend across the entire field of view to include a horizontal distance of more than 1200 km, assuming a 250 km emission height. Unstructured airglow covers most of the observable sky until 0515 UT and then rapidly falls in intensity, leaving only minor enhancements towards the southern and western horizons. The position of the eastern and western horizons. The position of the eastern and western edges of the depletion were mapped and a rather constant eastward motion with a velocity of 90 m sec^{-1} was determined.

The ionospheric sounder records taken during the passage of the depletion are shown in Figure 7. The traces marked A-J and the corresponding range change shown in the graph in the lower right corner are backscatter echoes which closely track the depletion. The approaching sounder echoes (traces A-E) are from the trailing edge of the depletion, while the receeding echoes (traces H-J) are from the leading edge of the depletion. During the passage of the depletion over the aircraft the virtual height of the layer rose from initially 220 km to 265 km, and returned to a height of 240 km after the passage. The good correlation of sounder returns and 6300 Å images is shown in Figure 7, with the white dots representing the location of the approaching echoes, the black dots the location of the receding echoes. The dots were positioned assuming a 250 km height of the scattering region, to the west during the approaching phase and to the east during the receding phase.

The existence of an airglow depletion and the strong backscatter from the regions bordering the depletion suggest the existence of an electron density depletion at least in the bottomside F-layer, moving to the east with the speed and having the physical dimensions of the optical phenomenon.

The 50 MHz backscatter measurements (Figure 8) show the time history of the development or drift of 3 meter irregularities above Jicamarca. The picture can be understood either as the time history of an eastward drifting irregularity region observed from a fixed location or, assuming a time stationary ionosphere, as an instantaneous east-west cross-section of such a region.

Some irregularities are seen in the lower F-region from the beginning of the observations until 0300 UT. Starting at 0357 UT the first echoes from an extended region of irregularities are observed at 500 to 600 km height. This disturbance eventually involves the F-region between 175 and 670 km. Irregularities in the F-region start to disappear beginning below the 500 km level (0430 UT) and ending at the 200 km level (0450 UT), while some very weak irregularities above 500 km are observed until 0535 UT. The relatively uniform diffuse background appearing at the first digital level (0 to 6 dB above threshold) was caused by a computer malfunction, and does not indicate the existence of a diffuse background of weak irregularities. The vertical dashed lines in Figure 8 indicate the times, when the leading and trailing edges of the depletion moved over the Jicamarca site. The good correlation between these times and the passage of the backscatter plume is evident. A scintillation event (up to 14 dB) on the aircraft-LES 9 path started at 0400 UT lasting until 0553, during which time the depletion moved through the ray path. The eastward drift of the depletion region moved the ray path from initially lower to later higher altitudes. P1 is the ray path from the aircraft to the satellite at the start (0400 UT), P2 the ray path at the end (0553 UT) of the scintillation event with respect to the depletion. Thus, if irregularities responsible for the observed scintillations were confined to a volume within and vertically above the depletion, the outlined area would be a cross-section through the responsible irregularity region.

A good picture of the structure which is produced by this set of data is a bottomside Ne depletion, of 165 km width and at least 1200 km North-South extent, coincident with irregularities of scale sizes of 3 m to 1 km. These irregularities are found from 250 km to at least 700 km and imbedded in a generally undisturbed ionosphere they are moving to the east at 90 m/sec. None of the available data allowed assessment of the average electron density within this structure above the 300 km limit of the 6300 Å airglow emissions. (A more detailed account of this event is given by Buchau et al., 1978)

2.3 Multiple Structures

Besides isolated depletions, which result in isolated scintillation events, we have observed sequences of several depletions within the field of view of the all-sky photometer. A typical example of 4 depletions existing simultaneously within the field of view is shown in Figure 9. The image, taken at 0445 UT on 20 March 1977 shows one depletion to the east, one overhead and two to the west of the aircraft. As a comparison with Figures 5 and 6 shows, these multiple depletions are narrower than the single one observed on 17 March 1977. Like the latter they extend from north to south across the whole field of view. In Figure 10 we show the results of mapping the time histories of all depletions observed during this flight. All-sky photometer observations were made from 0230 until 0550 UT. Eight well defined depletions

were observed. The figure shows clearly that the depletions observed in the early part of the night (I and II) are considerably wider (2 to 3° east-west extent) than those observed later (III through VII), which have an east-west extent of .5° (or 50 km).

Depletion II shows a narrowing trend from initially 2° down to .5° east-west extent. A similar, but not as drastic narrowing is seen on Depletion I. The data suggest that depletions are initially wide, possibly during the formation stage, and gradually become more narrow. Another feature clearly evident from Figure 10 is the slowing down of the eastward motion later in the night. Depletion I moved east at 120 m/sec^{-1}, Depletion III moved east at 97 m/sec^{-1}, Depletion IV at 73 m/sec^{-1}, Depletion V at 43 m/sec^{-1} and finally Depletion VI and VII appear to be stationary. (A complete reversal of the direction of motion from eastward drift to westward has been observed in one single case (out of 7 nights of observations) on 26 March 1977 at 0040 LT.) Depletion II, which is observed for a considerably longer time than any of the others, initially moved east at 140 m/sec^{-1} but slowed down to 60 m/sec^{-1} after 0330 UT. The average speed from horizon to horizon was 90 m/sec. It is not clear from the data, if the Depletions III to V simply disappeared around midnight local time or if problems of aspect made them invisible to the all-sky photometer.

Scintillation data of the GOES/Ancon link (ionospheric intersection point 77°W) and of the GOES/Huancayo link (ionospheric intersection point 75.3°W) are schematically superimposed on the time history of the depletions. The data are shown in the top panel at the longitudes of the respective ionospheric intersection points. Since the satellite is approximately to the north of both stations, the ray paths are approximately parallel to the north/south aligned depletions. This permits a comparison of the passage of the depletion across the respective ionospheric intersection point and the observed scintillations.

Strong scintillations >10 dB at both ground stations are associated with the passage of Depletion I. They stop when the intersection points are approximately halfway into the undisturbed region between Depletions I and II with the more easterly Huancayo showing the expected time delay. Scintillations abruptly start with the occurrence of Depletion II first over Ancon and then over the more easterly Huancayo. For the rest of the observations scintillations continue. It is likely, that the close spacing of Depletions II to IV and lingering effects similar to those observed after the passage of Depletion I result in uninterrupted scintillations.

Shown in the middle panel of Figure 10 are F-layer virtual heights (h'F) determined from airborne ionoionograms. The aircraft flight track in the top panel allows one to assess the h'F measurements in their relation to the depletions. Crossing Depletion I, a maximum h'F of 255 km was measured at 0230 UT, approximately in the center of the depletion. In the middle of the undisturbed region between Depletions I and II the observes h'F reached a minimum of 200 km. A second maximum of h'F (225 km) is reached in the middle of Depletion II. Later h'F changes cannot be clearly associated with crossing of depletions, most likely due to the narrowness of the features. The changes of h'F in relation to depletions observed at early local times show the same results as discussed for the 17 March 1977 observations and substantiate earlier findings by Van Zandt and Peterson (1968).

Finally, in the lowest panel, we show in schematic form the Jicamarca backscatter observations. Indicated is the presence of backscatter below 300 km, between 300 and 500 km and above 500 km. A large plume had developed as early as 0053 UT (1953 LT) from a thin layer of irregularities around 300 km. For comparison with the all-sky images this figure begins at 0200 and shows only the remainders of a plume, extending well above 700 km. The backscatter returns die out at 0315 UT. A second weak event, limited to heights below 500 km starts at 0400 UT and ends at 0455 UT. The Jicamarca location is indicated in the depletion time history as a dashed line. A comparison shows good agreement between the end of the backscatter plume at 0315 and the crossing of the trailing edge of Depletion I. The onset of backscatter coincides with the appearance of Depletion II over the station while again backscatter persists for some time after the passage of the trailing edge. No further backscatter was observed, even though continuing scintillations were observed close to the Jicamarca meridian on the GOES/Ancon link for more than one hour. This again shows the tendency for 3m irregularities to disappear prior to the km size irregularities during the lifetime of these events.

2.4. Relation of Scintillation and Spread F

Ground stations rarely permit the one-to-one correlation between ionospheric soundings and the effects observed on trans-ionospheric propagation links, since sounders and satellite receivers seldom monitor the same ionospheric volume. The availability of two aircraft during the October campaign was used to provide for extended observations by the Airborne Ionospheric Observatory of the ionosphere in the vicinity the 300 km intersect point of the ray path from the AFAL aircraft to the LES 9 satellite. Figure 11 shows in geographic coordinates the sub-ionospheric tracks (300 km altitude) of the ray paths from the two aircraft to the LES 9 satellite on 17 October 1976. The actual AFGL flight track closely followed the AFAL sub-ionospeeric track (the actual AFAL track is not shown). Black bars along the sub-ionospheric tracks indicate the presence of scintillations on the respective flight tracks. The three increasingly wider bars relate to the scintillation index ranges <6 dB, 6 to 12 dB and <12 dB.

It is considered likely that the irregularity structure, which resulted in the 0210 to 0306 UT scintillation event observed by the AFAL aircraft is identical to that which produced the 0247 to 0330 UT event observed by the AFGL aircraft. From the onset and end times and using the aircraft position information, the patch had an 250-300 km east-west extent and moved east at 100 m/sec^{-1}. Ionograms, which were taken by the AFGL aircraft flying along the AFAL sub-ionospheric track clearly show the instantaneous development of spread F at the onset of the 0210 UT scintillation event. The airborne digital ionograms in Figure 12 were selected to show the ionospheric changes accompanying the scintillations. The 0204 UT ionogram shows a typicaltundisturbed nighttime ionosphere with a rather high F layer critical frequency (foF2>16 MHz, the upper frequency limit of the airborne sounder). At this time (2020 LT) the aircraft was at 15°N magnetic

latitude approximately under the maximum of the Appleton anomaly. The next ionogram at 0214 UT shows that strong range-spread has developed along the major part of the trace, simultaneous with the onset of scintillations. The ionogram at 0239 UT shows, besides some range spread, a well defined oblique trace at a range of 330 km. This suggests the existence of Ne gradients similar to those seen in March 1977 (see Section 2.2) where they could be related to the airglow depletions. By 0304 UT (ionogram not shown), two minutes before the cessation of scintillations, the range spread at frequencies above 6 MHz suddenly disappeared (foF2 = 9.4 MHz) and the ionogram at 0334 UT again shows a completely undisturbed F trace (foF2 = 8.0 MHz). Good correlation between irregularities which scatter VHF signals and range spread F had previously been reported by Rastogi and Woodman (1977) who correlated VHF forward scatter and ionospheric data. The second smaller event (0421 to 0528 UT) observed by the AFAL aircraft was also accompanied by range spread F. The spread F appeared between 0444 and0454 UT and lasted until 0524 UT. The time difference between the onset of the scintillations and that of spread F suggests that irregularities at larger heights, affectedthe ray path before the ionospheric sounder moved under the disturbance.

2.5 Symmetry of the Scintillation Region

Figure 1, which schematically combines a large scintillation data base collected over more than a decade, shows the symmetry of the equatorial scintillation region. To establish this as an instantaneous as well as statistical fact, an attempt was made on 23 March 1977 to determine the latitudinal extent of the scintillation region by two aircraft flying simultaneously from the north and from the south. The southbound flight track of the AFAL aircraft (A/C 662), and the northbound track of the AFGL aircraft (A/C 131), as well as the scintillation index determined from the LES-9 signal are presented in Figure 13. The north/south symmetry of the scintillation region is clearly shown, although scintillations measured in the southern hemisphere are lower than those measured in the northern hemisphere. The southern boundary was observed at 15.2°S magnetic latitude, while the northern boundary was at 13.8°N magnetic latitude. Since the azimuth angle to LES-9 for AFGL aircraft (A/C 131) was 56° east of north, and for AFAL aircraft (A/C 662) approximately 90°, the proper spatial correction will position the southern boundary approximately 2° further north, improving the observed symmetry of the scintillation region.

3. ARCTIC MEASUREMENTS

The high latitude ionospheric region can be separated into three distinct zones: the polar cap, surrounded by the auroral oval, which in turn borders on the midlatitude F-layer trough in the night/hemisphere. While all three regions show the presence of spread F and thus of irregularities, the energetic particle precipitation the auroral oval, and in the polar cap is responsible for strong localized electron density enhancements associated with the aurora, the F-layer irregularity zone (FLIZ) (Pike, 1972) and the auroral E-layer (Whalen et al., 1971). Polar orbiting beacon satellites such as the WIDEBAND atellite observed by high latitude stations are suitable to map scintillations in the various regimes. Figure 14 shows data from a WIDEBAND pass (Revolution 2874, 19 Dec 1976) which was received onboard the aircraft. The relative position of the aircraft, the satellite track and the auroral oval are shown in the lower half of the figure. Scintillations of 8 dB are observed on 137 MHz from prior to 0810 to 0813 UT, while the ray path was deep inside the polar cap.

As the satellite approached the auroral oval, scintillations ceased. No scintillations were observed as the path traversed the auroral oval region and entered the trough. The all-sky camera pictures taken during this pass show complete absence of discrete auroral forms in the field of view, which covered the statistical oval belt and reached far into the polar cap. The large signal fluctuations starting at 0818 UT on 137 MHz and 0821 UT on 378 MHz are most likely due to interference by a direct mode and modes reflected from the aircraft tail surfaces. A detailed correlation between scintillations and auroral forms has been made for several WIDEBAND satellite passes, when 6300 Å auroral forms were imaged by the all-sky photometer.

During a flight on 21 January 1977, designed to investigate noontime aurora, two orbits of the WIDEBAND satellite were within the all-sky photometer field of view. This provided an opportunity to investigate the effect of various auroral forms on UHF radio propagation from satellite-to-aircraft. Figure 15 shows in corrected geomagnetic (C.G.) latitude and C.G. time the sub-satellite tracks for two orbits, one near local noon and one at late evening. Time (in UT) in one minute intervals is shown along the sub-satellite tracks. The field of view of the all-sky photometer, projected to 250 km, is shown for times when the satellite is close to the aircraft. Within the field of view is the 250 km projection of the sub-satellite track for each minute, and a schematic representation of the location of the 6300 Å auroras. The actual 6300 Å image, in negative, is also shown. The black bands along the sub-satellite track (and along the 250 km projection) delineate the regions where scintillations in excess of 1 dB were encountered. When projected to auroral altitudes (250 km) these regions correspond to regions of soft particle precipitation.

The interpretation of the all-sky photometer images at other wavelengths confirm, that the noon sector aurora at 0538 UT resulted solely from soft precipitation and was confined to F-region heights, while the 0916 UT aurora was produced by more energetic particles resulting in F- and E-layer aurora. The scintillations observed in the noon sector (up to 8 dB at 137 MHz, 4 dB at 378 MHz) were thus produced by F-region irregularities. The brighter auroral forms seen in the night sector resulted in somewhat smaller scintillations (7 dB at 137 MHz, 2 dB at 378 MHz).

Figure 16 shows the respective satellite tracks projected into the all-sky photometer images. Dots at 1 minute intervals allow comparison of the scintillation data shown in the top of the figure with auroral forms. From this figure it is evident, that the more significant UHF scintillations are produced, when the ray path traverses the auroral form, while VHF scintillations are observed also in the vicinity of the arcs, even though the stronger VHF scintillations correlate well with the aurora.

4. SIGNAL CHARACTERISTICS

The morphological studies and investigations of the geophysical mechanisms leading to and controlling the scintillation events are by themselves important, since they define global regions which are subject to the disturbances. Through understanding of the causative processes these studies permit the improvement

of predictions beyond the pure probability of occurrence. But to design systems which are capable of coping to a certain extent, with the disturbed environment, additional information about the effects of the disturbed environment on the system is required.

Scintillations cause both enhancements and fading about the median level as the radio signal transits the disturbed ionospheric region. When scintillations occur which exceed the fade margin, performance of the communications link will be degraded. This degradation is most serious for propagation paths which transit the auroral and equatorial ionospheres. The degree of degradation will depend on how far the signal fades below the margin, the duration of the fade, the type of modulation and the criteria for acceptability.

The amplitude, phase, and angle-of-arrival of a signal will fluctuate during periods of ionospheric scintillation. The intensity of the scintillation may be characterized by the variance in received power. The measure S_4 is defined as the square root of the variance of received power divided by the mean value of the received power (Briggs and Parkin, 1963). Attempts have been made to model the observed cumulative amplitude distribution functions (cdf). Whitney et al. (1972) have constructed model distribution functions based upon the use of the Nakagami-m distribution (Nakagami, 1966) $m = (S_4)^{-2}$ and have shown that the models provide a reasonable approximation to their observed empirical distribution functions. The cdf is a first order statistic and is useful for defining the minimum margin requirements for the communications link of nondiversity systems. The Nakagami m-distribution has been shown to be practically useful for describing the effects of scintillations on satellite communication links (Whitney et al., 1972).

The Nakagami distribution is characterized by a single parameter which is related to the rms value of the intensity of the scintillations. It completely describes the distribution. Whitney et al. (1972) has shown that the experimentally determined distribution function of ionospheric scintillations closely approximates the theoretical Nakagami distribution. While m can have any value ≥ 0.5, the empirical results indicate that m=1 is the limiting case for intense scintillations and is a Rayleigh distribution.

In addition to the information on the amplitude of the fades which is given by the cdf, statistical descriptions of the fading rate is needed in order to fully characterize the effects of scintillations on the communications channel. The application of time diversity techniques such as coding or interleaving depends on information about the power spectra or time correlation functions of scintillations. The autocorrelation function characterizes the rate of scintillation fading. It is the Fourier transform of the power spectrum and has a width which is inversely proportional to the bandwidth of the power spectrum.

When correlation data is available it can be used to determine the improvement in performance that can be obtained through the application of diversity techniques. Autocorrelation data can be used to evaluate time diversity techniques; cross-correlation data from multifrequency measurements can be used to evaluate the effectiveness of frequency diversity; and, if spaced receiver measurements are available, then the cross-correlation information can evaluate space diversity.

Diversity schemes attempt to reduce the effects of fading during a scintillation event by combining two signals that are fading independently. Figure 17 shows the improvement in performance that is possible with dual-diversity techniques. The probability of achieving diversity gain is given for several values of ρ, the correlation coefficient. It is based on slow, multiplicative Rayleigh fading and equal signal-to noise ratios in both branches of a dual diversity system. Most of the diversity improvement is achieved for the condition $\rho<0.6$. For example the improvement at the 1% point is approximately 8 dB for $\rho=0.6$ and 10 dB for $\rho=0$ or complete decorrelation coefficient of 0.6 may be used as a threshold for diversity action for a fading process although it strictly applies only to Rayleigh fading (strong scintillation.

In the following sections we will discuss in some detail the signal characteristics for a typical scintillation event, observed on 19 and 20 October 1976 at Ancon (see Figure 4). The data from this event was processed to give the variation of the S_4 index and the autocorrelation and cross correlation functions. Since two antennas spaced on a 366 meter east-west baseline were used to record scintillation at Ancon, the spatial correlation function and drift velocity of the irregularities can be measured.

The variations with time of the S_4 index, of the autocorellation interval ($\rho = 0.5$), and the cross-correlation coefficient for a 3 ¼ hour period (0030-0345 UT) are shown in Figure 18. The S_4 index shows an abrupt rise at the onset of scintillations, and indicates the drift of several irregularity regions through the antenna beam. The S_4 index reaches unity during the passage of the first two regions.

The autocorrelation interval was short (~0.5 seconds) following the onset of scintillations, but in general varied between 1 and 2 seconds. The bandwidth or rate of scintillation varied by more than a factor of 4. Generally the autocorrelation interval was lowest during the most intense scintillations. The data indicates that time diversity techniques would have to provide delays of a few seconds to significantly reduce the effects of scintillations.

The crosscorrelation coefficient also showed great variability, ranging from a low of approximately 0.2 following the onset of scintillations to almost unity even though the S_4 index was approximately 1.0, an indication of very intense scintillations. While the antenna spacing of 366 meters is sufficient to provide significant space diversity improvement for some periods of intense scintillation, a much larger spacing would be required to provide the necessary decorrelation for the entire period.

The scintillation pattern at the west antenna leads the east antenna showing that the velocity of the irregularities is eastward. Measured values of the delay time over the 366 meter spacing gives velocities that typically vary from approximately 50 to 200 meters per second.

Diversity schemes are normally only required to overcome the effects of intense scintillations; weak scintillations are offset by a reasonable fade margin. Frequency diversity is not generally effective unless bandwidths of several tens of MHz are available. Measurements are sparse, but show that carrier frequencies separated by a factor of 2.5 at UHF are decorrelated under conditions of intense scintillations. Simultaneous fluctuations on orthogonal-polarized channels are highly correlated for frequencies above

100 MHz; therefore polarization diversity is not a viable technique. Time diversity requires that the same message be sent at two different times separated far enough for the correlation coefficient to be less than 0.6. Reference to Figure 18 shows that delays of the order of several seconds are required. Space diversity requires that the same message be detected at two receiving points separated far enough for the correlation coefficient to be less than 0.6. Again by reference to Figure 18 it can be seen that antenna spacing of 366 meters is sufficient to provide diversity improvement for some periods, but a much larger spacing would be required to provide decorrelation under all conditions of intense scintillations. Therefore space diversity seems only practical for ground stations and possibly shipboard application.

Of special interest to designers of airborne systems are those effects, which are specific for the airborne situation. It was previously shown that the field strength pattern, which results in the observed scintillations, moves towards the east at speeds between 50 and 200 m/sec^{-1}. This speed is comparable to the speed of the larger airglow and electron density depletions described in Section 2. To establish how the aircraft heading would affect the observed fading rate, flight legs were flown across ground stations with headings in increments of 45° covering the full 360° circle. Figure 19 is a typical example of the scintillations observed on an easterly and a westerly heading. The top panel shows data taken while the aircraft was on an easterly heading, flying with the drift of the field strength pattern and in effect slowing the fading rate down to as low as 1 fade per 45 seconds. The bottom panel is taken while the aircraft was flying due west, against the drift of the field strength pattern. The fading rate is substantially enhanced with fades as fast as one fade per 2 seconds observed. From data taken over several hours during 3 nights of operation it was established, that fading periods ranged from <1sec on westerly headings to >1 min on easterly headings. Figure 20 shows the results from the 20 October 1976 flight. Shown are the mean fading rates determined for the various aircraft headings plotted in a polar coordinate system of fades/minute and aircraft heading. Each flight leg was approximately 10-15 minutes long and the fading rate was established by counting positive crossings of the mean signal level in two minute segments. The bars show the ± 1 range. The figure shows a maximum mean fade rate of 20 fades/minute (.3 Hz) for a due west heading aircraft and a minimum mean fade rate of 2.5 fades/minute (.04 Hz) for a southeast heading.

Variations in the scale size of the field strength pattern as well as changes in the drift speed will naturally also result in fading rate changes. Though it is in principle possible to determine the drift speed and pattern scale size from the measured fading rates on various headings, the data base shown here including a total of 3h 22 min. covered too long a time to allow for the deduction of these parameters.

The spectral density estimate (Figure 21) determined for two seven minute samples, one from data taken on a due west heading, a second one taken from data collected minutes later on a NE heading show drastic effects in the shift of the knee frequency to lower frequencies, when the aircraft turns from a west to an east course. In addition a change of the frequency dependence of the roll-off from f^{-3} to f^{-2} is observed.

5. CONCLUSIONS

Airborne and ground based observations of scintillations and geophysical phenomena have produced a new large scale picture of the ionospheric structure during equatorial scintillation events. Structures of kilometer size electron density irregularities which are responsible for scintillation of trans-ionospheric VHF/UHF signals, form after sunset in large, field aligned volumes. These volumes are identified in the bottomside of the F layer by a depletion in electron density which results in a depletion in the 6300 Å airglow emission. This airglow depletion permits the mapping as well as the determination of drift, dynamics and lifetimes of the irregularity structures. They have a large north-south extent of more than 1200 km (recent new observations indicate that the depletions extend for more than 3000 km across the magnetic equator). The east-west width varies between 50 and 200 km. The Jicamarca 50 MHz Radar shows that 3 meter irregularities are co-located with the kilometer irregularities. The height extent of the irregularity structures (plumes) encompasses the F region from below 200 to well above 600 km. They occur as single depletions or appear in groups. All-sky images and ionospheric backscatter soundings permitted measurement of their eastward velocity as being as high as 140 m/sec early in the evening. At later local times they tend to slow down, or even reverse their drift direction after local midnight.

Airborne and ground based measurements have shown that the irregularity structures maintain their integrity as they drift eastward. Several airglow depletions have been observed for as long as three hours, moving from west to east across the field of view. The depletions show a tendency to become more narrow later in the evening and die out around local midnight. Even after their disappearance, scintillations and thus kilometer size irregularities seem to persist for some time. The scintillation region, the area in which the irregularity structures are found, is symmetric with respect to the magnetic equator. The signature of the irregularity structures in ionograms is the development of range spread F preceded often by the appearance of strong oblique echoes.

Arctic measurements have shown, that during quiet conditions (no discrete aurora present) scintillations were observed only in the polar cap. During the presence of well-defined 6300 Å (or F-region) auroras two DNA WIDEBAND satellite passes showed good correlation between aurora and VHF/UHF scintillations.

The investigations of the signal characteristics of scintillated signals have shown, that the Nakagami-m distribution is a reasonable approximation of the observed amplitude distribution functions. Using a two antenna system, auto and cross correlations were determined to investigate the effectivenes of diversity techniques to overcome scintillation related communications problems. The autocorrelation interval determined from a selected typical scintillation event varied between one and two seconds, suggesting that time delays of a few seconds would be required to significantly reduce the effects of scintillation. The cross-correlation coefficient was rather variable, but the data suggest that much larger than the used antenna spacing of 366 meters is required to provide the necessary decorrelation for the entire period.

Signal characteristics determined from the airborne studies show a strong heading dependence of the fading rate with fadings as low as 1/minute observed on east headings and as high as 1/sec observed on west headings, a direct result of the eastward drift of irregularities and thus of the field strength pattern. This suggests that for time diversity purposes delays in the order of minutes would be required. Space diversity onboard an aircraft is not feasible considering the results of the ground based measurements.

A follow-up study will address the development phase of the irregularities as well as the reasons for the reported strong longitudinal differences in seasonal occurrence patterns and in severity of equatorial scintillations. Results from the reported and the planned work will feed into the predictive scintillation modelling.

ACKNOWLEDGEMENTS

We thank the following members of the Air Force Geophysics Laboratory for their contributions: Dr. J. Aarons and Lt Col A.L. Snyder for their support of and interest in the scintillation research program. J.P. Mullen for providing scintillation data from Huancayo. R.W. Gowell, J.B. Waaramaa and J. W.F. Lloyd for the engineering and flying support of the airborne missions. R. Carnevale for providing logistics support. MSgt G.A. Coolidge and AlC P.J. Diroll for the enthusiastic support of the all-sky photometer data analysis. We thank the personnel of the Jicamarca Radio Observatory for taking and processing the radar information and Dr. J.P. McClure for providing the radar maps. Ing. A. Bushby, Instituto Geofisico del Peru, was instrumental in providing the scintillation data from Ancon. The strong support from the air and ground crews of the 4950th Test Wing, Wright-Patterson AFB, Ohio, made it possible to achieve the goals of this program.

This research was sponsored, in part, by the Air Force Laboratory Independent Research Fund of the Air Force Geophysics Laboratory, Air Force Systems Command and, in parts, by the Defense Nuclear Agency under Subtask Code 125AAXHX633, Work Unit 33.

REFERENCES

Aarons, J., J.P. Mullen and H.E. Whitney, 1969, The Scintillation Boundary, J. Geophys. Res., 74, 884-889.

Aarons, J., H.E. Whitney, R.S. Allen, 1971, Global Morphology of Ionospheric Scintillation, Proc. I.E.E.E., 59, 159-172.

Aarons, J., 1976, Equatorial Scintillations: A review, Air Force Survey in Geophysics, No. 341, AFGL-TR-76-0078.

Basu, Santimay, Sunanda Basu, J. Aarons, J.P. McClure and M.D. Cousins, (submitted 1978 to J. Geophys. Res.), On the Co-Existence of Km-and M-Scale Irregularities in the Nighttime Equatorial F-Region.

Briggs, B.H. and L.A. Parkin, 1963, On the Variation of Radio Star and Satellite Scintillations with Zenith Angle, J. Atmos. Terr. Phys., 25, 330-366.

Buchau, J., E.J. Weber and J.P. McClure, 1978, Radio and Optical Diagnostics Applied to an Isolated Scintillation Event, Proceedings of Ionospheric Effects Symposium, 24-26 January 1978, Arlington, VA., in print.

McClure, J.P., W.B. Hanson and J.H. Hoffman, 1977, Plasma Bubbles and Irregularities in the Equatorial Ionosphere, J. Geophys. Res., 82, 2650-2656.

Mende, S.B. and R.H. Eather, 1976, Monochromatic All Sky Observations and Auroral Precipitation Patterns, J. Geophys. Res., 81, 3771-3802.

Mullen, J.P., 1973, Sensitivity of Equatorial Scintillations to Magnetic Activity, J. Atmos. Terr. Phys., 35, 1187-1194.

Nakagami, M., 1960, Statistical Methods in Radio Wave Propagation, edited by W.C. Hoffmann, pp. 3-36, Pergamon, N.Y.

Pike, C.P., 1972, Equatorward Shift of the Polar F-Layer Irregularity Zone as a Function of the Kp Index, J. Geophys. Res., 77, 6911-6915.

Rastogi, R.G. and R.F. Woodman (accepted for publication 1977) VHF Radio Wave Scattering due to Range and Frequency Types of Equatorial Spread F, J. Atmos. Terr. Phys., in print.

Rottger, J., 1973, Wave-like Structures of Large Scale Equatorial Spread-F Irregularities, J. Atmos. Terr. Phys., 35, 1195-1206.

Steiger, W.R., 1967, Low Latitude Observations of Airglow, Aurora and Airglow (B.M. McCormac, ed.), Reinhold Publ. Co., 419-433.

VanZandt, T.E. and V.L. Peterson, 1968, Detailed Maps of Tropical 6300Å Nightglow Enhancement and Their Implications on the Ionospheric F2 Layer, Ann. Geophys., 24, 747-759.

Whalen, J.A., J. Buchau and R.A. Wagner, 1971, Airborne Ionospheric and Optical Measurements of Noontime Aurora, J. Atmos. Terr. Phys., 33, 661-678.

Whitney, H.E., J. Aarons, R.S. Allen, and D.R. Seeman, 1972, Estimation of the Cumulative Amplitude Probability Distribution Function of Ionospheric Scintillations, Radio Science, 7, 1095-1104.

Woodman, R.F. and C. LaHoz, 1976, Radar Observations of F-Region Equatorial Irregularities, J. Geophys. Res., 81, 5447-5466.

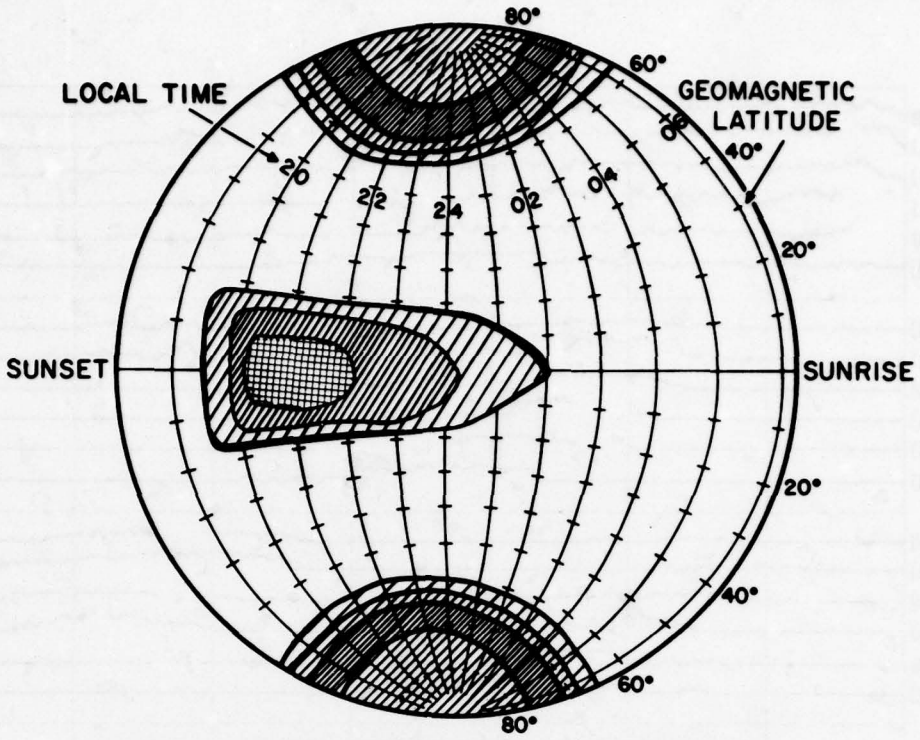


Figure 1 Global Scintillation Regions

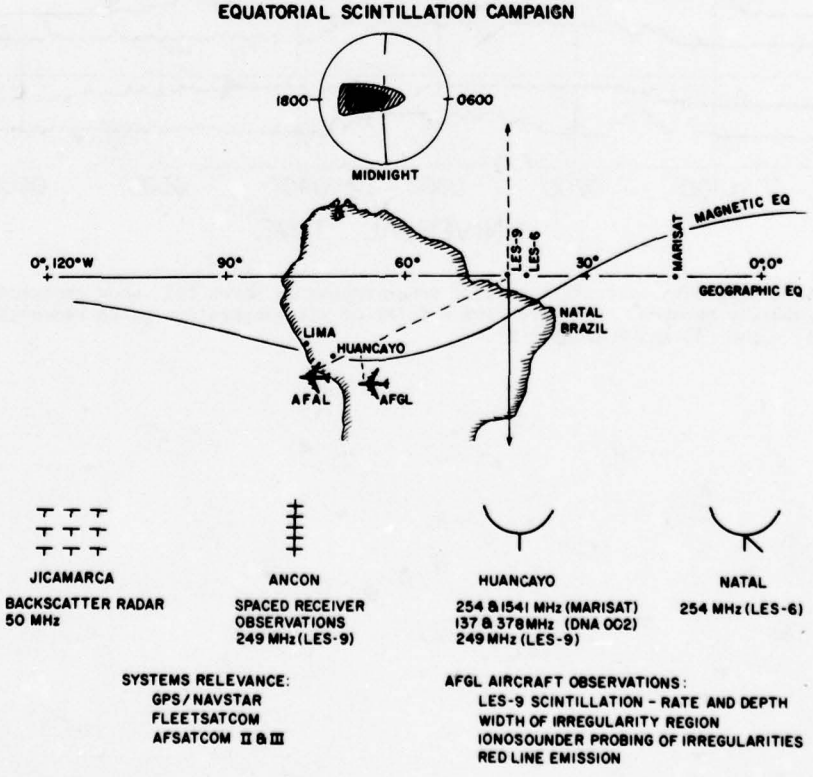


Figure 2 Equatorial Scintillation Campaigns conducted in October 1976 and March 1977 combined two aircraft and several ground experiments.

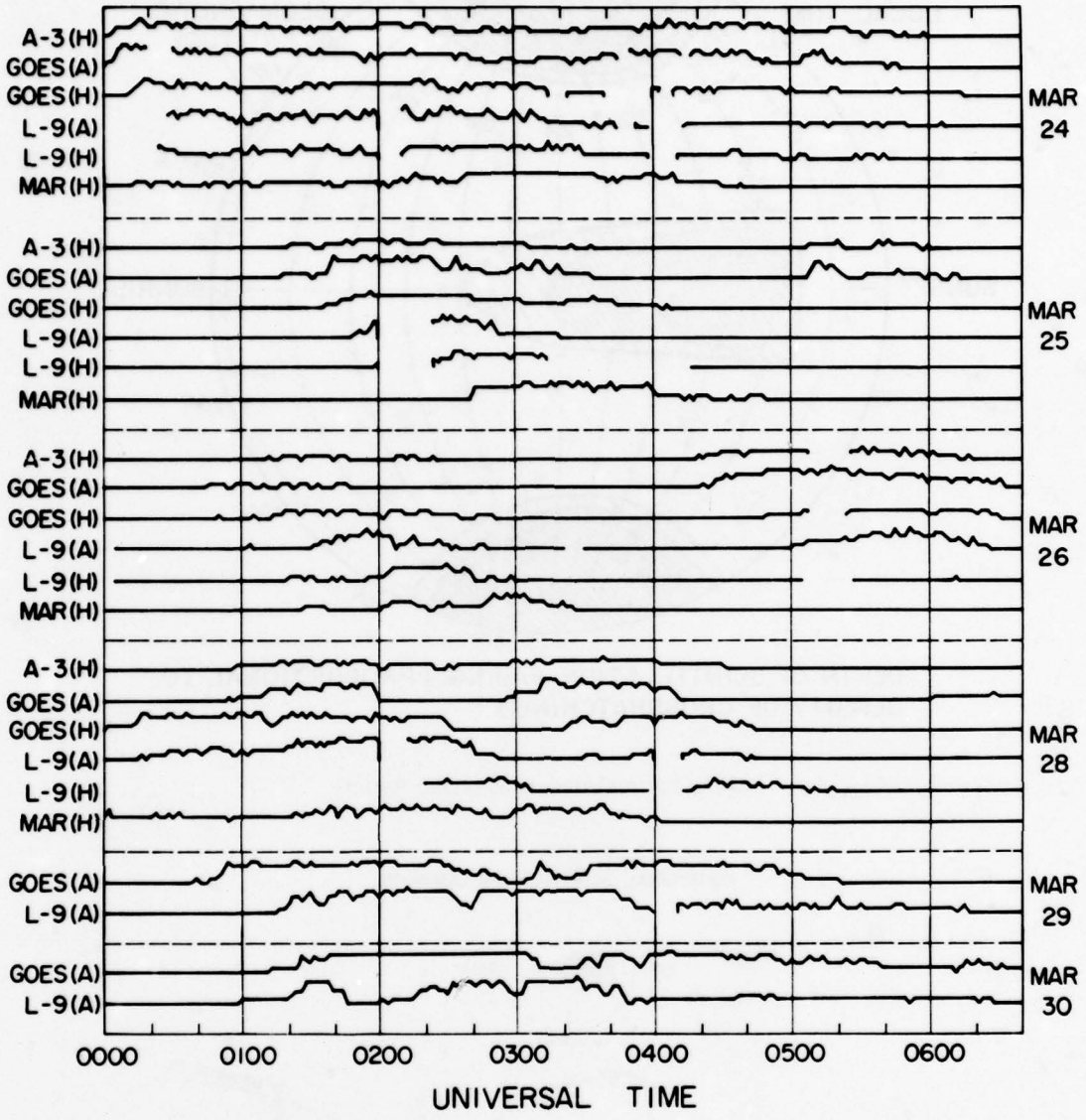


Figure 3 Scintillation data covering 6 days of measurements in March 1977 show eastward drift of irregularity patches. Data are from a total of six propagation paths recorded at two ground sites, Ancon (A) and Huancayo (H).

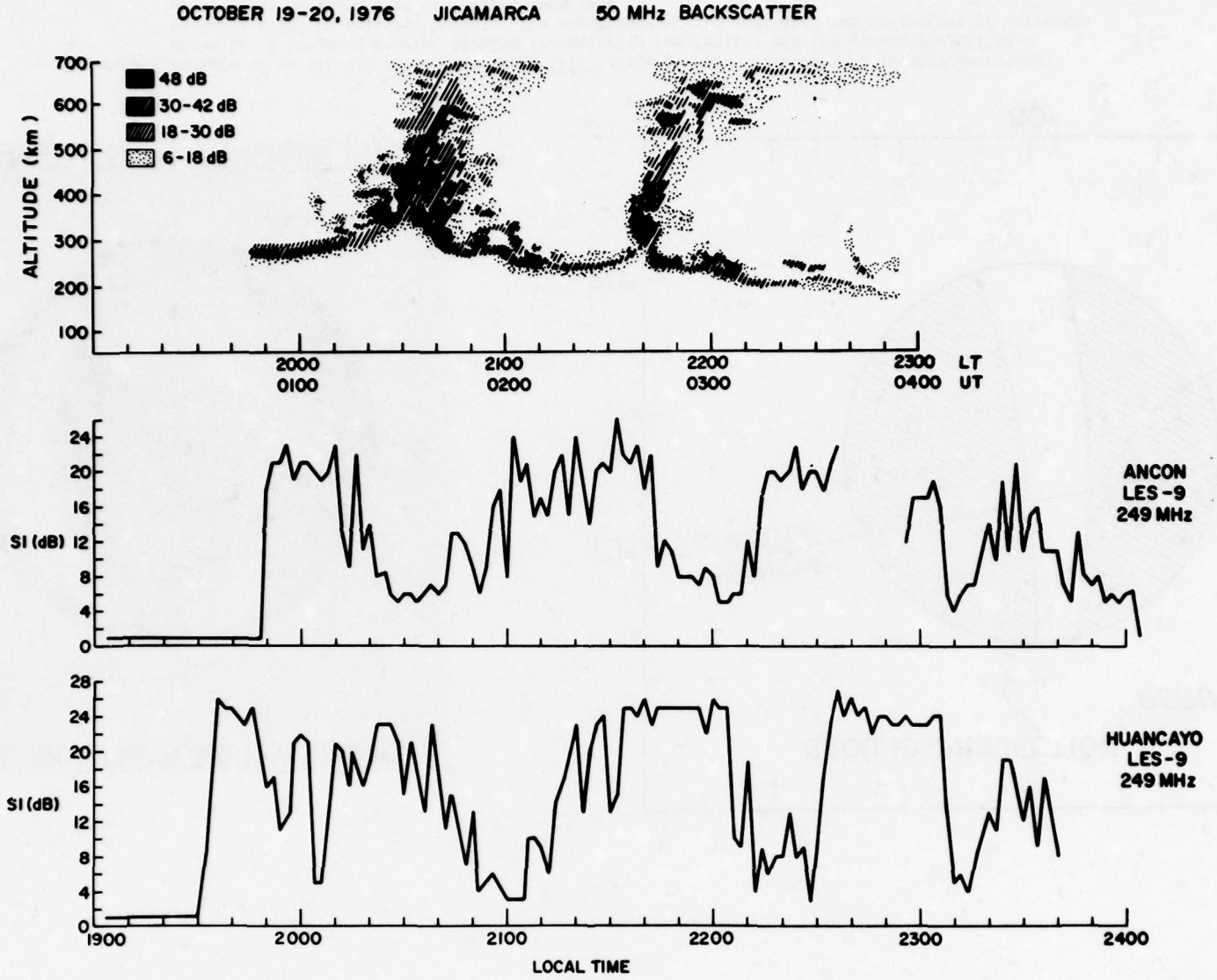
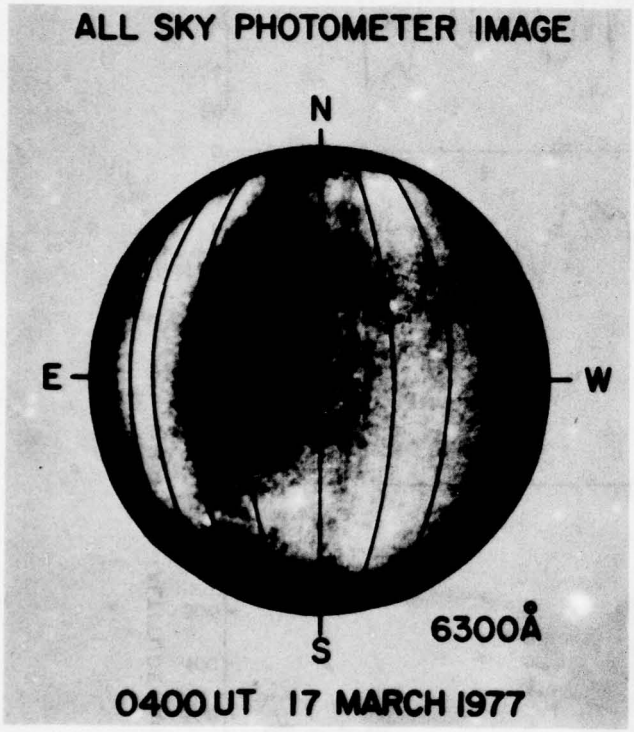


Figure 4 Jicamarca radar backscatter map and corresponding scintillations observed at Ancon and Huancayo.



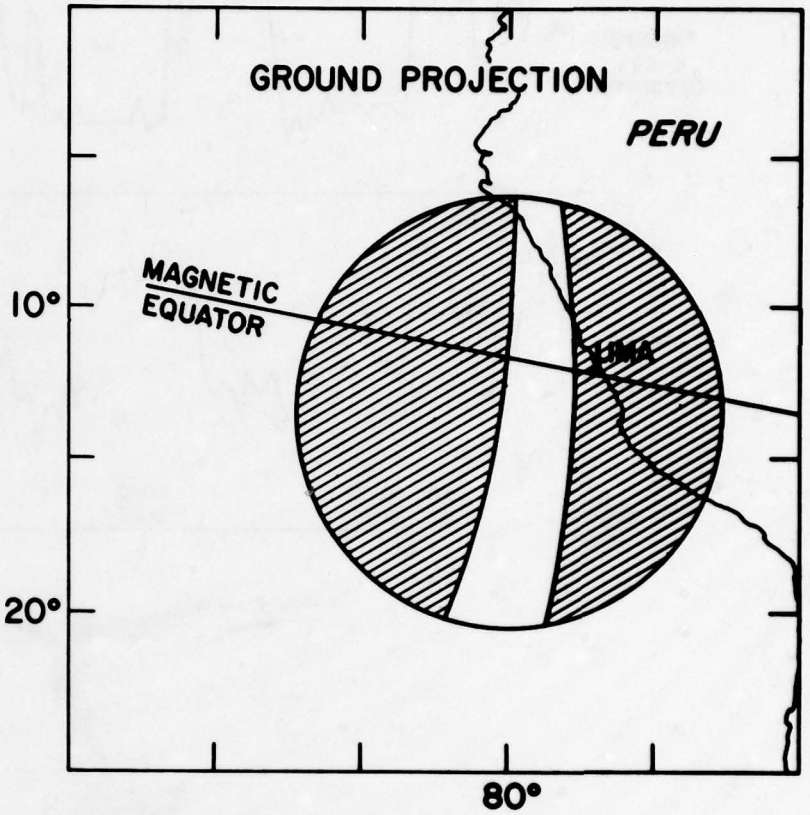


Figure 5 Example of an All Sky Photometer image (155° field of view) of 6300 Å OI airglow emissions. Shown is a large equatorial airglow irregularity (depletion) and its ground projection assuming a 250 km emission height. The superimposed grid indicates the projection of corrected geomagnetic meridians, at one degree intervals.

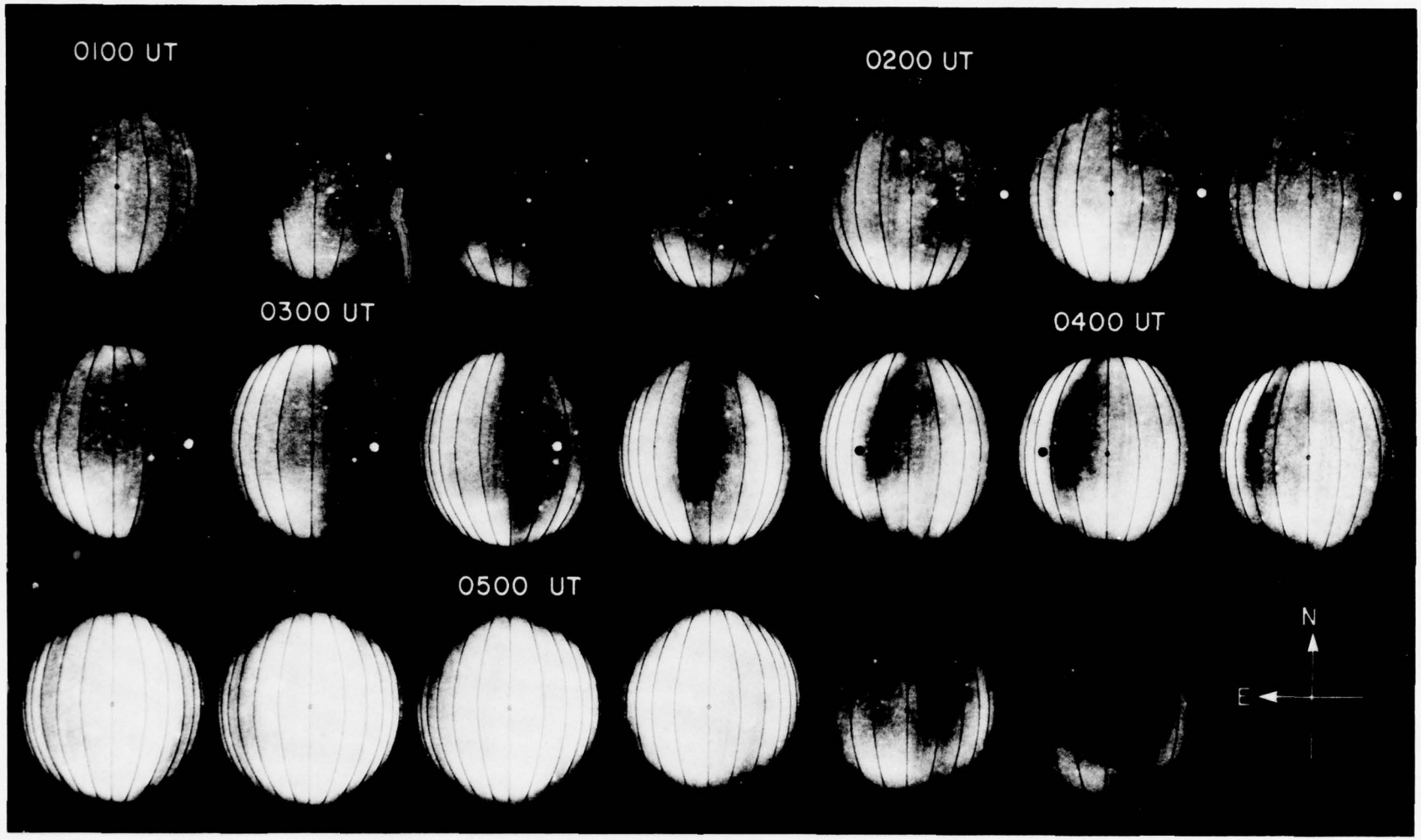


Figure 6 All Sky 6300 Å OI airglow images at 15 minute intervals, from 0100 UT to 0545 UT, 17 March 1977. The figure shows the motion of a large airglow depletion from west to east between 0200 and 0500 UT. The black and white dots represent respectively the location of approaching and receding oblique F-region ionosonde backscatter returns. The backscatter returns follow closely the motion of the airglow depletion.

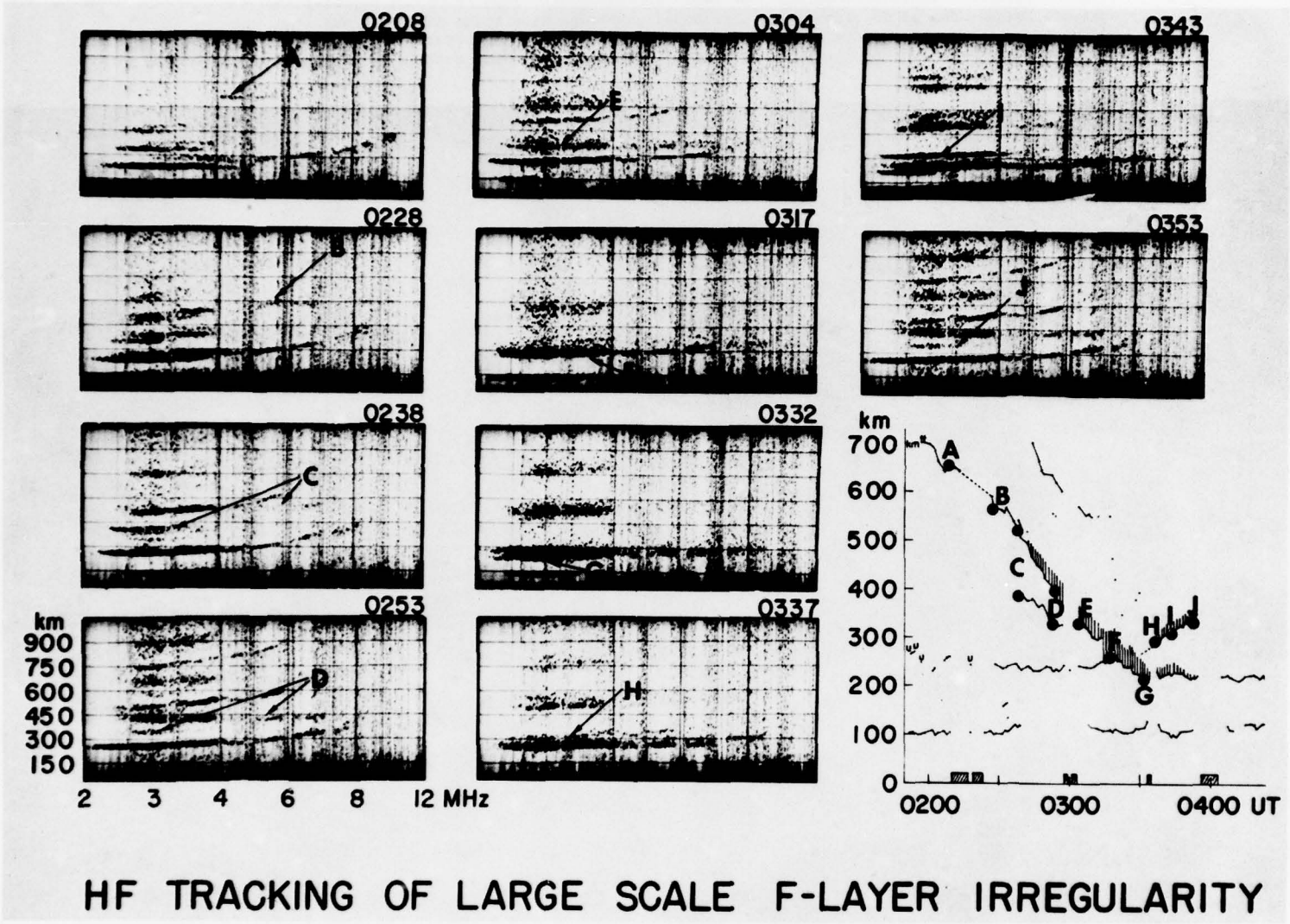


Figure 7 Airborne ionograms taken during the passage of the large airglow depletion (Figure 6) show backscatter echoes associated with the trailing edge (echoes A-E) and the leading edge (H-J) edge of the depletion. The minimum virtual height of the F-layer increases from 220 to 265 km as the depletion moves overhead.

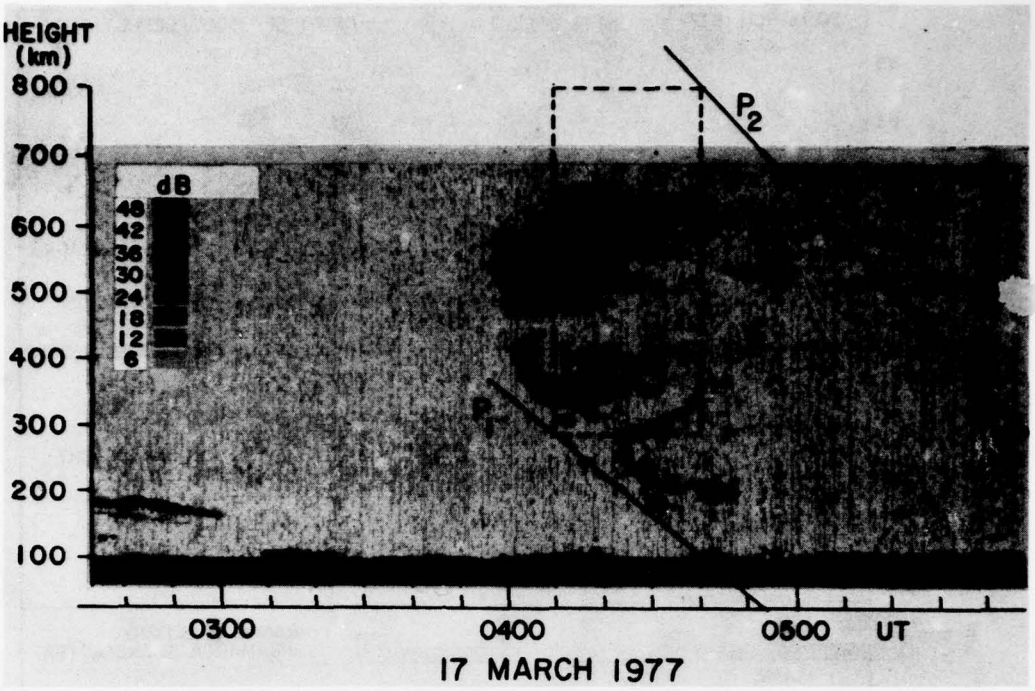


Figure 8 Range-time intensity map of isolated F-region disturbance passing over the Jicamarca radar. The dashed rectangle represents a cross section of the irregularity volume determined from airglow and scintillation measurements. P_1 and P_2 denote the trans-ionospheric ray paths at the beginning and end of the associated scintillation event.

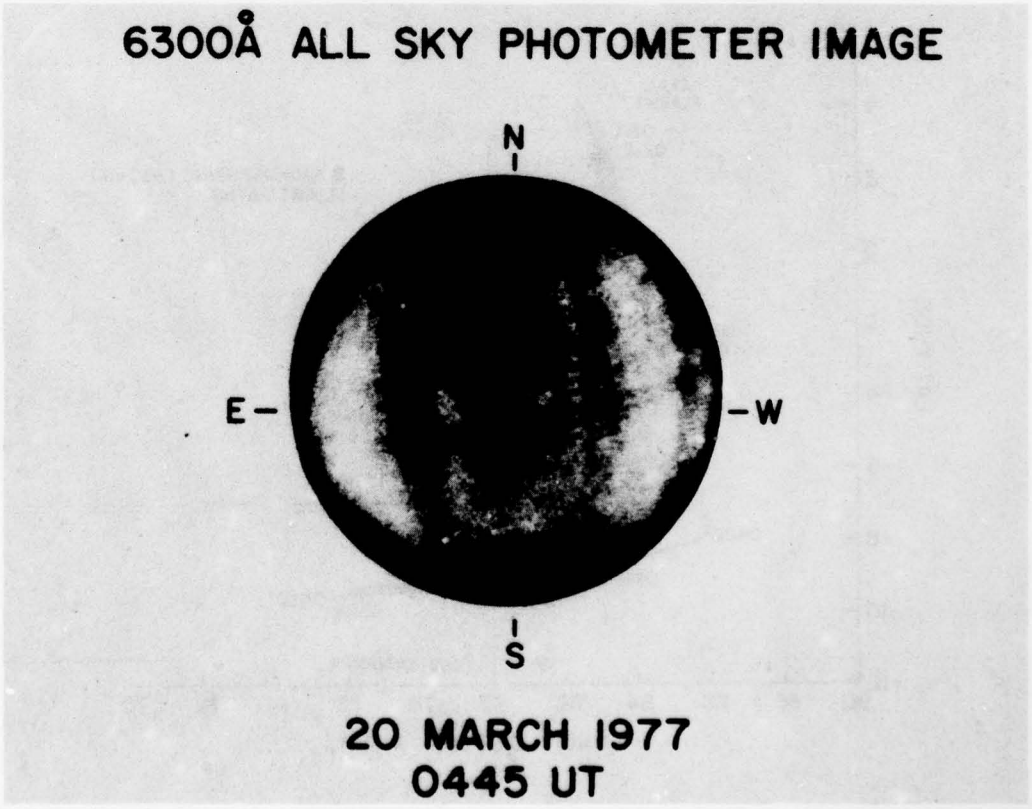


Figure 9 6300 Å All Sky Photometer image taken at 0445 UT on 20 March 1977 shows the existence of 4 well developed airglow depletions.

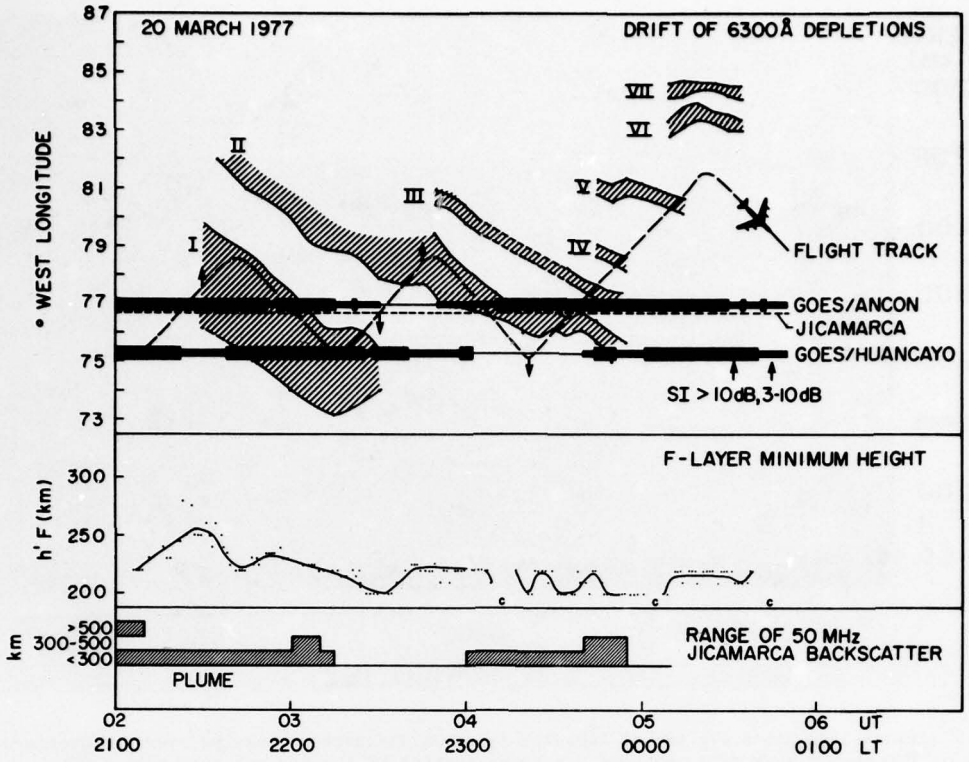


Figure 10 Time histories of eastward motion of multiple depletions in relation to scintillation measurements, F-layer virtual height changes and the occurrence of 50 MHz backscatter.

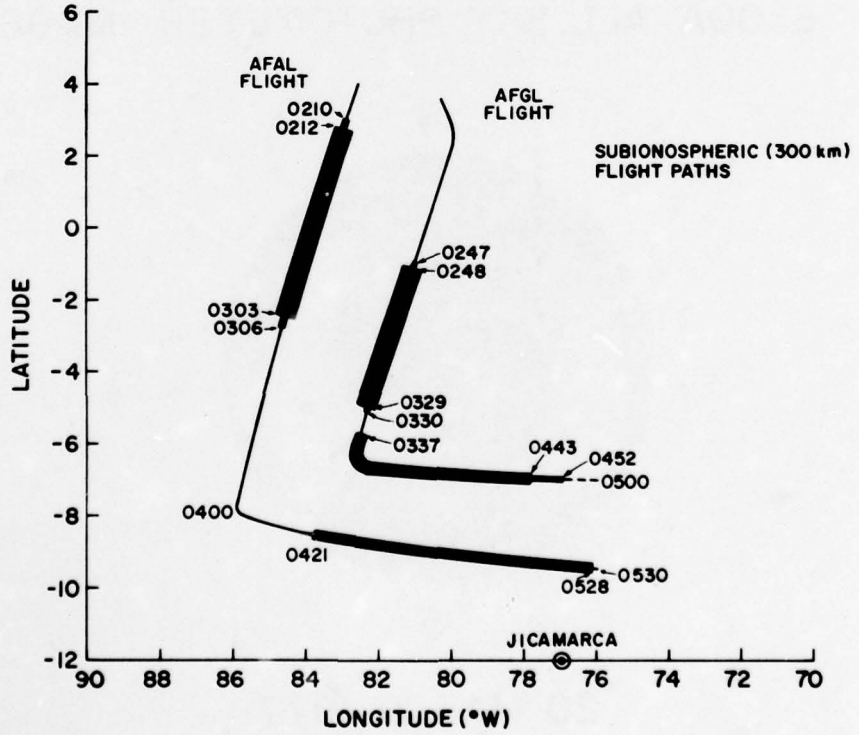


Figure 11 Scintillation measurements on the 249 MHz LES 9 transmissions, made on 17 October 1976, are superimposed on the sub-ionospheric (300 km intersection height of the ray path to LES 9) flight paths of the AFAL and the AFGL aircraft. The actual flight track of the AFGL aircraft is coincident with the sub-ionospheric track of the AFAL aircraft. The three increasingly wider bars relate to the scintillation index ranges <6 dB, 6 to 12 dB and >12 dB.

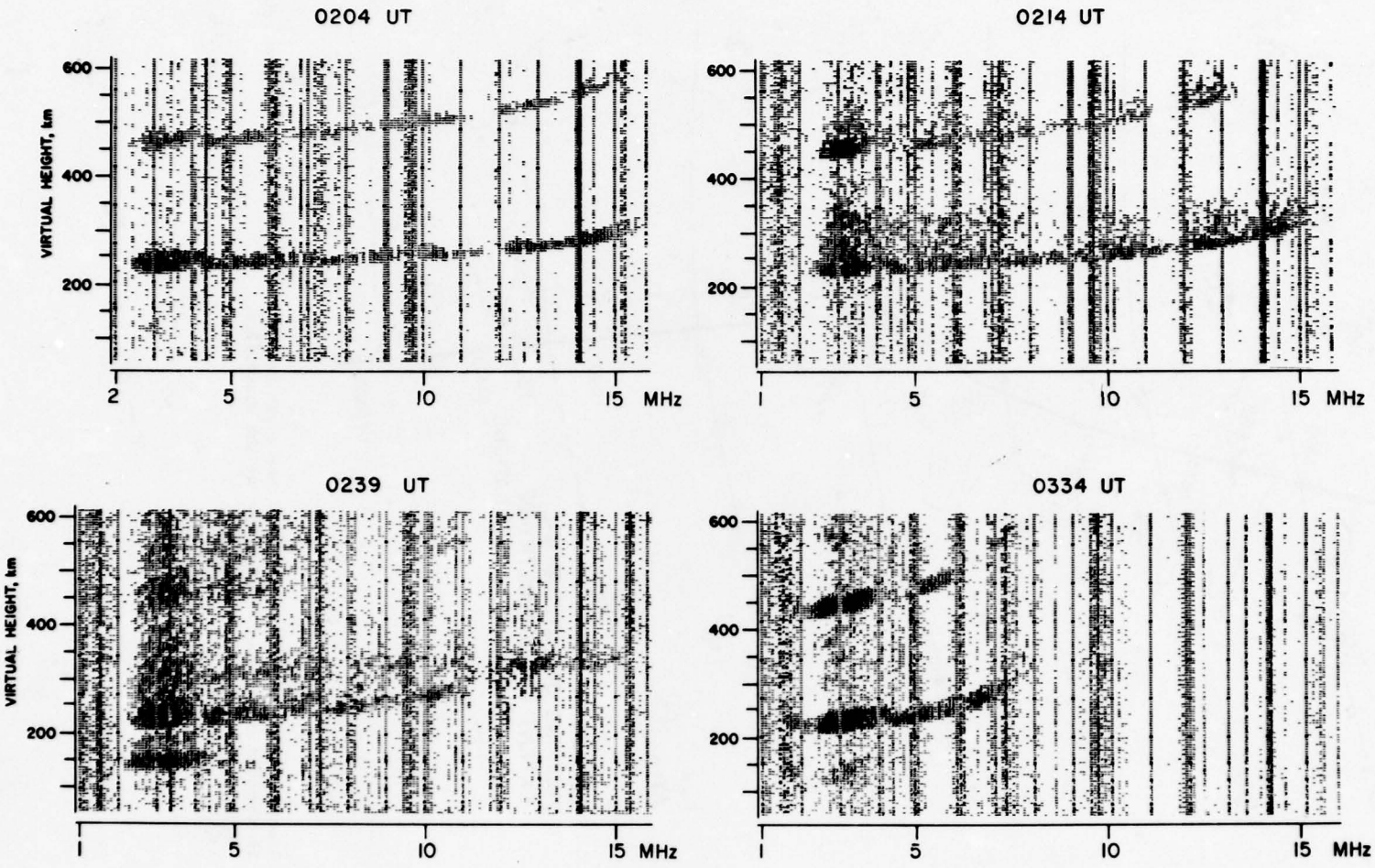


Figure 12 Ionograms taken by the AFGL aircraft along the AFAL sub-ionospheric track show fast development of range spread F (0214 UT) with the onset of scintillations (see Figure 11). After the scintillation event the ionosonde shows again undisturbed conditions (0334 UT).

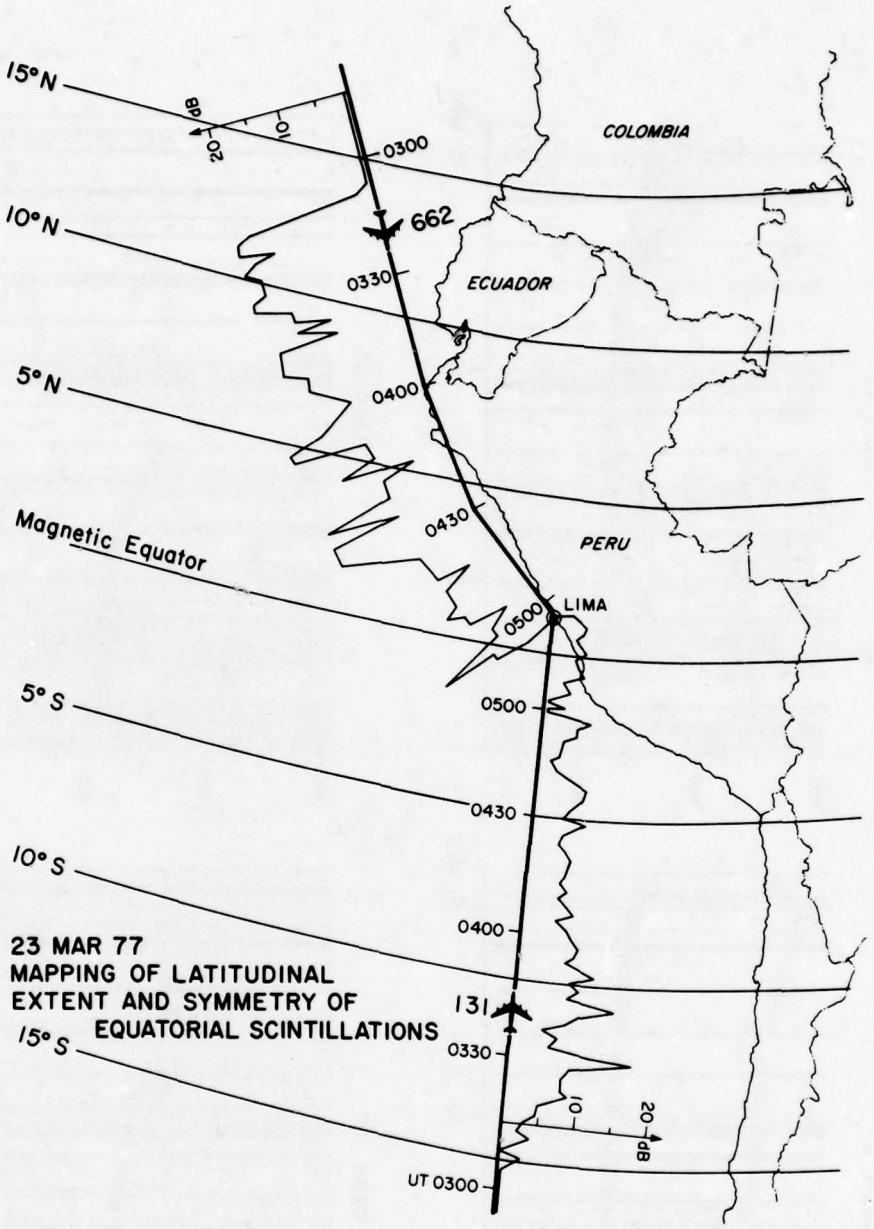


Figure 13 Simultaneous flights from north and from south into the scintillation region show actual symmetry in magnetic latitude of the location of the boundaries.

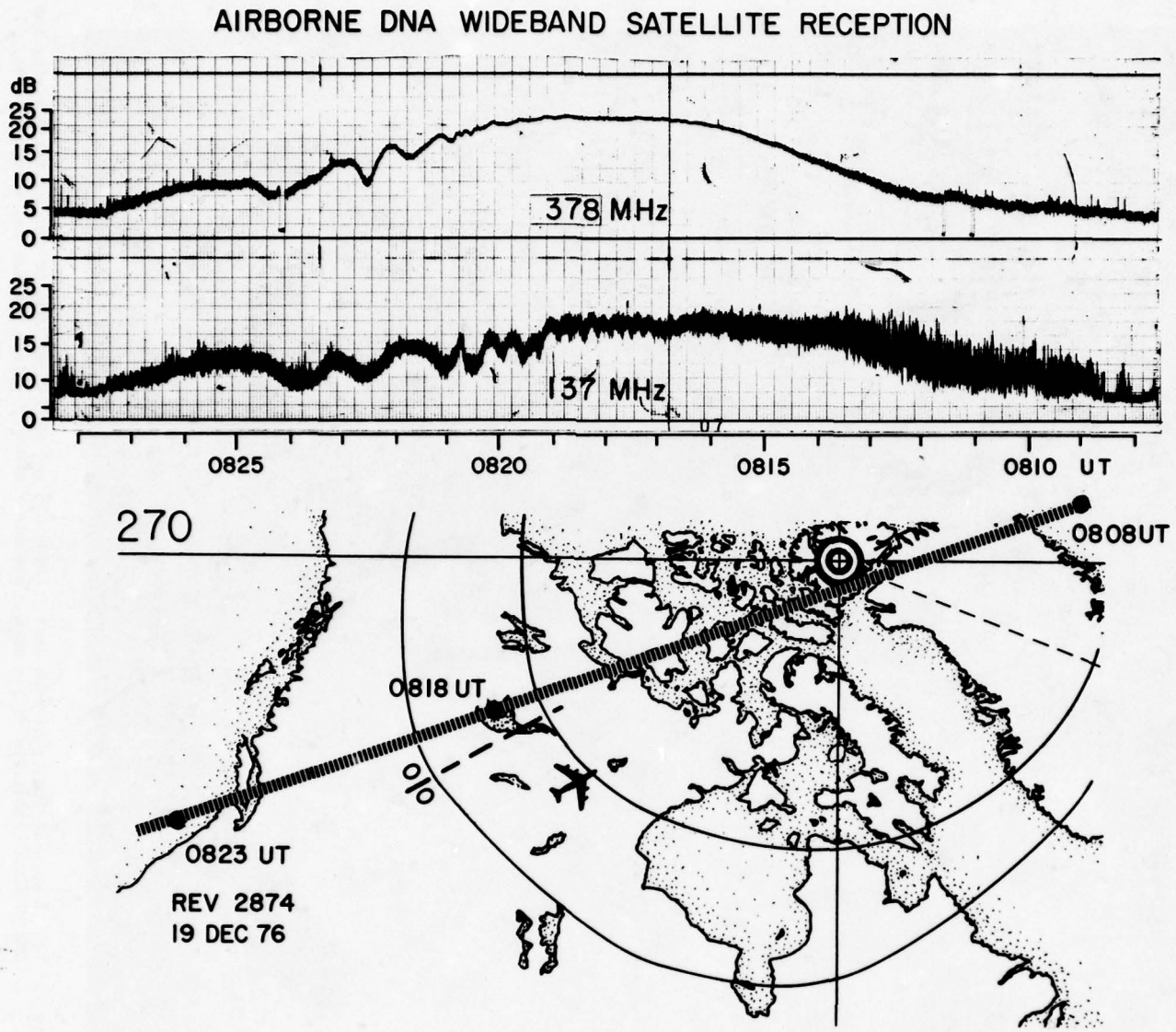


Figure 14 A WIDEBAND satellite path permits the mapping of polar cap, oval and subauroral effects on VHF and UHF signals. On this path scintillations were seen only inside the polar cap.

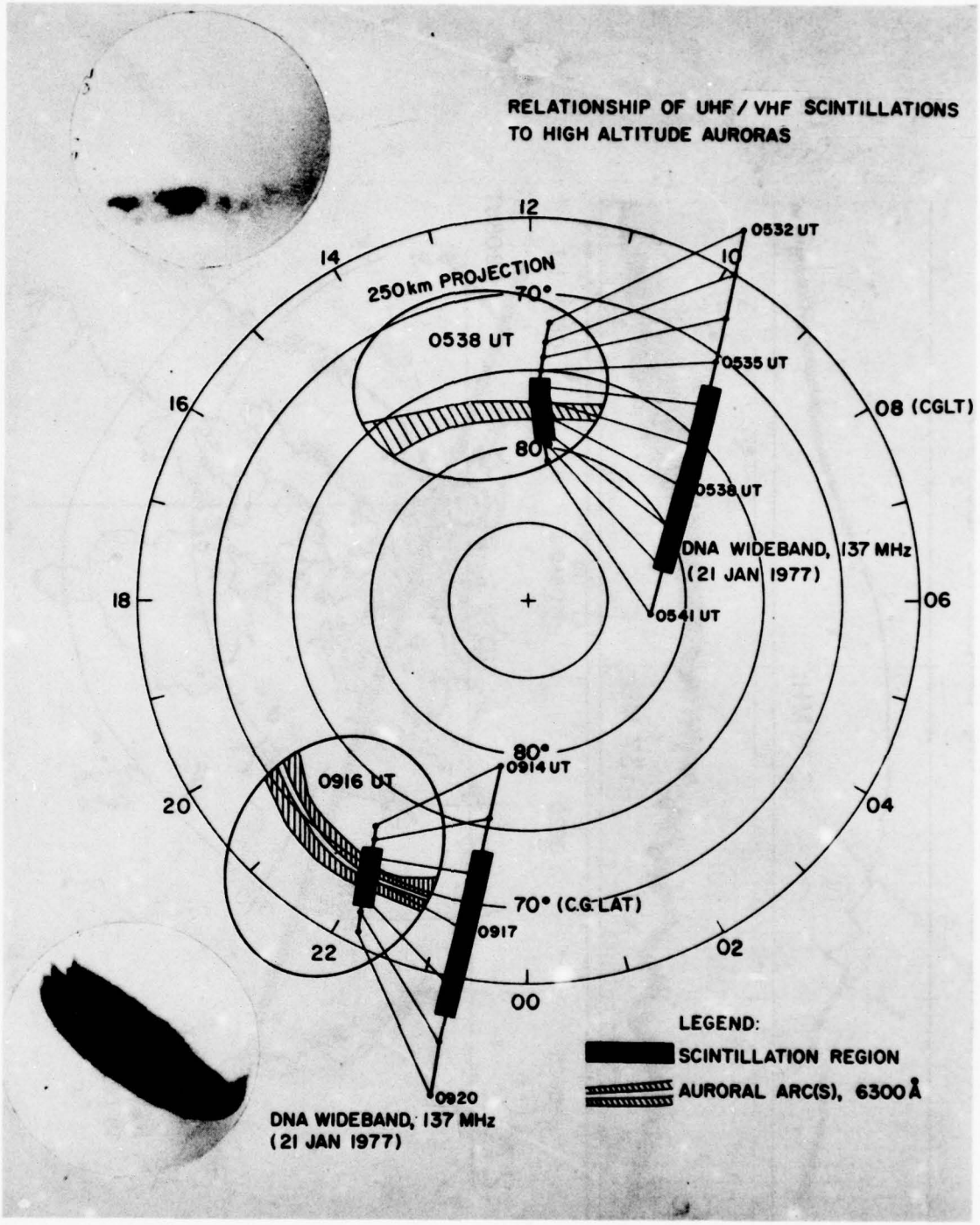


Figure 15 6300 Å auroral images of noon sector (0538 UT) and nighttime (0916 UT) aurora taken during two DNA WIDEBAND satellite passes show close correlation of F-layer aurora and scintillations.

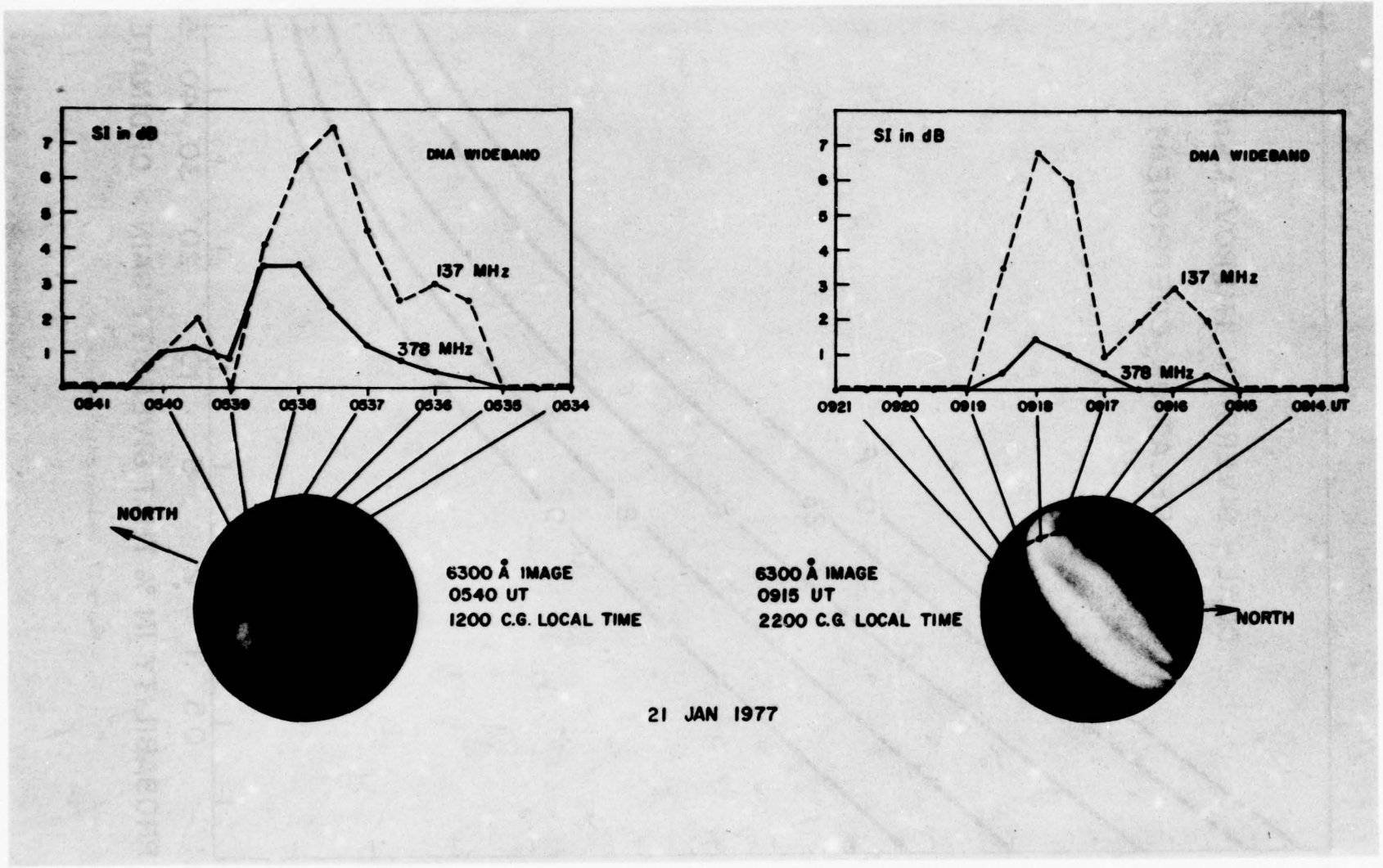


Figure 16 WIDEBAND satellite passes superimposed on the 6300 Å all-sky photometer images shown in Figure 15. The close correlation between the 6300 Å aurora and UHF and strong VHF scintillations is evident.

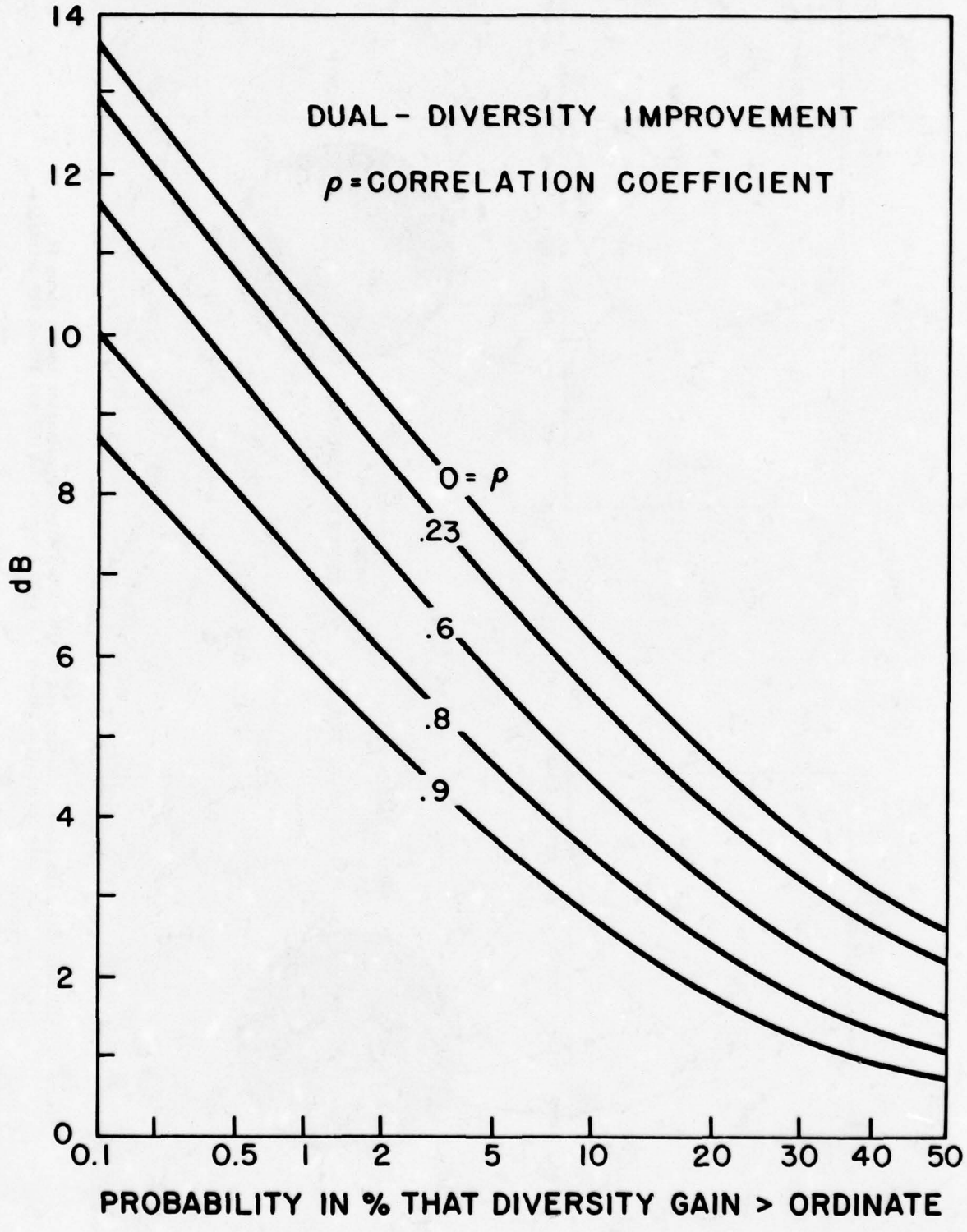


Figure 17 Dual-diversity improvement.

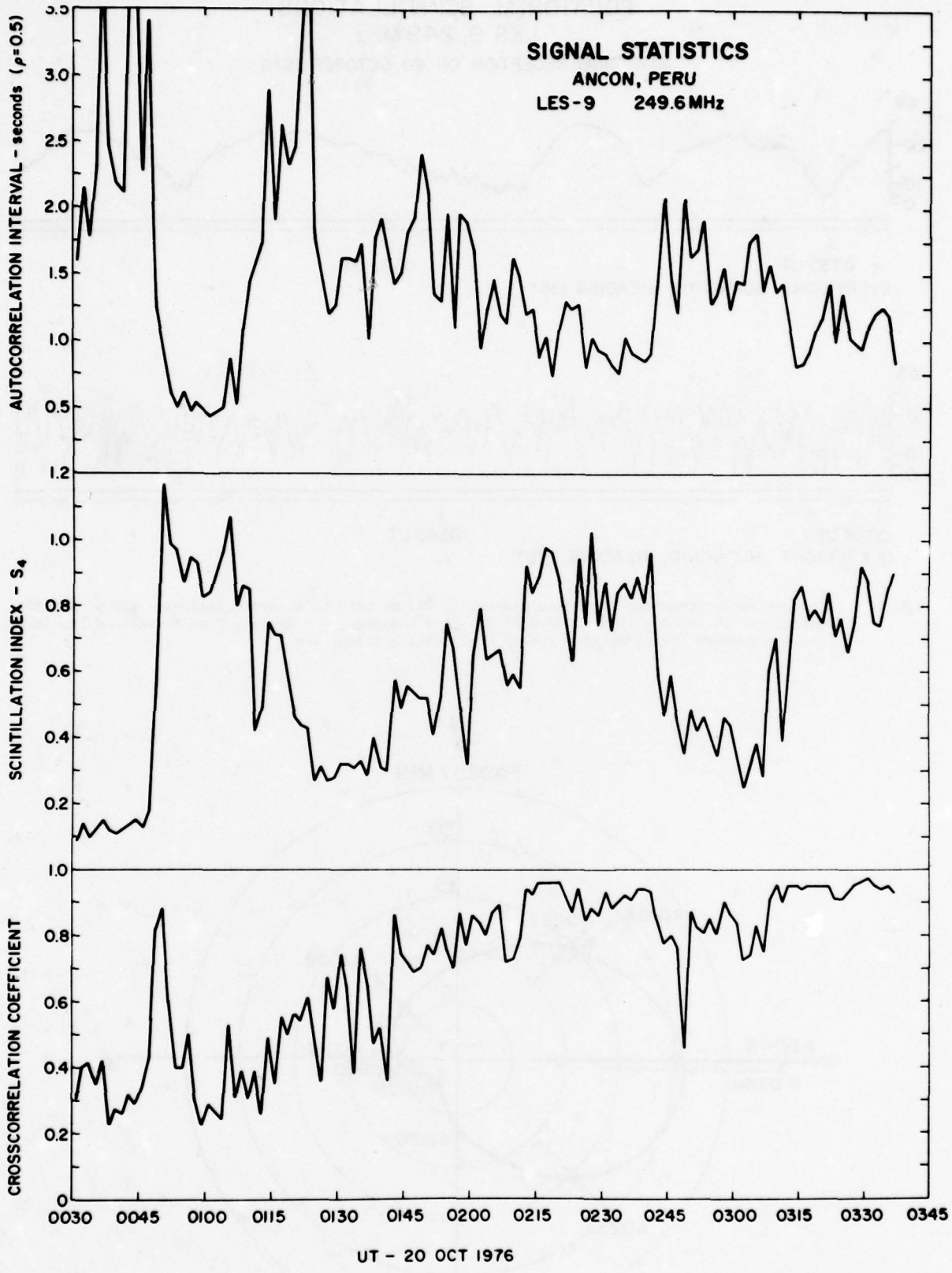


Figure 18 Variations of S_4, autocorrelation interval and cross-correlation coefficient for an extended period of scintillations on 20 October 1976.

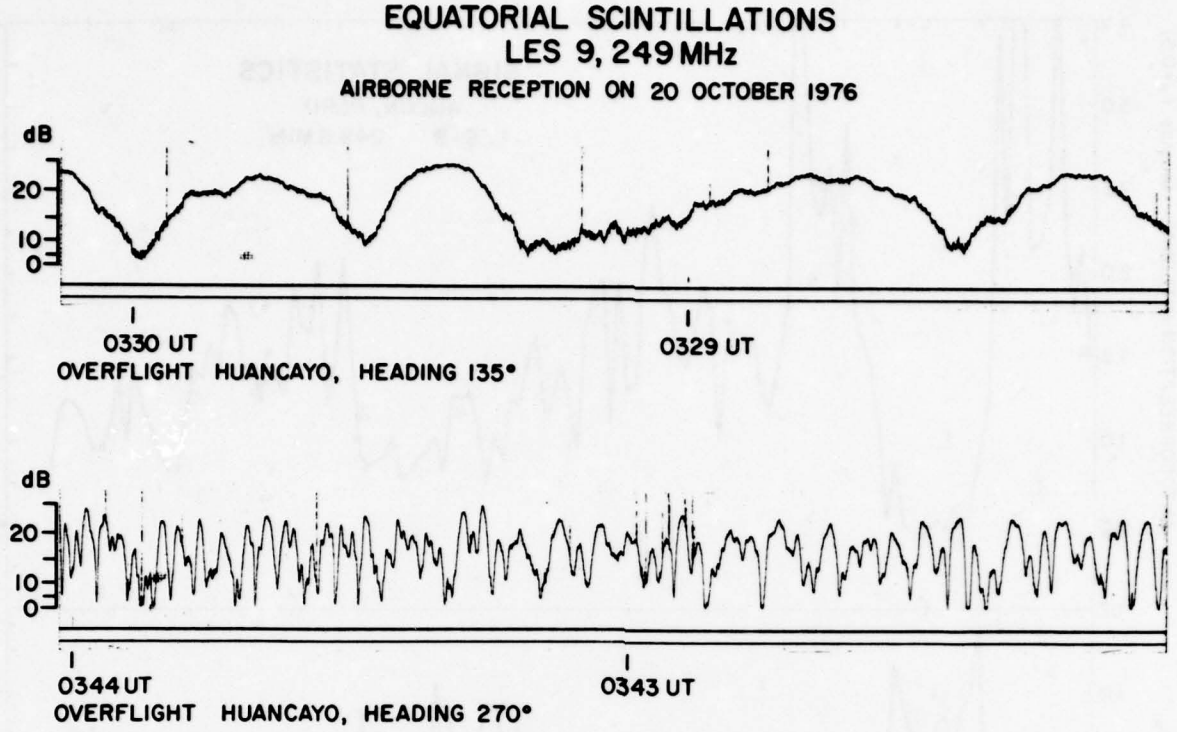


Figure 19 Airborne reception shows strong dependence of fading rate on aircraft heading. LES 9, 249 MHz, recorded on 20 October 1976, 0032-0354 UT. H = Huancayo, A = Ancon. Time = start of leg in UT. Dots are average fading rates for each 10-12 minute flight leg.

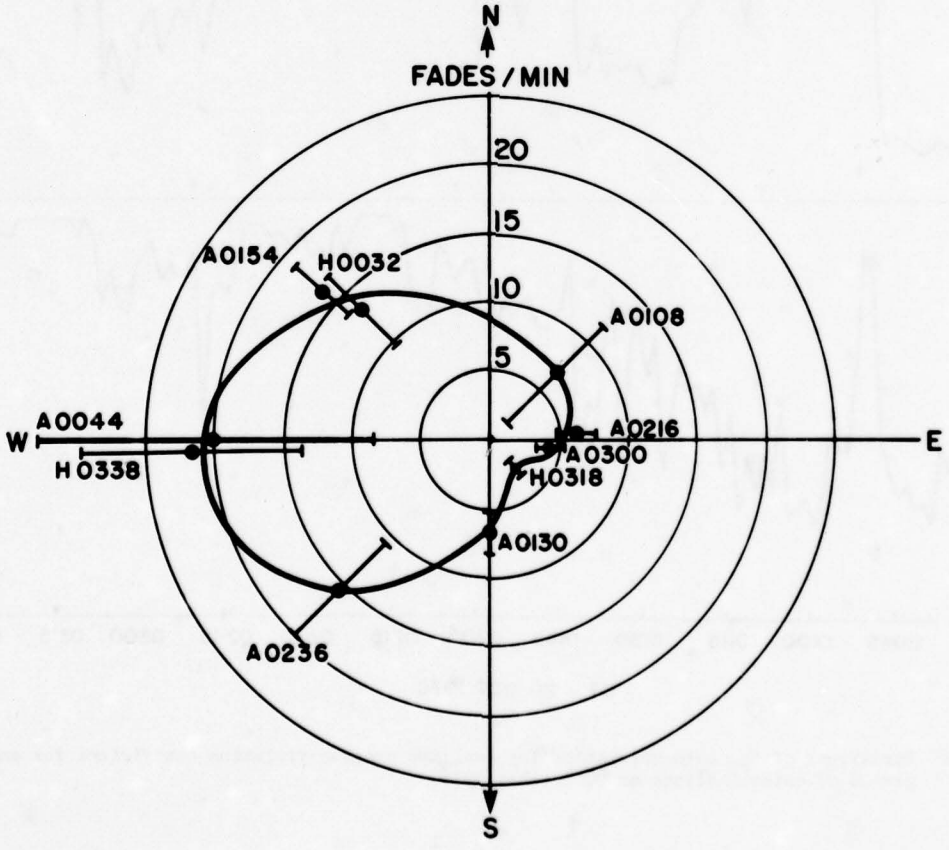


Figure 20 Fading rate as function of aircraft heading.

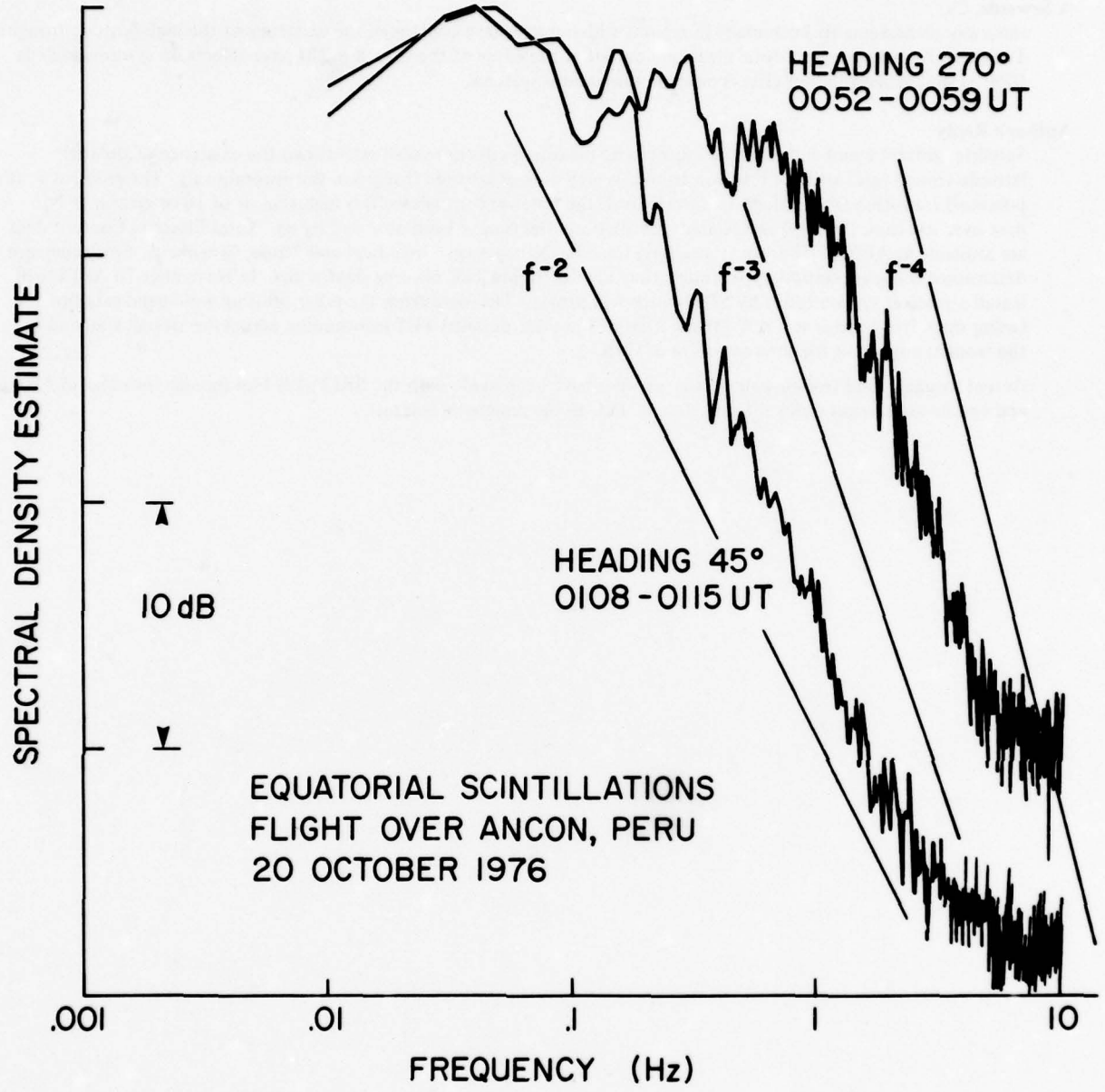


Figure 21 Sample spectra showing effects of aircraft heading.

DISCUSSION

A.Sewards, Ca

Have any measurements been made in regions which might have confirmed the existence of the high-latitude trough? The sharp rate-of-change of total electron content at the edges of the trough might have effects on systems such as GPS, as well as scintillation effects on communications systems.

Author's Reply

Satellite, ground based and airborne ionospheric measurements have well established the existence of the high latitude trough (mid latitude F-region trough as well as high latitude troughs in the auroral oval). The gradient at the poleward transition from trough to auroral oval (the poleward trough well) is high (factor of 10 or greater in N_e max over less than 100 km) and related scintillation effects have been observed by us. Total Electron Content data are available at AFEL/PHP from Goose Bay, Canada; Narssarsuag, Greenland and Thule, Greenland. Since they are determined using geostationary satellites, they are describing TEC on very slant paths. In November 78 AFEL will install a receiver system (built by SRI) which will provide TEC data from the polar orbiting wide band satellite (using three frequencies at – 400 MHz). This will provide detailed TEC information across the trough wall and in the trough, important for the correction of GPS.

Actual trough – and trough wall $N_e(h)$ profiles have been made with the SRI Poker Flat incoherent radar in Alaska and can be used to get quick information on TEC in the regions of interest.

Multipath Propagation Measurements by Doppler Technique

by

P. Form, R. Springer, Technical University
Braunschweig, Sonderforschungsbereich 58
in cooperation with H. Bothe, K. Klein,
DFVLR Braunschweig (FRG)

SUMMARY

In this paper some MLS-features are discussed with respect to multipath propagation. For illustration of specific multipath effects, a doppler-shift measurement technique with high angle resolution is presented, which needs simple antennas and equipment and offers easy interpretation.

1. INTRODUCTION

Multipath propagation and its influence on guidance accuracy got increasing importance during ICAO's MLS-competition. At first flight tests with single reflector screens /1/ should give information about multipath immunity of the different system proposals and, essential characteristics of the systems, like scanning rate, accord only to the need of aircraft guidance /2/. Later on MIT-Lincoln Lab. fed its experience on multipath propagation and a multipath simulation program into the process /3/. By merit of this institution this simulation program /4/ proved to be an efficient tool both for system comparison and system development with respect to multipath propagation effects /5/.

These programs also were used in the FRG for simulation of the destinated ICAO-airport scenarios, flight pathes and system simulation /6, 7, 8/.

The german MLS-system proposal is DME-derived /6, 9/. That means the ground station measures the amplitude and phases of the airborne interrogation pulses received by multiple antennas and receivers. These amplitudes and phases deliver aircraft azimuth and elevation information by an iterative computer evaluation. Both informations are transmitted to the aircraft by additional time coded reply pulse pairs. Because of the chosen straight forward L-Band-hardware technique and the digital signal processing the german system could easily be simulated with a high degree of confidence.

The MIT-multipath simulation program /5/ is the most complete one and contains a couple of details. Nevertheless multipath propagation is very complex and additional information about multipath propagation in real environments and under real circumstances was desirable. For this purpose a L-Band multipath propagation measurement technique was developed, which supports field tests of DLS /10, 11/ and discussions about extended applications of MLS in strong multipath environments such as mountainous areas /12/.

2. MULTIPATH PROPAGATION AND SYSTEM FEATURES

Multipath propagation means transmission of a signal not only by direct signal path DP (Fig. 1), but also by several additional pathes. Additional pathes are given by the reflecting ground (GR), buildings, aircraft, hills or mountains and other reflection objects in the vicinity of the system. For instance a building can generate not only one, but mostly two or up to four different pathes (Fig. 1 /5/). All additional path lenghts are longer than the direct path producing different delayed and phase shifted signals with respect to the direct transmitted signals. Due to the small wavelength (L-Band: 30 cm; C-Band: 6 cm) even a small change of aircraft position causes strongly and differently changed phase

shifts. Therefore detailed system comparison needs same scenarios and flight pathes, which can be guaranted only by simulation, but not by succesive flight tests.

By aircraft motion the direct path signal and all additional path signals get a different doppler shift D due to the path angle α with respect to the actual speed vector of the aircraft (Fig. 2). Even in a simple continuous wave transmission system without any modulation aircraft motion generates a maximum spectrum width M, which depends on the carrier frequency fc and the speed v of the aircraft. For approach speeds of 145 Knots the maximum spectrum width M is

$M = 2\ \Delta f_{max}$ = 500 Hz for a L-Band system (DLS) or

M = 2500 Hz for C-Band systems (TRSB, DMLS).

In an extreme case for instance the direct signal gets a maximum positive doppler-shift while a reflection behind the aircraft gets a maximum negative doppler-shift. Both signals interfere with the frequency M of 500 Hz (L-Band) or 2500 Hz (C-Band) respectively. From this point of view scanning systems should have appropiate scanning rates /13/, which would be clearly higher than presently intended /2/. Limited coverage of the ground station and of the airborne antenna naturally decrease the maximum spectrum width M, but operational aspects make this solution less attractive.

Like EVANS /13/ pointed out, by smaller scanning rates certain interference frequencies are aliased down to zero frequency or to one of the frequencies in 0,25 Hz - 0,5 Hz range which are particularly bad for autopilot couplers.
This aliasing effect generates so called "bad angle-orientations", at which errors cannot be averaged (Fig. 3).

Two tools can be distinguished for multipath signal selection:

1.) multipath angle discrimination by large antenna apertures on the ground,

2.) multipath time delay discrimination by impulstransmission.

While TRSB and DMLS depend on angle discrimination only, the DLS-proposal uses both techniques /4/. Because amplitudes and phases of the DME-interrogation pulses are measured during the rising first pulse, all later received multipath signals are attenuated or eliminated. If the 80% point of the specified DME-Pulse (Fig. 4) is measured, multipath time delay discrimination accords to Fig. 5. Locations of the same multipath delay are ellipsoides with the aircraft in the one focus, the groundstation in the other (Fig. 6 /12/). By this effect a limited transmission cell is established, which ends closely behind the groundstation and the aircraft and commonly does not touch the ground in the vicinity of the aircraft. If a too small sample rate generates bad angle orientations, these orientations only exist within these ellipsoid-shaped transmission cells of DLS (Fig. 7). Because reflectors beside and behind the aircraft can not produce high interference frequencies, the theoretical necessary sample rate can be clearly decreased.

The reflections of a building for instance occur in sectors only, which are marked in Fig. 8 by edge rays assigned by numbers. While approach and landing the aircraft passes these sectors within certain time intervals of limited duration. This limited duration commonly further reduces the need for high scanning rates, if prediction type filtering is used (Tracking Filter). But the efficency of such filtering is limited by the flight path and the multipath distribution versus time.

3. MULTIPATH MEASUREMENT EXPERIMENTS

For further investigation and discussion of the relations mentioned above two types of multipath measure-

ments were made in different real environments:

1.) Continuous wave transmission tests for illustration of angle dependend multipath propagation,

2.) Impuls transmission tests for illustration of distance/time delay dependend multipath generation.

3.1 Continuous wave transmission doppler-shift measurement technique.

For this type of field tests a continuous wave signal of a 1 GHz-carrier and of a high spectral purity and stability was transmitted by the test aircraft. Test aircraft and airborne antenna diagrams for different attitudes are shown in Fig. 9. Inspite of the engines and the fixed landing gear the antenna-diagram ensures a good circular coverage.

According to Fig. 10 the transmitted signal passes in the ground receiver three superimpositions, which superimpose the signalfrequency down to 1000 Hz, which can easily be stored by a taperecorder. Both signal generation on board and superimposition on the ground are stabilisized by rubidium clocks. After tests the taperecorder signal can be evaluated versus time and frequency by filtering and spectrum analysis. The ground antenna was a vertical $\lambda/4$-Monopol with 360° degree coverage.

Aircraft motion generates a doppler-shift diagram corresponding to Fig. 11, which illuminates reflecting obstacles and also the ground station with different frequencies due to their direction angle α with respect to the actual speed vector of the aircraft. By narrow band filtering and by spectrum analysis the amplitude and frequencies of direct path and multipath signals can be distinguished.

3.2 Continuous wave transmission field test results in a flat environment with building reflections

Several continuous wave transmission field tests were taken in Braunschweig Airport (see Fig. 8). In this environment the reflections of buildings within the airfield dominate very clearly in this almost flat environment. They are concentrated to the runway, but that did not affect the study of multipath propagation, if the testaircraft overflies the runway in different low altitudes after an eastbound straight in approach. Representing an Azimuth-MLS-Station the experimental ground station was located in this case at the stop end east of the runway 09.

The lower part of Fig. 12 shows the received voltage versus time including the whole doppler-spectrum. The three diagrams above contain versus the same time narrow band filtered components of the spectrum. Each of these filtered components represent reflections, which are generated in a certain angle sector with respect to the actual speed of the testaircraft. Due to the filter frequency and band width of 30 Hz these angle sectors are 40 ± 6 degreee, 50,5 ± 5 degree and 59 ± 4 degree. Additional dashed lines in the diagram and their number indicate the aircraft passing multipath sector edge rays of Fig. 8 and the identification letter indicates the corresponding building. The concrete-glass building L (Luftfahrtbundesamt), the mainbuilding M and the hangar H cause reflections, which represent nearly 100% of the sum signal in that moment, while the one floor terrace T, building G and others generate 80%, 25% and smaller reflections. Inspite of the large amplitudes, however, the duration of reflections is short. In peculiar a segmented building like building L generates a couple of large, but extreme short spikes.
Additional plots (Fig. 13) show better reflection angle separation by 10 Hz-bandwidth filters corresponding to an angle-bandwidth of ± 1,5 and ± 2 degrees. These sharp filters still follow the time functions in a sufficient manner and seperate the earlier reflection of the left part and the later reflection of the right part of buildings M and L each.

Multipath filtering and plots versus time can be supported by spectrum analysis. In this case the tape

recorded signal is sampled and stored during a certain time interval ti. Plots Fig. 14 show both the FFT-Analysis and the stored time function (sum signal) taken in selected moments. In Fig. 14 plots a) show the aircraft in a relative multipath free region close to the airfield edge (▽, compare Fig. 8, 12). Only some propeller modulation side bands are to be seen. In plot b and c aircraft passes the multipath sector of hangar H, plot d corresponds to the sector of the terrace T and the plots e) and f) to the sector of building L respectively.

The chosen time interval ti = 0,2 sec as the basis of Fast Fourier Transformation makes the results approximately true, because during this short time a multipath signal seems to be rather stationary. The multipath signal amplitude can be considered only in an absolut scale. A relation to the direct signal (DP) cannot be found here, because the frequence resolution of 5 Hz of the spectrum analysis is still too small for a separation of the direct signal and the ground reflexion (GR). This problem only can be met by a further improved evaluation technique or by an appropiate ground antenna without illumination of the ground.

3.3 Multipath measurements in a mountainous site.

The same equipment and the FFT-Analysis was utilized for multipath measurements at Salzburg Airport (Austria, 1401' MSL). This airport is surrounded by mountains of considerable height in the West (2500' MSL), south (6000' MSL) and east (2500' MSL) in a 2 to 5 NM distance. Because of this environment initial approach direction is from north (160°) irrespective of wind.

Final approach to RWY 16 gets guidance by an ILS with limited coverage (± 40 degree clearance, ± 10 degree precision) and without any back course information. The specified decision height 700' ensures a safe pull up still well clear of the valley (Fig. 15, 16).

For approach to the RWY 34, aircraft at first follow ILS guidance till SI NDB and then switch to a 130 degree visual circling approach (see Fig. 16). In a steady descent this circling path leads over the city in a 0,8 NM radius turn to RWY 34.

This airport represents a type of sites, where higher MLS-accuracies alone will not permit to decrease the decision height. If a future MLS would decrease the decision height (700') for RWY 16, it has to give overshoot guidance according to the reciprocal circling approach procedure; a future MLS for RWY 34 approaches has to offer circling approach- and descent guidance between SI NDB and the RWY 34 threshold. In such environments increased MLS-coverage (360 degree) and omnidirectional airborne antennas are necessary. Because of the banked descent the airborne antenna must provide sufficient transmission capability also in banked attitudes /14/.

In the next Figures 17, 18 and 19 each three signal plots were taken in intervals of one or two seconds distance /15/. In these moments test aircraft is in the flight positions C, D or E (compare Fig. 16) while completing the last quarter of the circling turn to RWY 34.

Each plot shows signal amplitudes versus frequency on the left side, signal versus corresponding time to the right. In each left plot highest amplitude represents the direct path signal sometimes accompanied by ground reflections of very small difference. The other amplitudes represent multipath signals generated in front of the aircraft (doppler-shift + 240 Hz), aside the aircraft (doppler-shift 0 Hz) and behind the aircraft (doppler-shift -240 Hz). While test aircraft continues the turn the next spectral plots of Fig. 18 and 19 contain reflection families, which also turn back (to lower doppler-shifts). In general 10% to 30% reflections are typical in the Salzburg area, if omnidirectional airborne antenna patterns (Fig. 9) are utilized. Front beam horn antennas, however, would amplify high frequency reflections and decrease direct path signal at the same time during circling or curved approaches and departures.

Reflections are smaller compared with building reflections, however, since they occur in families generated in various directions at the same time, the sum signal gets deep and very fast interference modula-

tions like corresponding right plots illustrate. These plots also demonstrate the need for high sample rate, if scalloping frequencies should be avoided in omnidirectional scanning systems without impulse transmission techniques and multipath time delay discrimination.

4. IMPULSE TRANSMISSION TESTS.

In completing the illustration of multipath effects airborne DME-Interrogation pulses (ILS, DME 109,0; channel 36X, 1060 MHz; antenna and pattern see Fig. 9) were observed on the ground, while aircraft approaches to RWY 16 and RWY 34 (circling) of Salzburg airport /15/. Some photos (Fig. 20) show short and long multipath delays sometimes generating pulse distortions. Frequently chains of reflections of smaller amplitudes and delay times in the range of 2 to 20 µs can be observed. These results illustrate the use of impulse transmission techniques in new guidance systems for all purposes and class of services.

5. CONCLUSION

Experimental doppler-shift measurement technique gives some detailed information about multipath propagation and interference.

Utilizing real airborne MLS-antennas and simple test- and recording equipment, evaluation delivers complete information about signal in space versus time and frequency and high multipath angle resolution and identification of reflecting sources in real environments. Since the influence of attitude dependend airborne antenna pattern and complex terrain is included, the results give useful support to the simulation work.

REFERENCES

1 ICAO AWOP, WGA Report London 1974

2 RTCA-SC 117. Standard Adjustment Criteria for Airborne Localizer and Glide Slope Receiver. 3-14-63, RTCA-Document Do 117.

3 ICAO AWOP Report Melbourne 1975

4 MIT-Lincoln Lab. ATC 44 WP-5023. 30. July 1975:
R. Burchsted, J. Capon, R. Orr. Preliminary Description of MLS-Simulation Program-Version 1

5 J.E. Evans, R. Burchsted, J. Capon, R.S. Orr, D.A. Shnidman, S.M. Sussman. MLS-Multipath Studies. MIT-Lincoln Lab. Proj.Report ATC 63, prepared for FAA. FAA-RD-76-3. 1976

6 DLS-Documentation for ICAO presented by the FRG September 1975.
App. B 4 Multipath Simulation Results (H. Ecklundt)

7 H. Ecklundt. Simulation of Multipath propagation for DLS. Mitteilungen aus dem Sonderforschungsbereich 58 "Flugführung" der Technischen Universität Braunschweig, M1, Sept. 1976

8 H. Ecklundt, Kabiersch. WP-100 AWOP-Meeting Montreal, April 1978.
Computer Simulation Results for Multipath Performance Comparison (presented by the FRG)

9 ICAO AOWP-Meeting Montreal, April 1978, WP-7 and Addendum No. 1. Federal Republic of Germany proposal for a New Non-visual Precision Approach and Landing Guidance System DLS

10 DLS-Documentation for ICAO presented by the FRG, Sept. 1975. App. B5, Multipath measurements in the test area (P. Form, R. Springer)

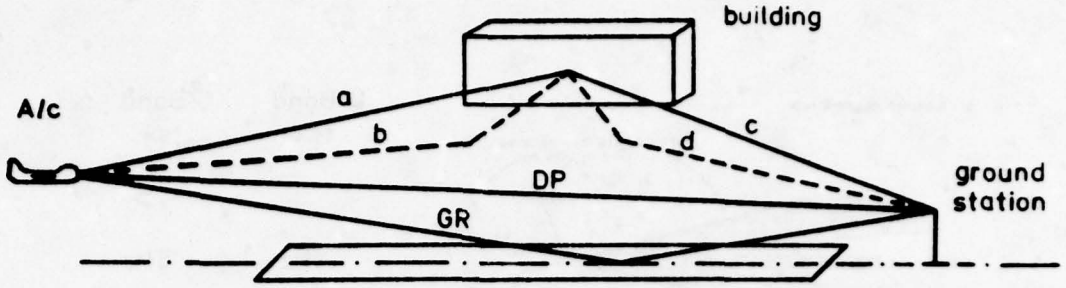
11 P. Form, R. Springer, H. Ecklundt. The DLS-Test-Airport - a non clean environment. WP AWOP WG 1, 6th meeting The Hague, 5-16. Juli 1976, presented by the FRG

12 P. Form. Multipath immunity of MLS in mountainous sites. WP. AWOP-WG. A, 7th meeting London, Nov. 1976, presented by the FRG

13 J.E. Evans, MIT-Lincoln Lab., System Selections Considerations 1974

14 P. Form, D. Brunner. The Need for integrated navigation systems in the TMA. WP 56, ICAO AWOP-meeting Montreal, April 1978, presented by the FRG

15 P. Form, R. Springer. Field tests for multipath propagation measurements in mountainous sites. WP 101 ICAO AWOP-meeting Montreal, April 1978, presented by the FRG.



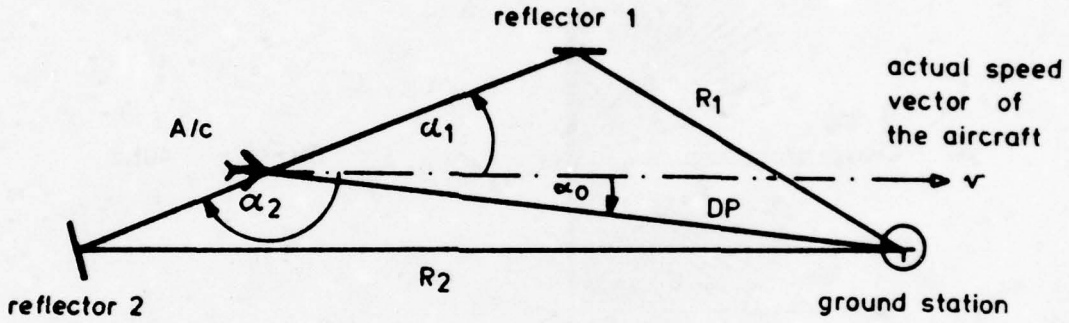
DP direct path

GR ground reflection

building reflection pathes:

1: a - c
2: a - d
3: b - c
4: b - d

Figure 1

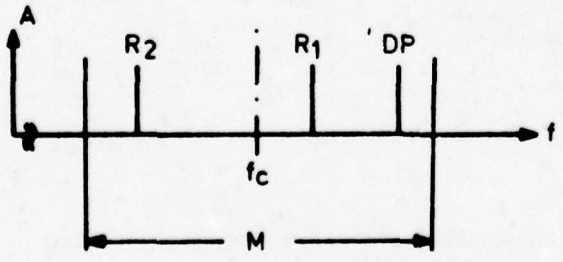


doppler shift of each signal by A/C-motion:

direct path $D_0 = \frac{f_c}{c} \cdot v \cdot \cos\alpha_0$

reflection R_1 $D_1 = \frac{f_c}{c} \cdot v \cdot \cos\alpha_1$

reflection R_2 $D_2 = \frac{f_c}{c} \cdot v \cdot \cos\alpha_2$

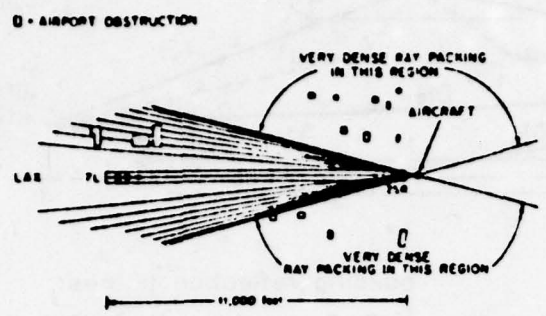


maximum spectrum width

$M = 2\frac{f_c}{c} \cdot v$

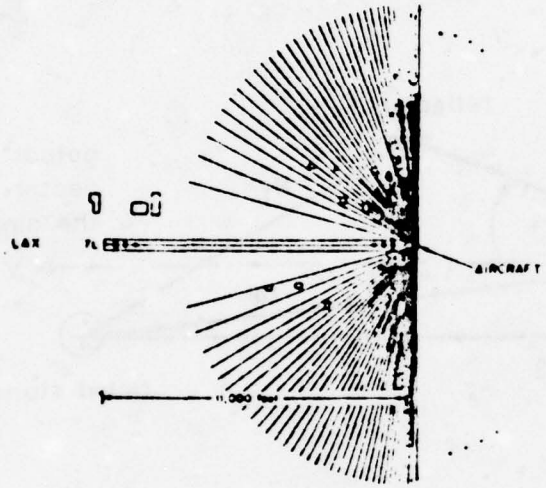
(f_c = carrier frequency)

Figure 2

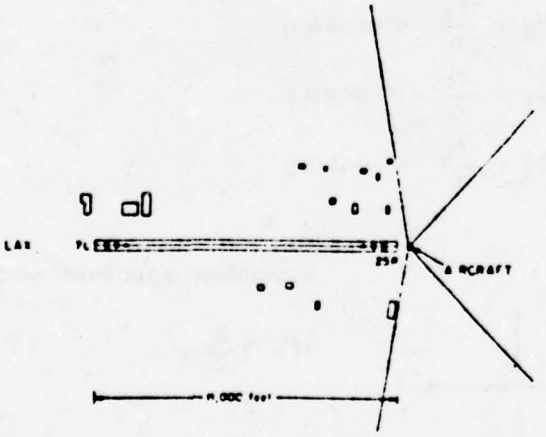


L-Band fs= 1Hz

C-Band fs= 5Hz



8Hz 40Hz



200Hz 1000Hz

Fig.3 Bad angle orientations as a function of scanning/sample rate fs (aircraft speed 140 knots)
Reference J.E.Evans, MIT LL/13/

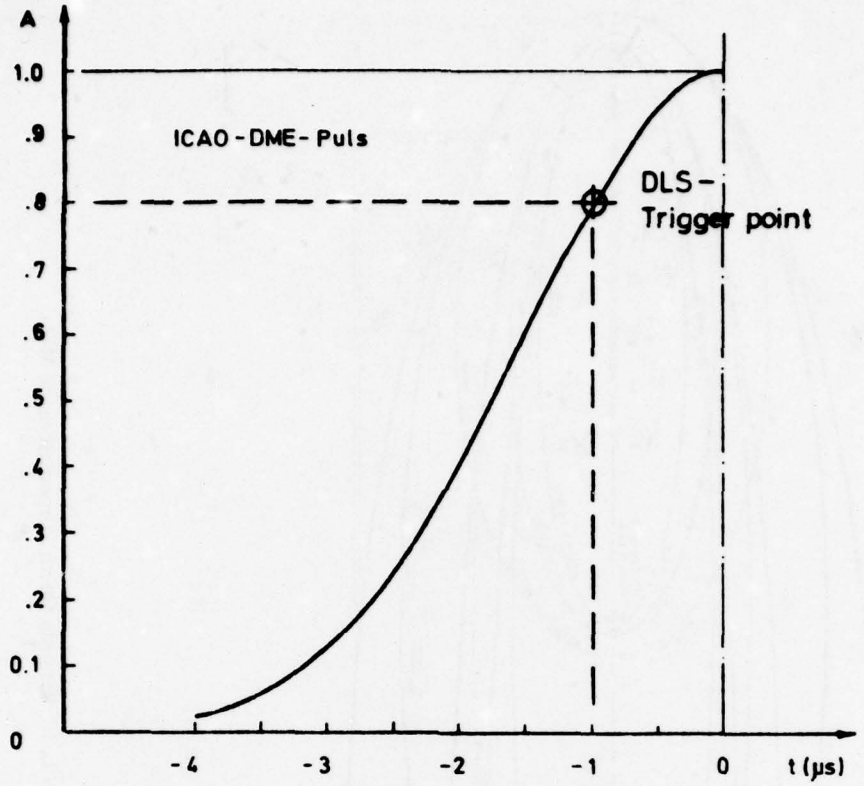


Figure 4

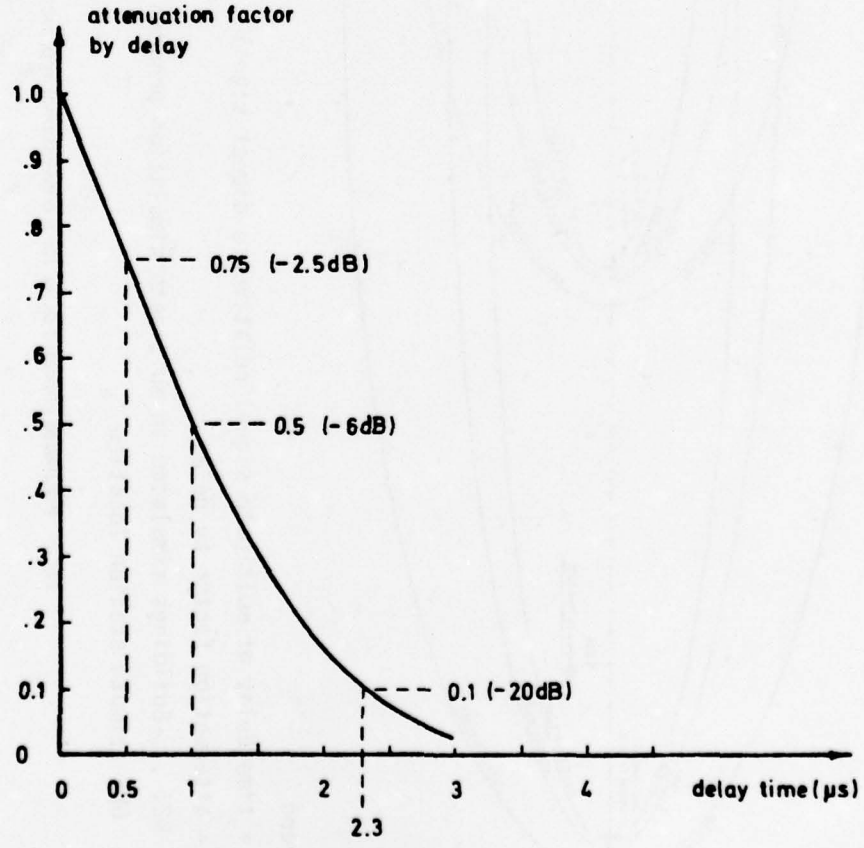
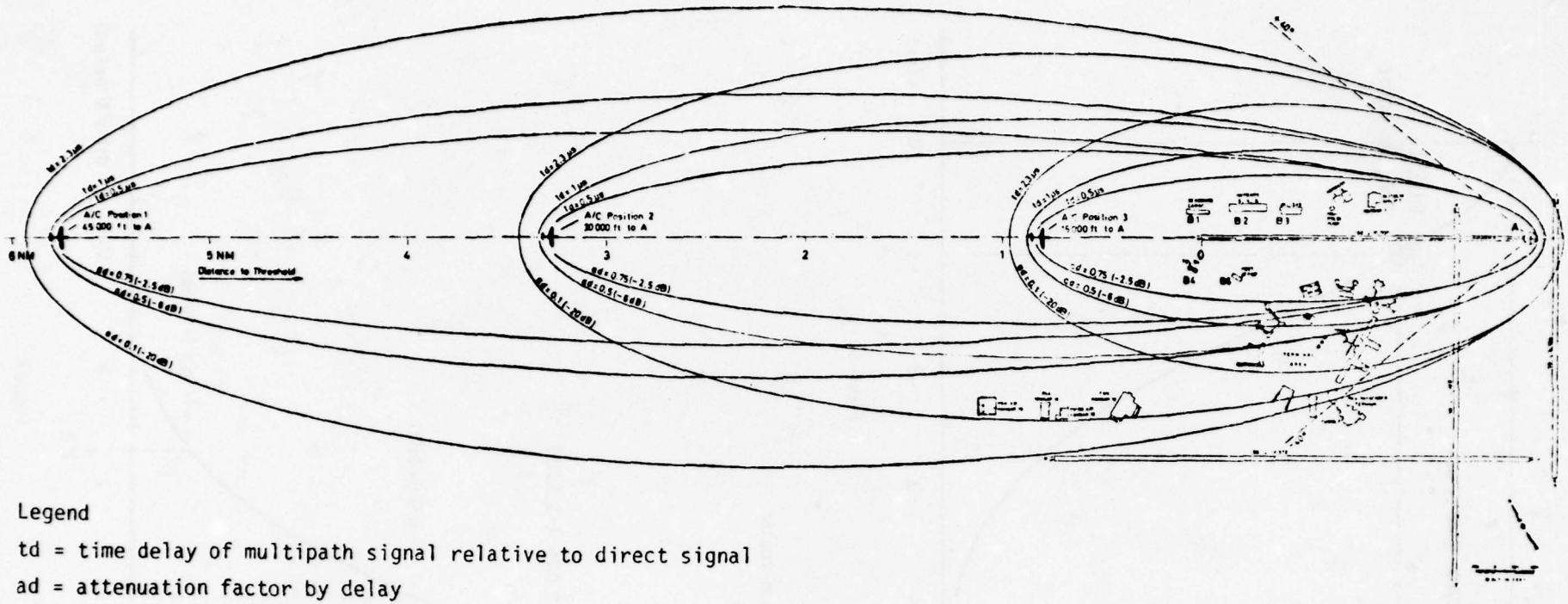


Figure 5



Legend

td = time delay of multipath signal relative to direct signal

ad = attenuation factor by delay

B1, B2, ... Buildings simulated in multipath simulation program (MIT)

A DLS-azimuth station location

Fig.6 Figures of constant delay for a conventional approach to runway 13 L, J.F.Kennedy Intern. Airport (scenario MIT)

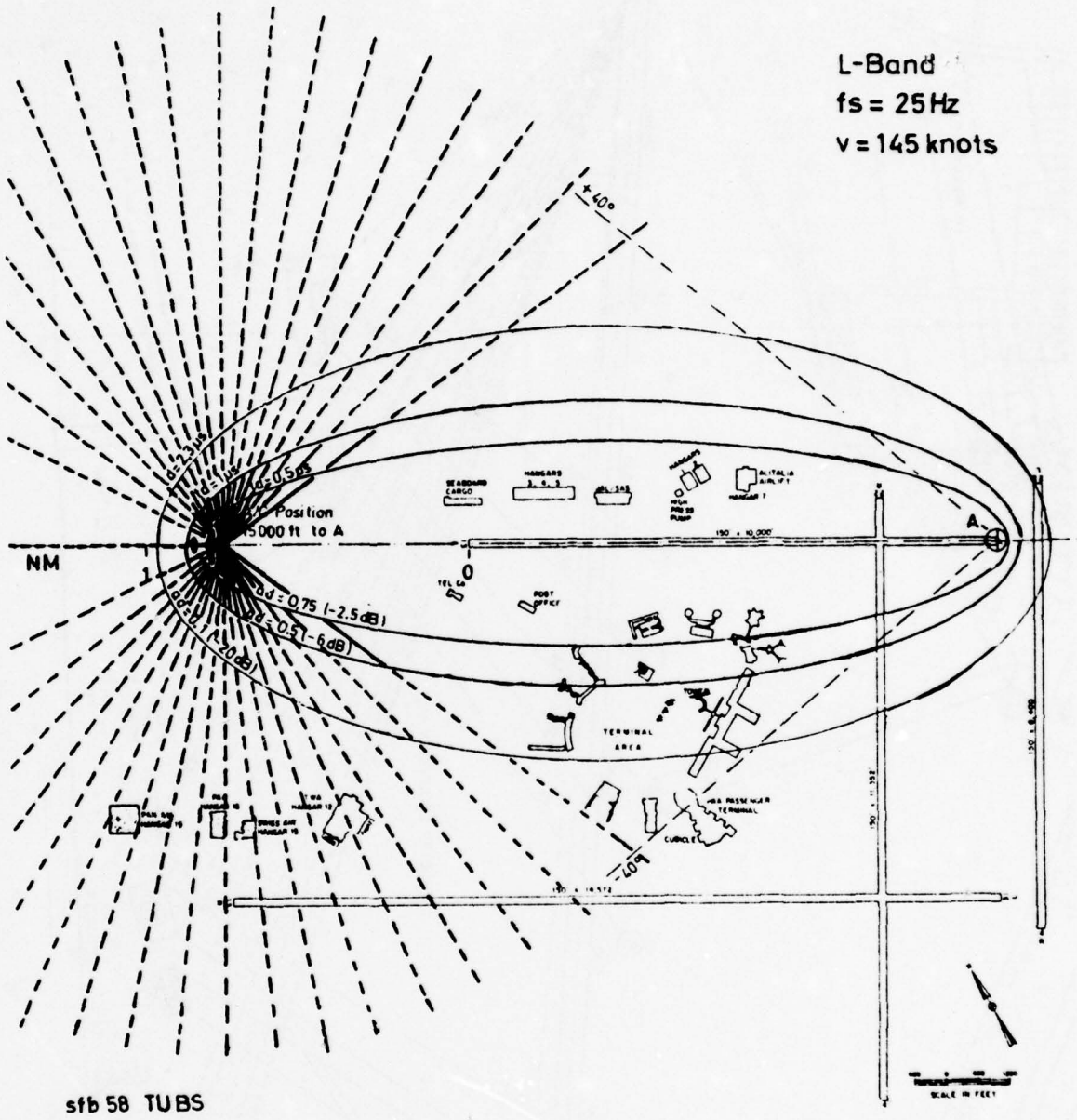


Fig.7 Bad angle orientations and multipath time delay discrimination

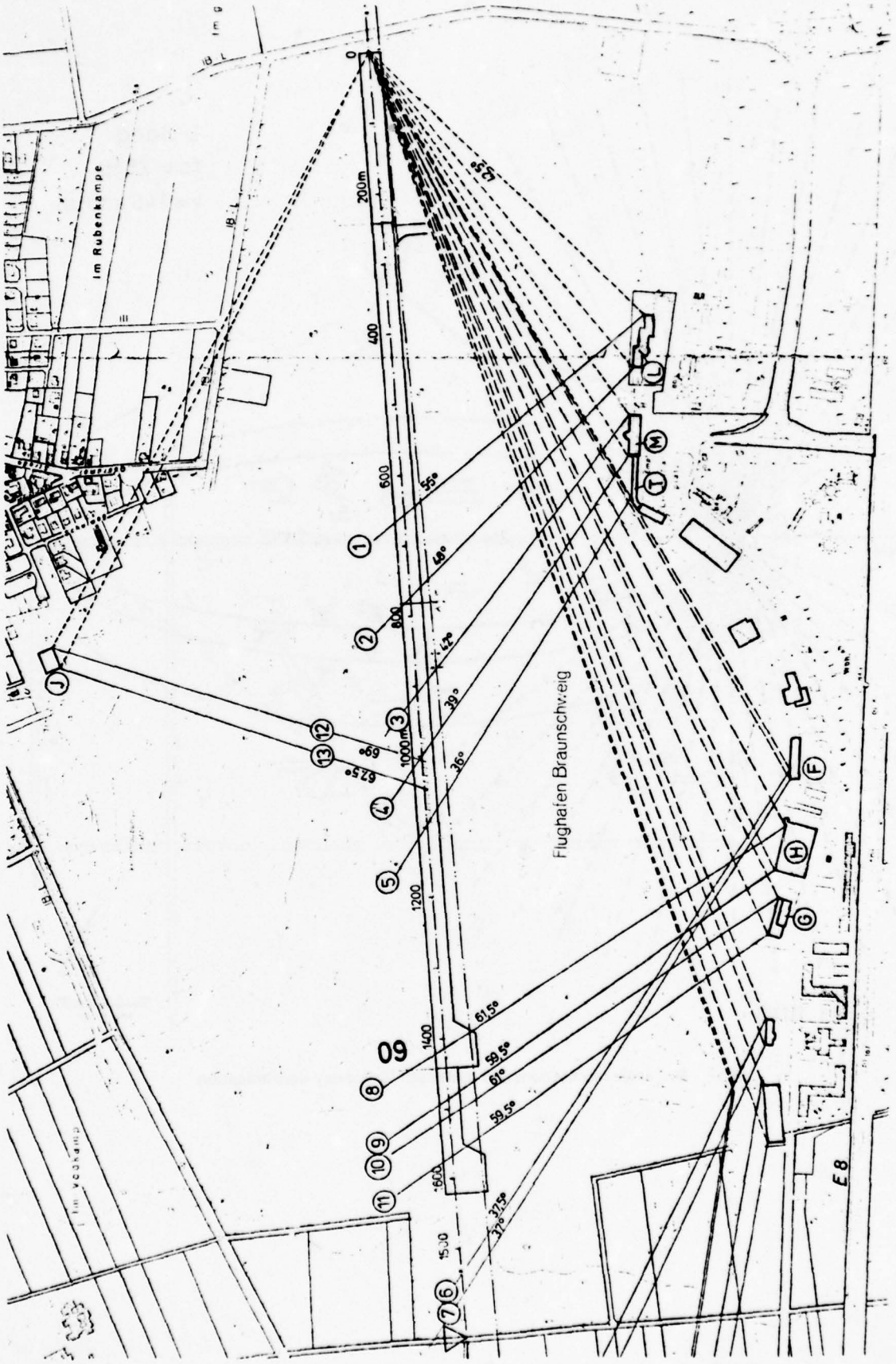
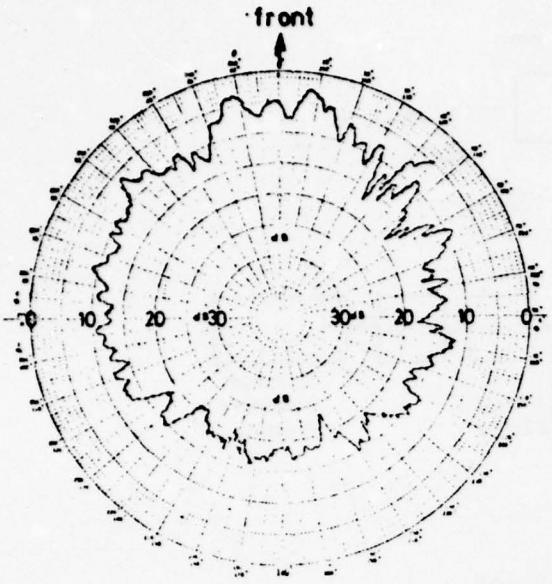
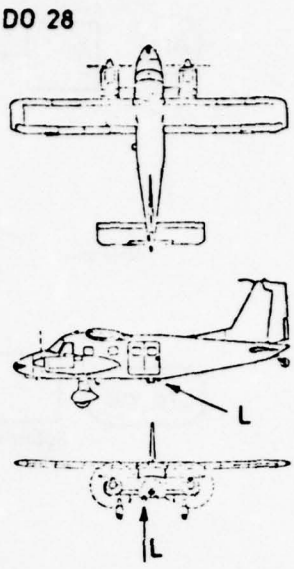


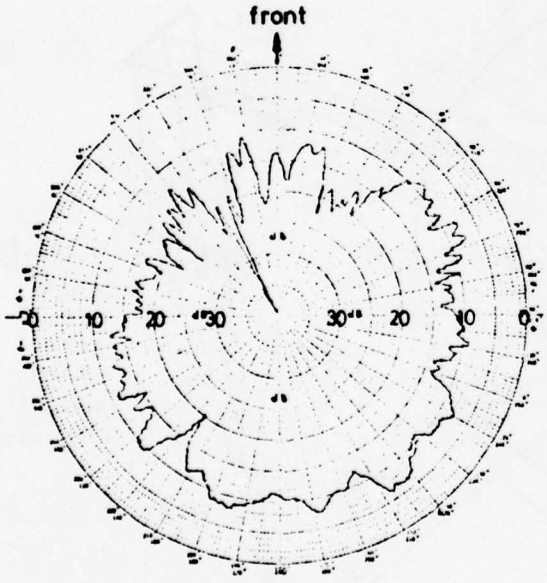
Fig.8 DLS-Testairport Braunschweig RWY 09 L



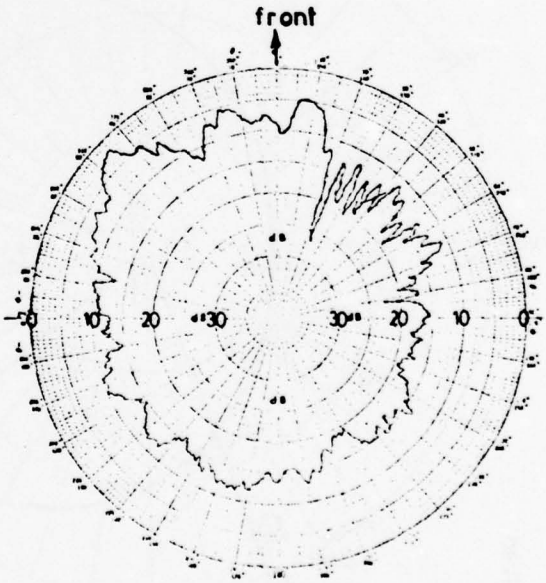
Horizontal Attitude



Test Aircraft and Antenna L



15° banked turn left



15° banked turn right

Fig.9 Installation and horizontal patterns of airborne antenna

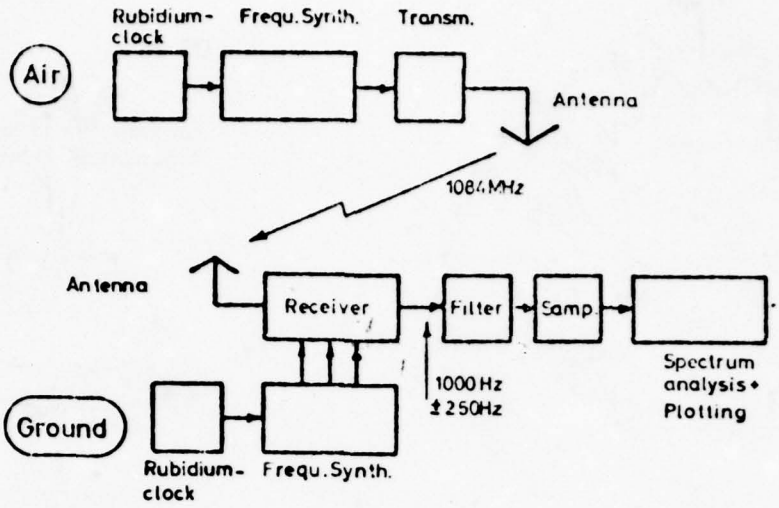


Fig.10 Test equipment

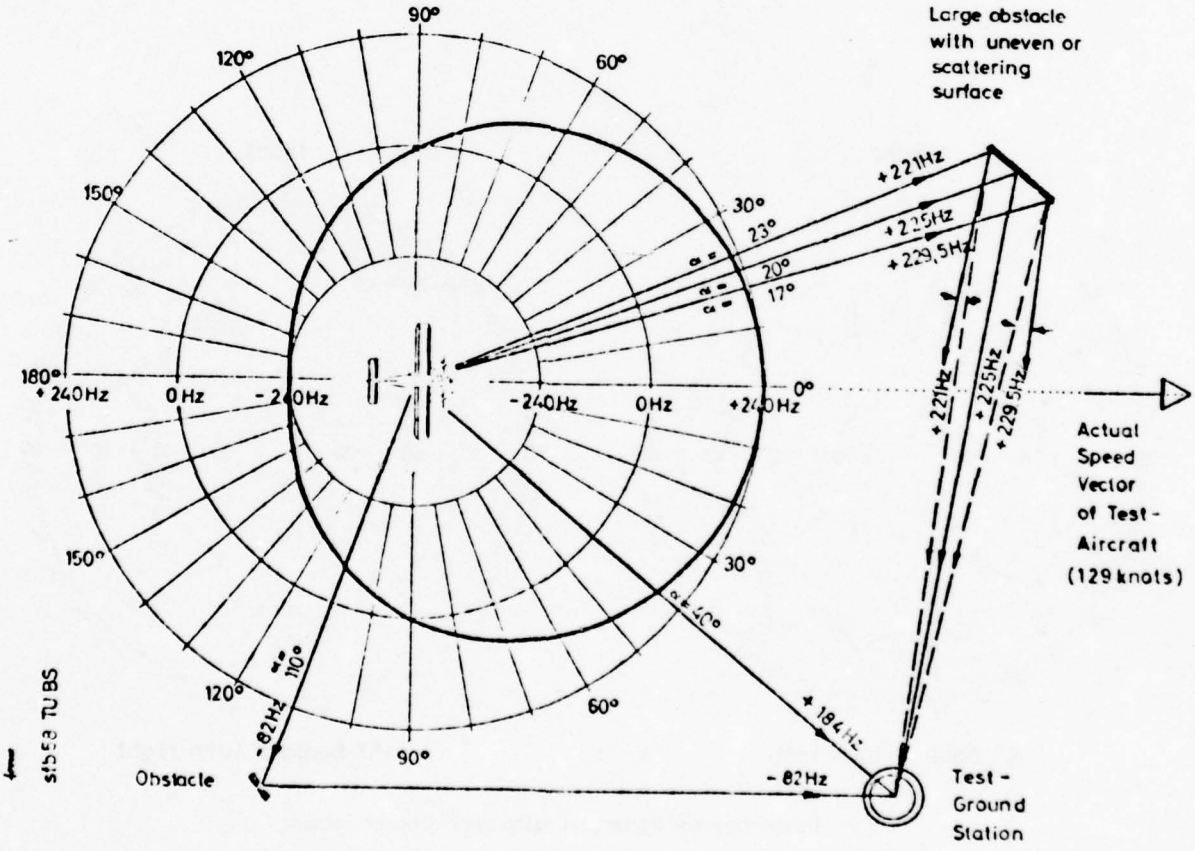


Fig.11 Doppler frequency shift pattern of a moving aircraft (L-band)

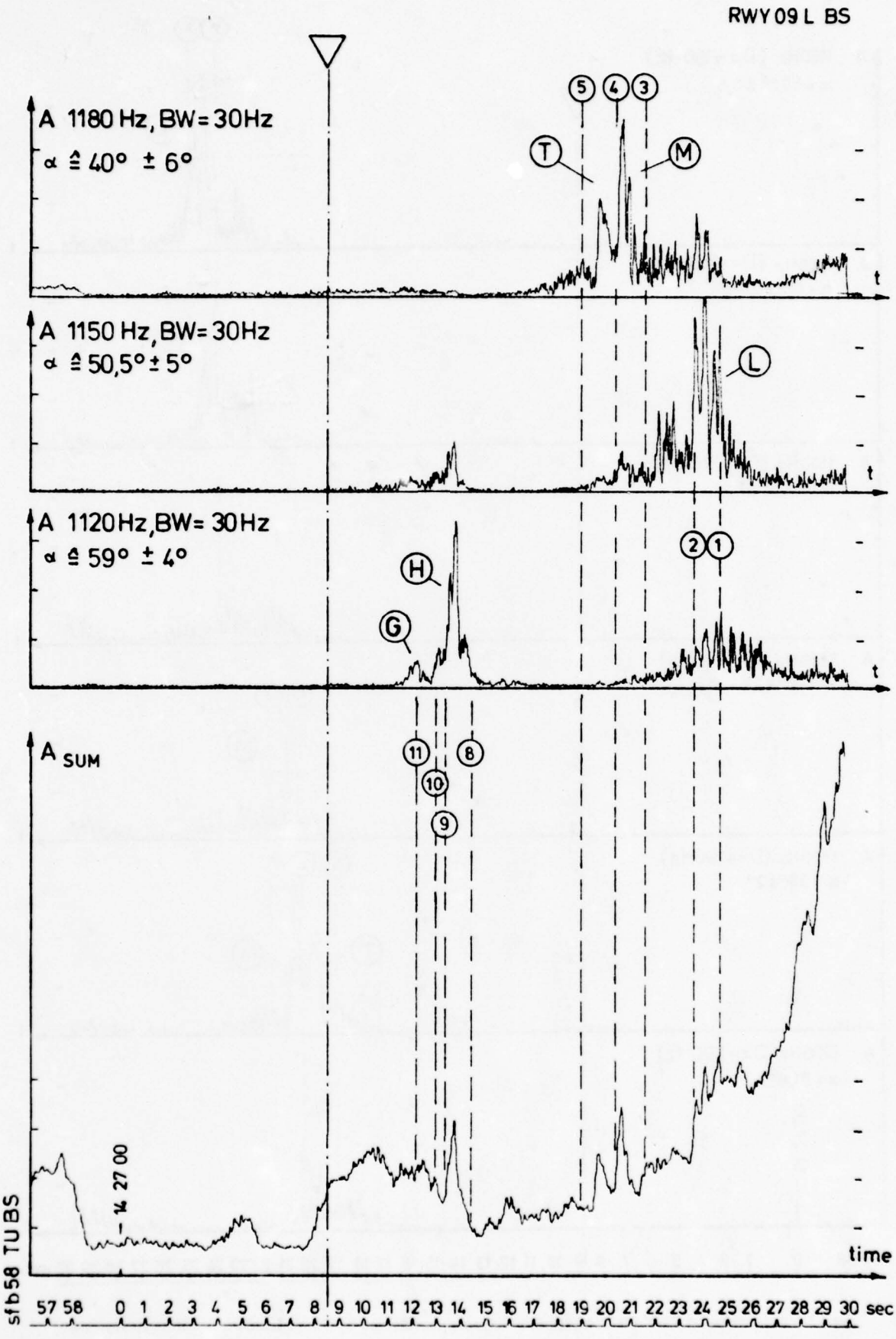

RWY 09 L BS
A 1180 Hz, BW = 30 Hz
α ≙ 40° ± 6°
5
4
3
T
M
t
A 1150 Hz, BW = 30 Hz
α ≙ 50,5° ± 5°
L
t
A 1120 Hz, BW = 30 Hz
α ≙ 59° ± 4°
2
1
H
G
t
A SUM
11
10
9
8
- 14 27 00
sfb58 TUBS
time
57 58 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 sec

Figure 12

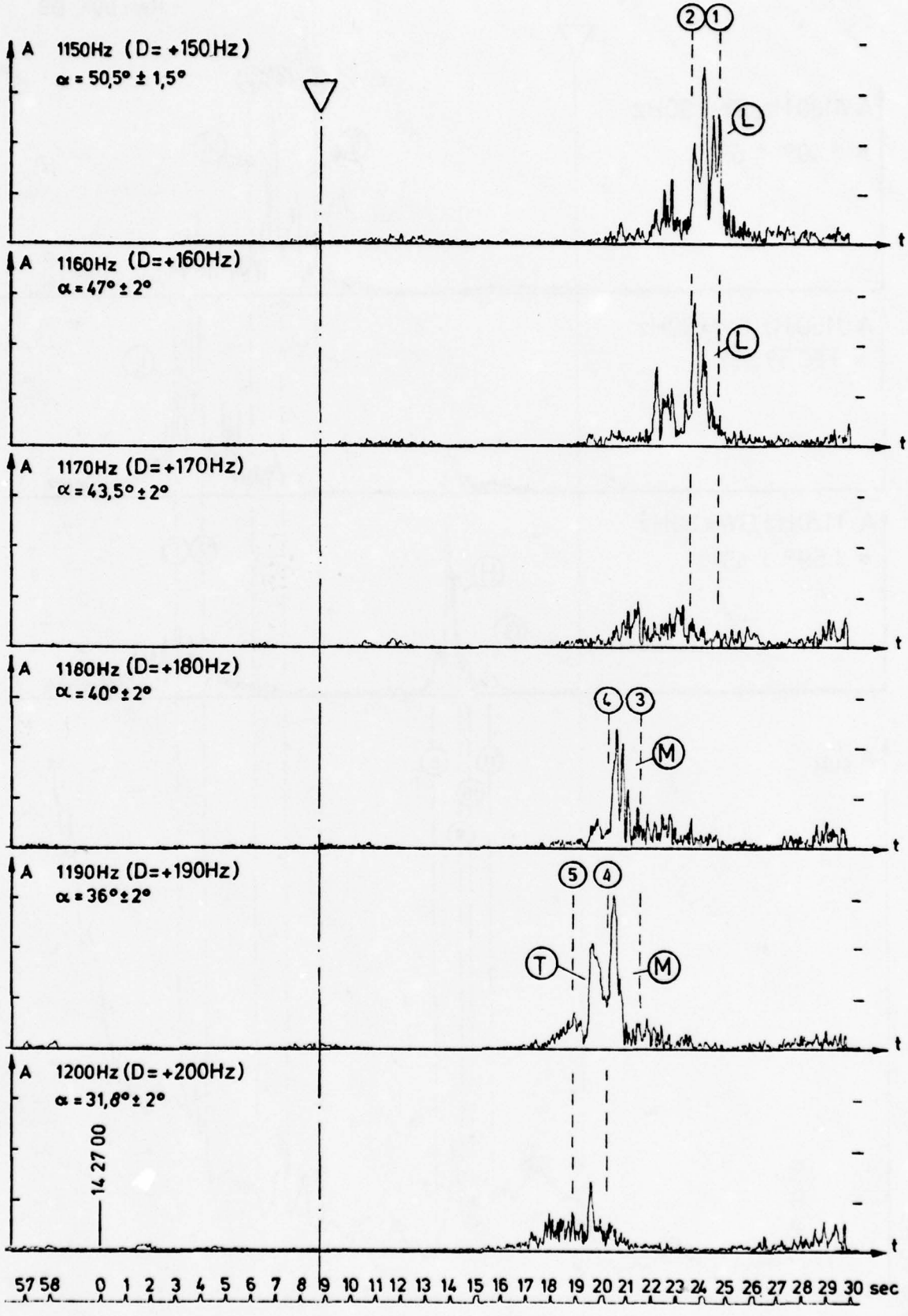


Fig.13 Braunschweig RWY 09 L, multipath components selected by 10 Hz bandwidth-filters versus time t

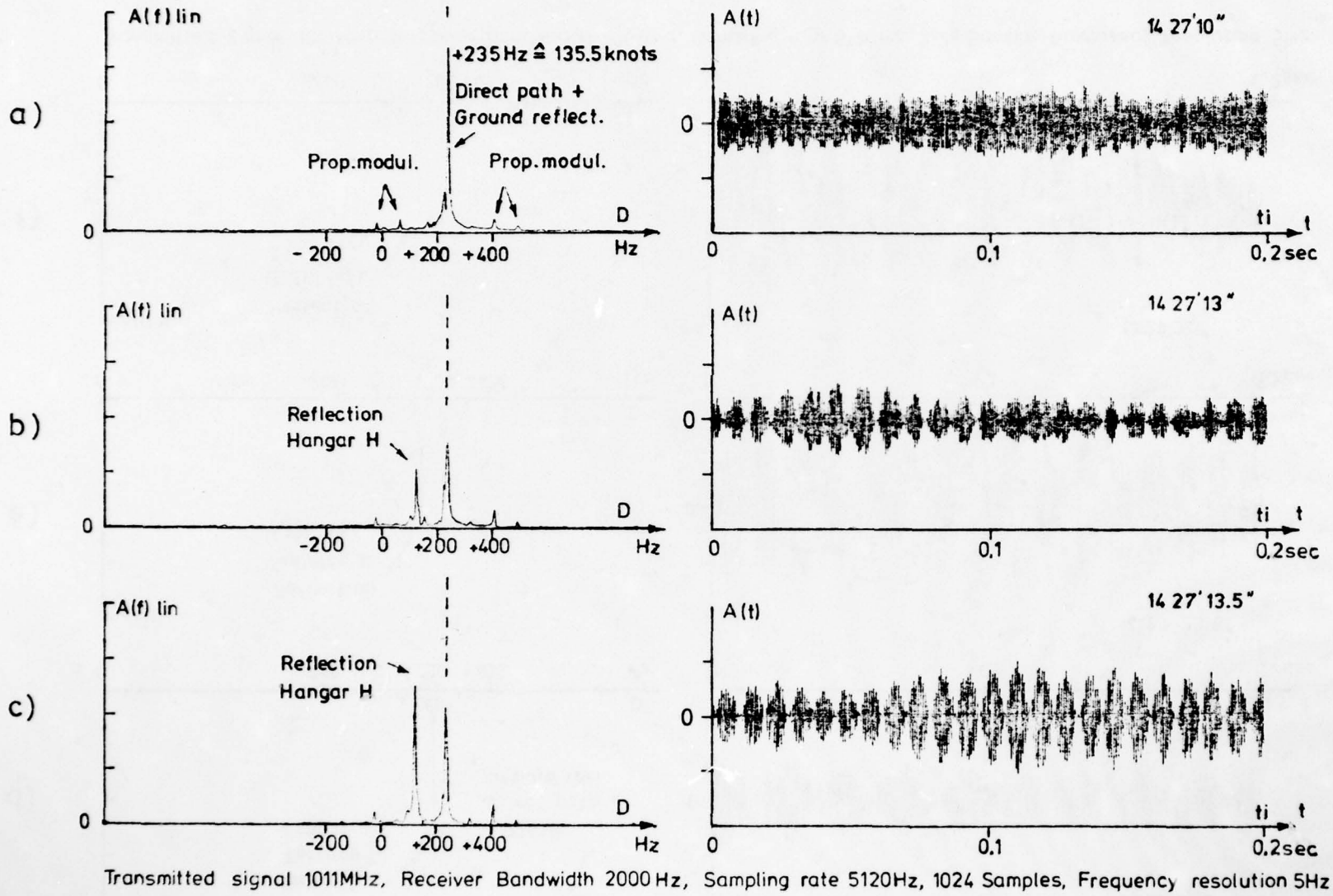


Fig.14 Braunschweig RWY 09 L, 1011 MHz received signal versus doppler shift D and time t

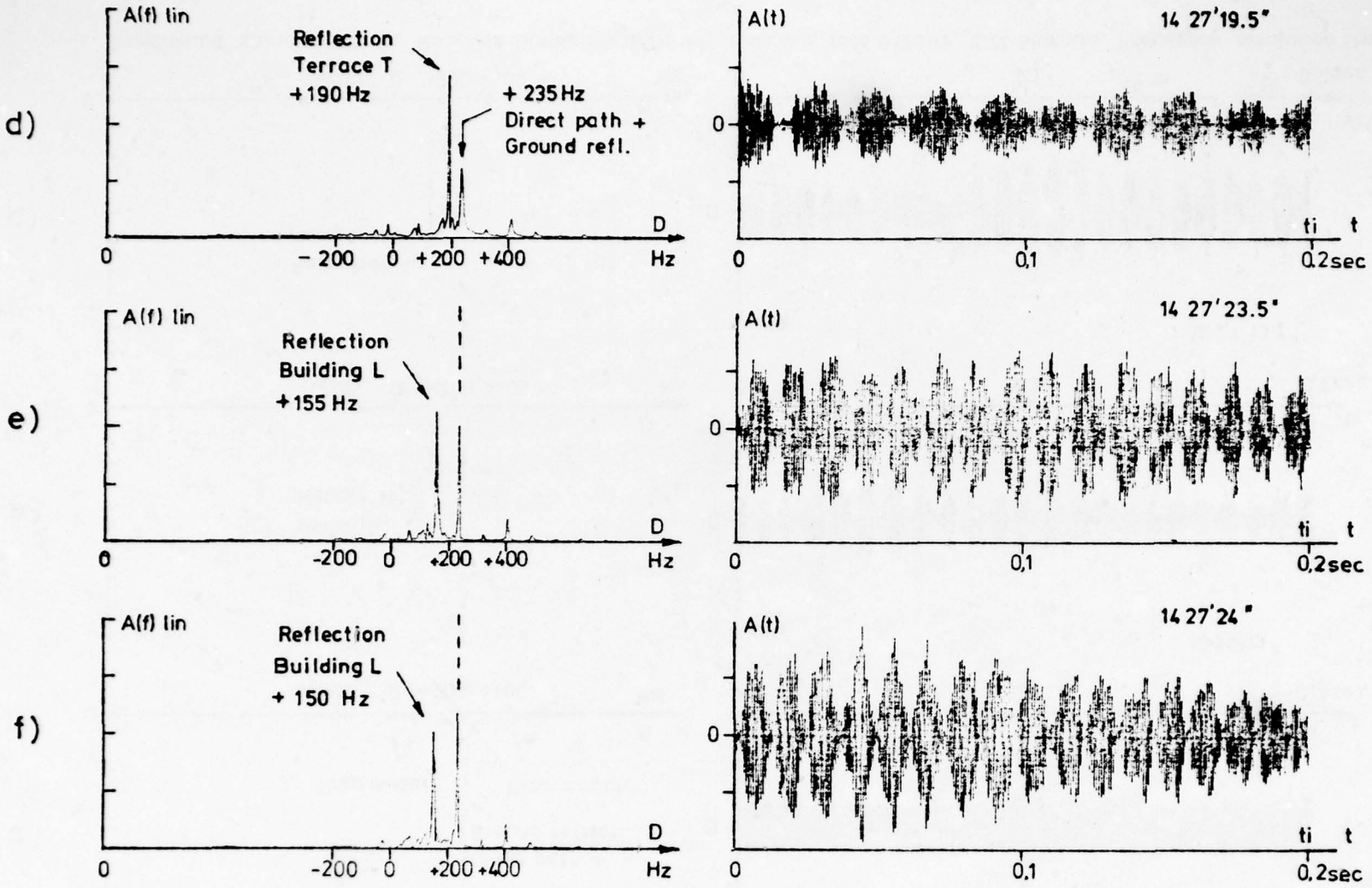


Fig. 14 (cont.) Braunschweig RWY 09 L, 1011 MHz received signal versus doppler shift D and time t

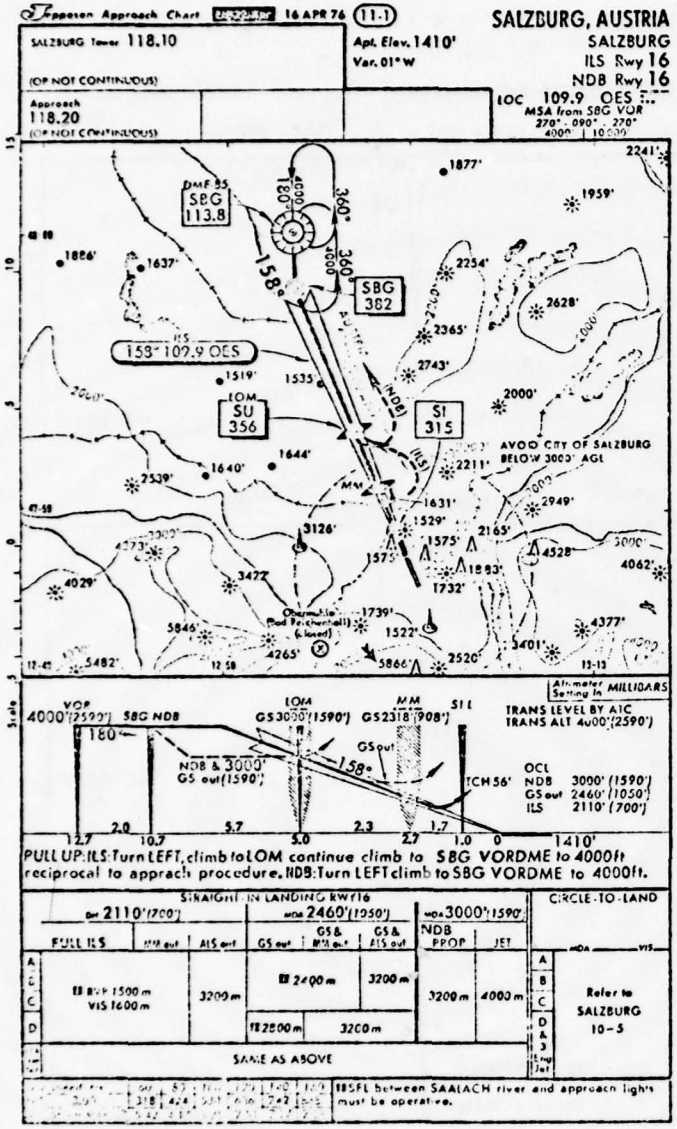


Figure 15

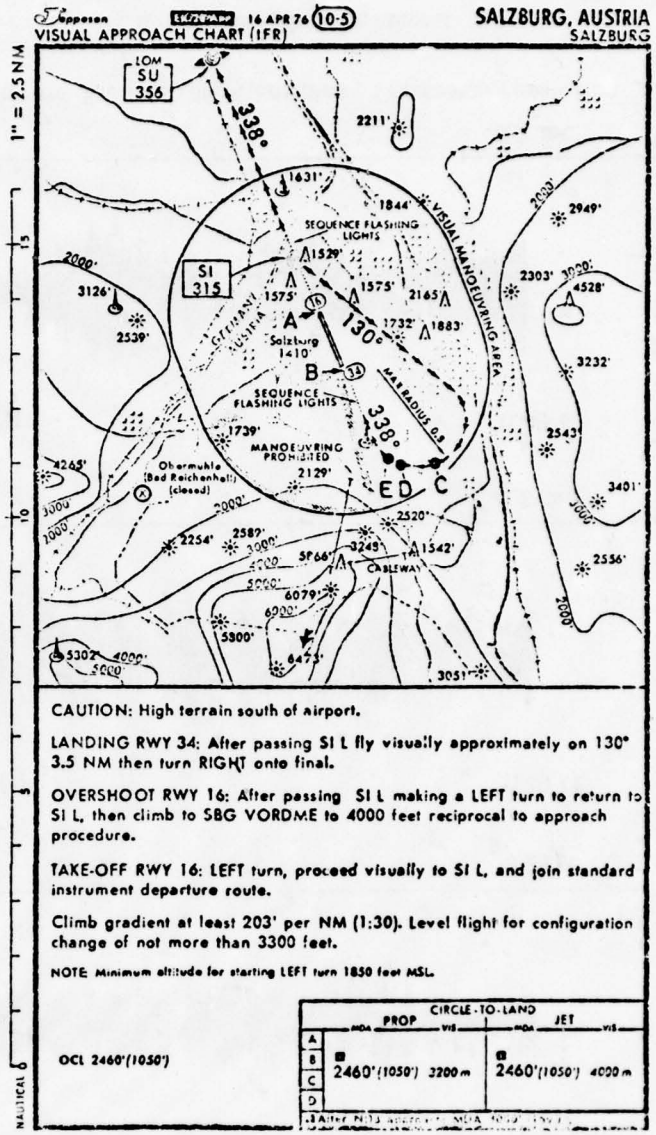


Figure 16

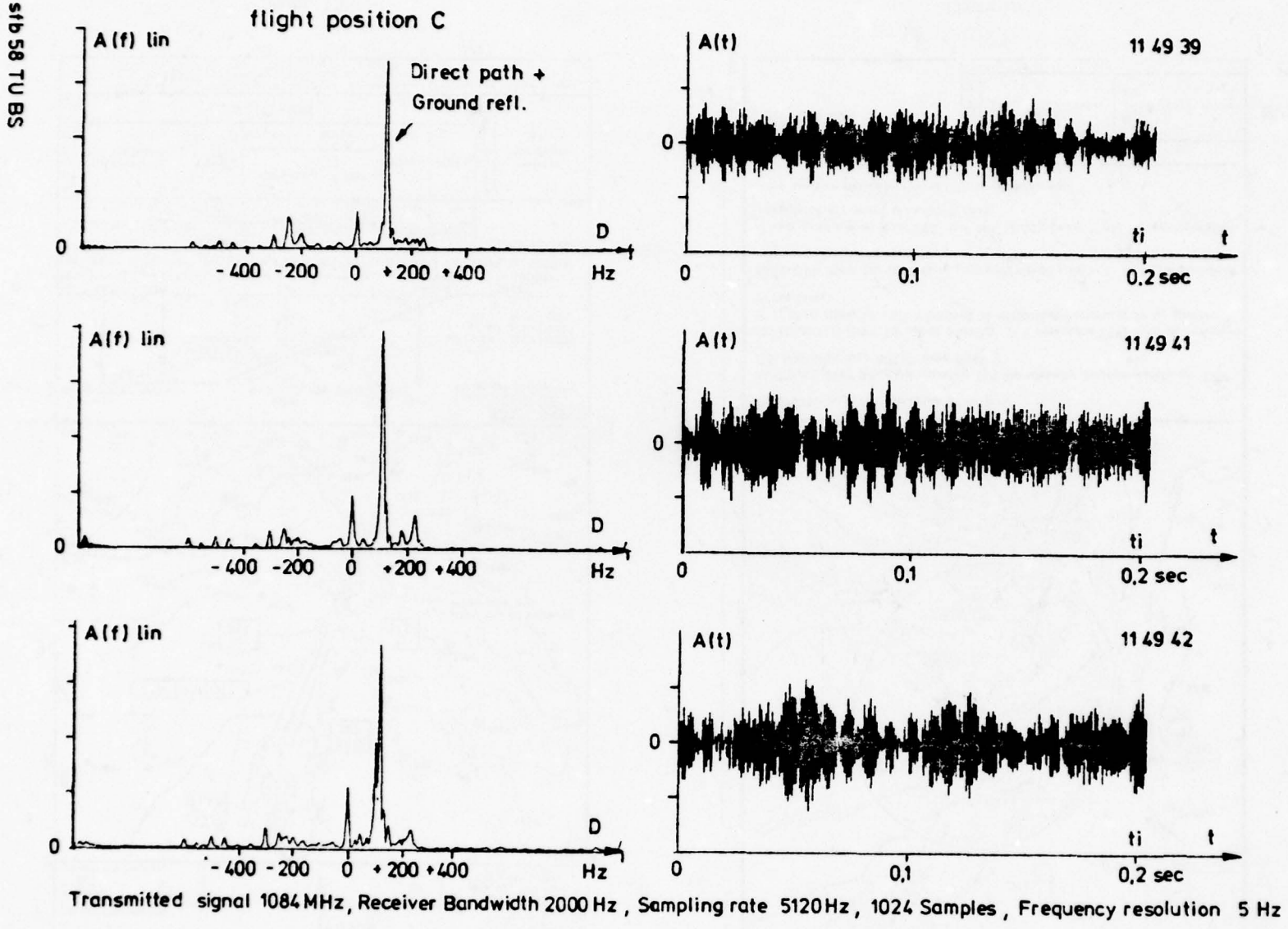


Fig.17 Salzburg RWY 34 circling approach, channel 123 Y, received signal versus doppler shift D and time t

sfb 58 TU BS

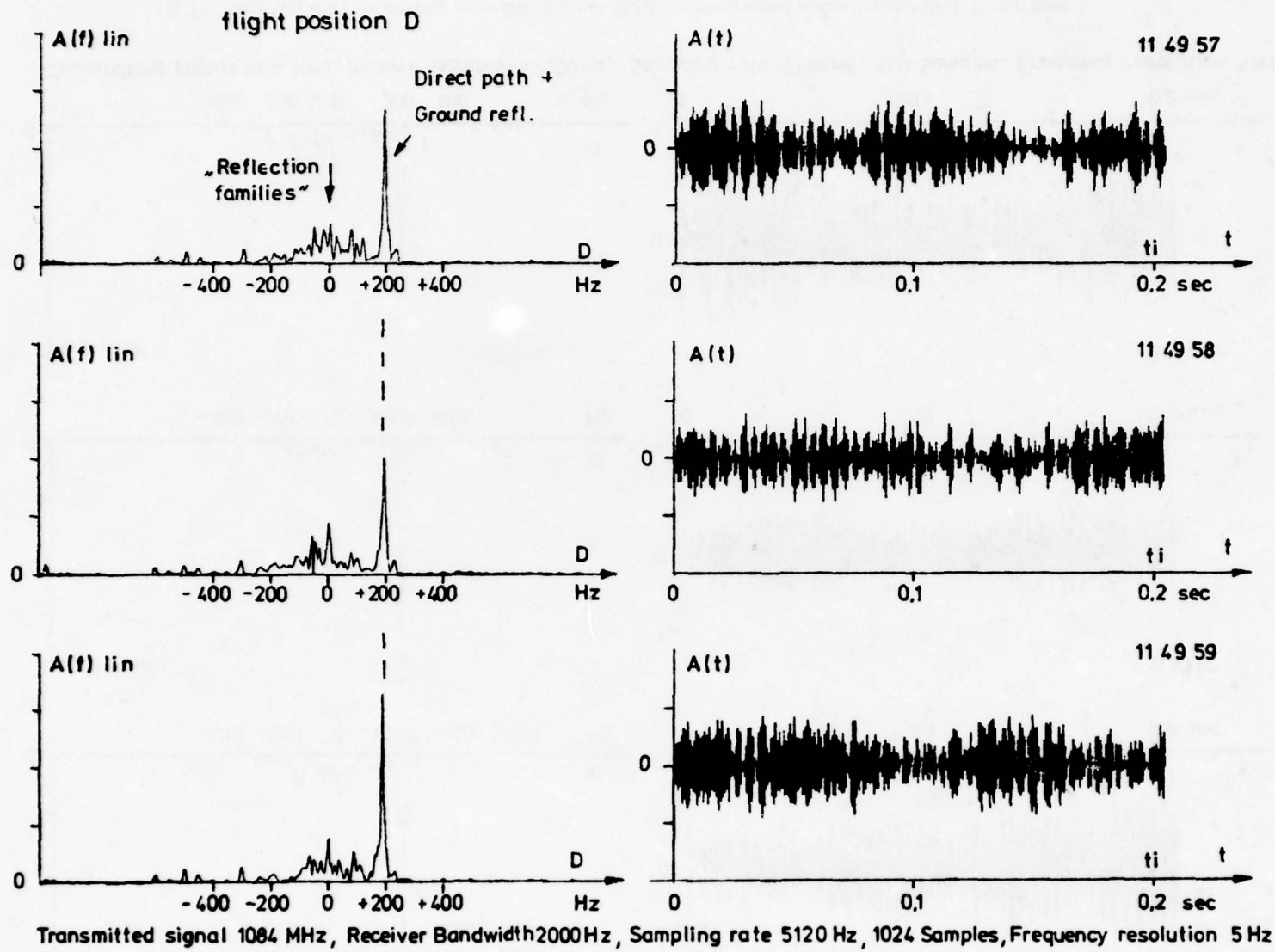


Fig.18 Salzburg RWY 34 circling approach, channel 123 Y, received signal versus doppler shift D and time t

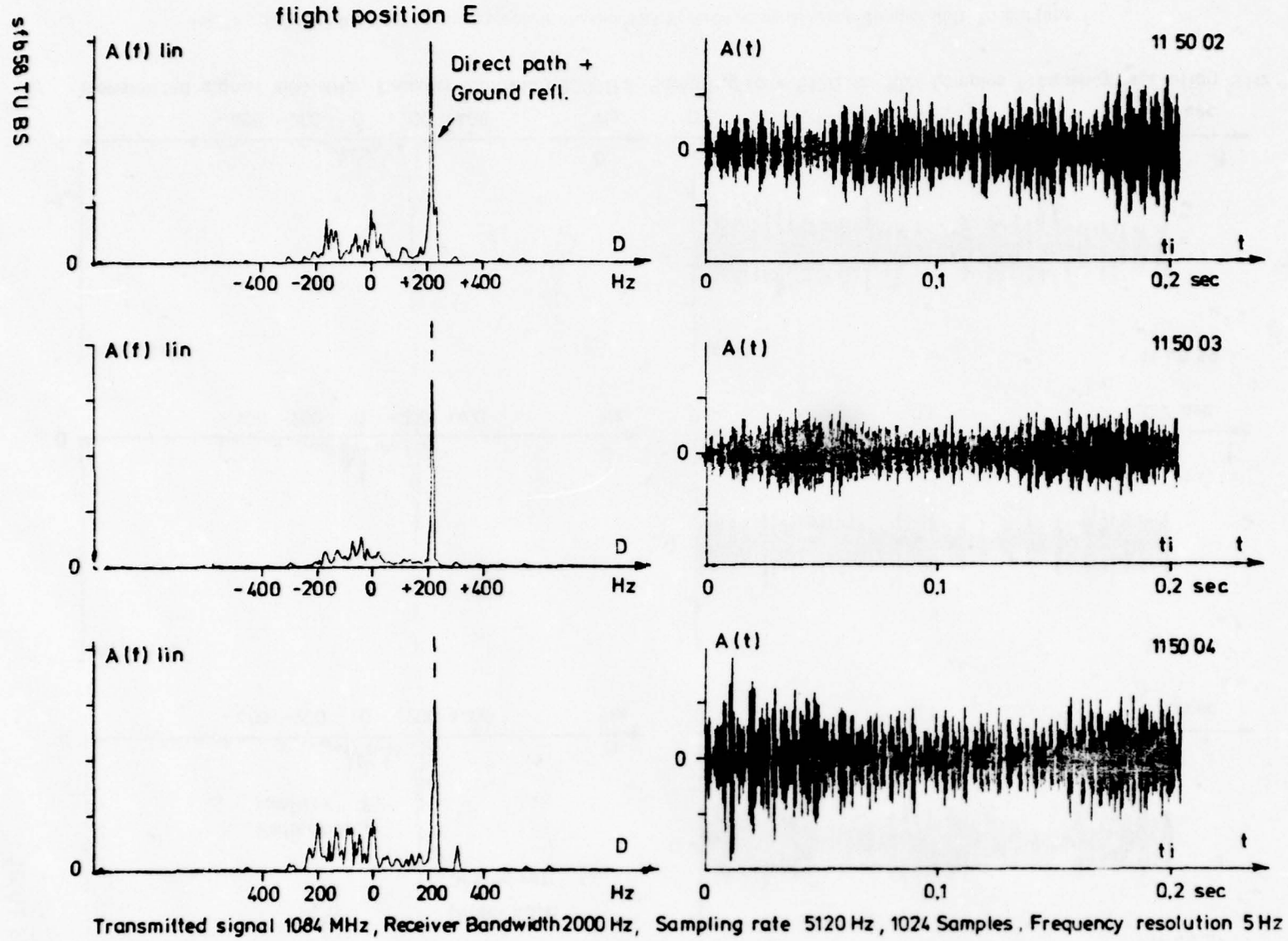
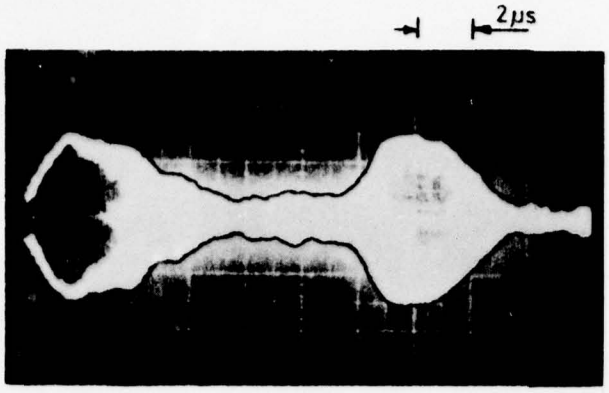


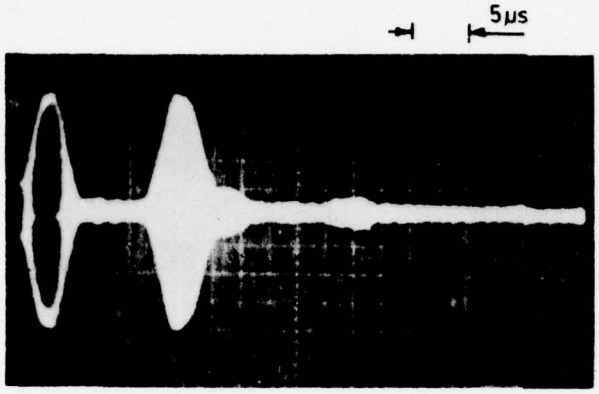
Fig. 19 Salzburg RWY 34 circling approach, channel 123 Y, received signal versus doppler shift D and time t



Rare case of large short delay reflections.

A/C north west SI NDB

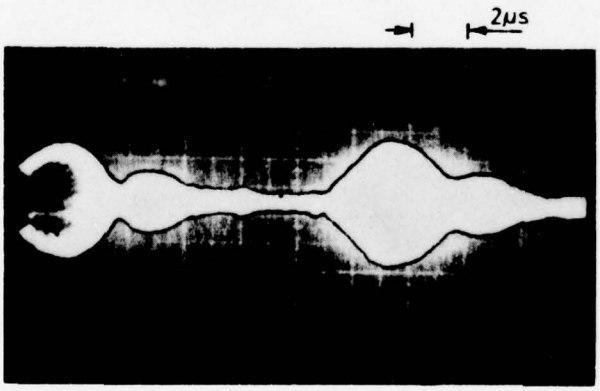
Ground station location A



25% reflection about 15µs delay.

A/C south east of SI NDB, circling

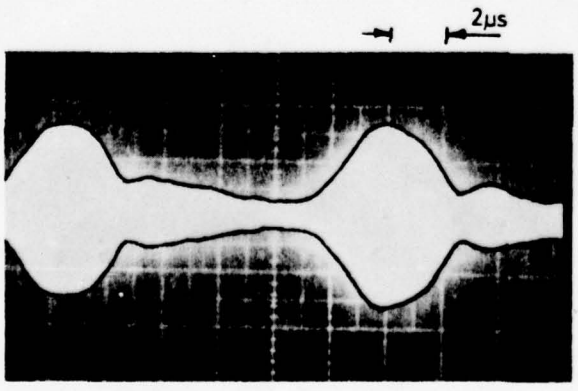
Ground station location A



50% reflection about 4µs delay.

A/C on the runway

Ground station location A



35% reflection about 4µs delay.

A/C circling south east of NDB SI

Ground station location B

Fig. 20 Airborne DME-interrogation pulses and delayed reflections (Salzburg)

A CHANNEL SIMULATOR FOR L-BAND SATELLITE-MOBILE COMMUNICATIONS

Peter D. Engels
Transportation Systems Center
Kendall Square
Cambridge, MA

SUMMARY

The Transportation Systems Center of the U.S. Department of Transportation has developed instrumentation which accurately simulates all important features of a fading, dispersive communications channel. In particular, this equipment provides a simulation of the multipath signals encountered in a satellite-mobile communications link operating at L-Band. This paper presents the results of this development program.

The presentation is divided into three main topics:

1) First, the essential elements of the theoretical background which provides the basis for the design are presented. The results of this analysis are the estimates of important channel parameters such as relative multipath level, bandwidth, etc.

2) The report then describes the final implementation. The simulator provides for simultaneous variation of the following parameters:

   a) Bandwidth of the reflected signal.

   b) Differential doppler between direct and reflected signal.

   c) Relative level of the reflected signal.

   d) Time delay of the reflected signal.

   e) Direct path Doppler.

   The two paths are then linearly combined, and additive white noise is also included so as to provide the correct signal-noise ratio.

3) Finally, the report presents the results of a test program which was designed to validate the simulator performance. This test program exercised several modems through the NASA ATS-6 satellite to a variety of vehicles. Data is presented for performance both through the simulator and through the satellite.

The importance of a validated laboratory simulator is manifest; any variety of experimental modems can be evaluated and compared under identical and precisely controlled test conditions. Such a procedure has clear advantages over conventional field test programs in cost savings, repeatability and comprehensiveness.

1. INTRODUCTION

The design of electronics for use in mobile communications systems utilizing a satellite link must include consideration of the effects of mobile satellite propagation characteristics. In many such links, the signal-to-additive noise ratio is marginally above that required by system performance criteria. Thus, the added degradation resulting from the presence of multipath components reflected from the earth's surface must be minimized. This can be achieved through careful modem design together with evaluation of the propagation effects.

Theoretical evaluations of modem performance in the multipath environment can provide significant insight into the degradation mechanisms. However, such analyses typically ignore part of the modem functions, such as bit synchronization or carrier extraction and their interaction with the multipath distorted signal. Alternatively, some analyses incorporate these second-level processes, but the number of approximations and assumptions necessarily employed limit the usefulness of the results.

The excessive complexity of a complete analytical modem performance evaluation in the presence of multipath and the very high cost associated with operational field tests of such modems lead to the desirability of a channel simulation facility for modem evaluation.

This paper summarizes the work carried out at the Transportation Systems Center of the U.S. Department of Transportation to characterize the performance of a variety of existing modems in the presence of simulated multipath and noise. The results obtained through the simulation experiments are designed to be used in a final analysis leading to an optimum terminal design.

In addition to substantial cost savings, the simulator approach offers several other advantages relative to field test evaluation procedures. For example, the simulator can be implemented in a manner which allows for exactly repeatable channel fluctuations. Thus it is possible to compare the performance of different modem techniques under identical channel conditions. In addition, the simulator can be used to obtain the performance of an experimental modem as a function of various modem parameters such as loop bandwidths, integration times, filter shapes, etc., under precisely repeatable conditions.

2. THEORETICAL BACKGROUND AND HARDWARE IMPLEMENTATION

The mobile-satellite channel has two major propagation paths, the direct path between the mobile and satellite, and a multipath component reflected from the surface of the earth. The satellite is assumed to be geosynchronous, although other orbits can be accommodated. The direct path is characterized by transmission loss (free space), time delay, scintillation and atmospheric fading, and Doppler shift. The transmission loss is simulated by adding noise at the system output so that an appropriate direct-path signal power-to-noise density ratio is established at the input to the demodulation equipment under test. The absolute time delay of the direct-path signal, scintillation and atmospheric fading of the direct path component are not simulated by the system, since it is designed to simulate channels in which multipath fading is the predominant effect. The received signal Doppler shift is simulated over a range of $\pm$1000Hz (nominal). This range accommodates a radial velocity component between mobile and satellite of 400 knots at 1.6 GHz.

The multipath component is characterized by its relative multipath power, relative multipath delay, relative multipath Doppler shift and Delay-Doppler scatter function. The relative multipath power is the average power in the total multipath component relative to the average direct path power. The simulator provides capability for the establishment of any ratio between direct and multipath power from all-direct (no multipath) to all multipath. Typical values of satellite link direct-to-multipath ratios vary from 0 dB to 20 dB, depending on the simulated system's antenna configuration and the characteristics of the terrain being traversed.

The relative multipath delay is the delay of the specular point multipath component relative to the direct path component. This delay approaches zero when the satellite is on the horizon relative to the mobile. It is maximum when the mobile is at the sub-satellite point. Most cases of interest occur when the satellite is at low elevation angle with respect to the mobile. Thus the simulator is designed to provide 3 discrete relative multipath delays, 5 μsec, 30 μsec, and 55 μsec.

The relative multipath delay, τ_{ms}, is approximated by assuming a flat earth and the satellite at infinity. Then,

$$\tau_{ms} \cong \frac{2h}{c} \sin \gamma$$

where h = mobile altitude above the reflecting surface
c = speed of light
γ = multipath reflection grazing angle

The relative multipath Doppler shift is a measure of the variation of relative delay. Assuming a flat earth and the satellite at infinity this component would be entirely due to vertical motions. For most cases, the relative Doppler, f_{ms}, is well approximated by:

$$f_{ms} \cong -v_h \left(\frac{2f_c}{c}\right) \sin \gamma$$

where v_h = mobile vertical velocity
and f_c = carrier frequency

The simulator provides a relative Doppler range of $\pm$100 Hz. This corresponds to a 180 ft/sec vertical velocity component at 1.6 GHz at a grazing angle of 10° or a 60 ft/sec vertical velocity component at 1.6 GHz at a grazing angle of 30°.

The Delay-Doppler scatter function has been the subject of intensive investigation (see References 1, 3, and 7 for examples) and thus will not be discussed in depth here. The overall delay spread, i.e. the integral of the Delay-Doppler scatter function over Doppler, at the -10 dB contour, is approximately (Reference 7, eq. (15)):

$$\tau_{ds} \cong \frac{2h}{c} \beta_o^2$$

where $\beta_0/\sqrt{2}$ is the r.m.s. surface slope. The delay power spectrum shape for various elevation angles is shown in Reference 1, Figure 5. Thus, assuming an aircraft at 50,000 ft. and $\beta_o \cong 0.2828$, the delay power spectrum will extend over several microseconds.

The overall Doppler spread, i.e. the integral of the Delay-Doppler scatter function over all delays, is given by (Ref. 1):

$$D \cong \left(\frac{2f_o}{c}\right) (2\alpha)\, v \sin \gamma \qquad \alpha = \beta_o/\sqrt{2}$$
$$v = \text{radial velocity of mobile}$$

assuming that vertical velocity components are negligible. An upper bound on r.m.s. Doppler spread can be obtained by assuming a Mach 3 aircraft (v = 3000 ft/sec) and α = 0.2 (i.e. β_o = 0.2828). In that case D at 1.6 GHz is approximately 3840 sin γ Hz or, for example, 1920 Hz when γ = 30°. The Doppler spectrum for multipath components which are delayed relative to the specular multipath component have been shown to be bimodal (Ref. 3) with most of the spectral energy piled up at the edges of the band and the bandwidth increasing with increasing delay.

The simulator implements the channel's Delay-Doppler scatter function by means of a tapped delay line. Specifically, 5 taps 2 μsec apart are employed. Thus valid channel models can be created for signal bandwidths of 200-300 kHz or less. At each tap, the delayed signal component is split into in-phase and quadrature components. These are modulated in balanced modulators by independent low-pass gaussian noise functions to create a Rayleigh fading process. The bandwidth of the noise at each tap is

independently controllable and can be set to create the desired Doppler-as-a-function-of-delay characteristic. The low-pass gaussian processes have second-order Butterworth filter shapes.

The Doppler spread, D, of the process is defined (Ref. 1, p. 559) as twice the standard deviation (r.m.s. bandwidth). The Doppler spread then becomes:

$$D = 2f_{3\ dB}\ \text{(Hertz)},$$

$f_{3\ dB}$ = 3 dB bandwidth of the Butterworth filter.

Front-panel controls are provided on the simulator to set the bandwidth of each of the five tap modulation processes. These controls are calibrated directly in r.m.s. Doppler spread (Hertz). Doppler spreads in the range from 10 Hz to 1.999 kHz can be generated. Each of the ten independent low-pass gaussian processes is generated from a pseudo-random sequence of length 2^{39}-1. The sequences can be reinitialized at any time by depressing a reset control. Thus the fading effects created by the simulator are exactly repeatable. In addition, the sequence generators can be stopped and restarted by a front-panel control. In this way, the simulator can provide a "frozen-channel" for experiment purposes. In operation, the instantaneous delay through the tapped delay line portion varies dynamically. This effect over-shadows the slow changes in relative multipath delay which are not simulated.

Figure 1 shows a block diagram of the channel simulator. The simulator input and/or output can either be at 70 MHz or at L-Band. Internally, the system has a common I.F. frequency of 40 MHz. Either input frequency is down-converted to 40 MHz, and split into two paths. One, called the direct path, passes the signal directly through, with no modification. The other path provides the relative delay, delay spread, Doppler spread and relative Doppler functions which characterize the multipath return. An attenuator sets the level of relative multipath power, after which the multipath component is summed linearly with the direct path component. The composite signal is then mixed back to 70 MHz and summed with additive noise to set the desired total signal-to-noise ratio. The 70 MHz signal can be mixed to L-Band if so desired. Figure 6 is a photograph of the simulator.

3. TEST CONDITIONS

Most of the modem tests described in this report are measurements of performance versus direct path carrier-to-noise density (C/N_o) as a function of two multipath parameters, the carrier-to-multipath (C/M) and the multipath Doppler spread, D. The relative power of the reflected multipath signal is a function of the multipath reflection angle and the characteristics of the reflecting surface. The range of this parameter over sea water is from -2 dB to -5 dB.* The multipath actually seen by the communication system is reduced by antenna discrimination. However, the multipath attenuation achieved by this means under low reflection angle conditions may not be significant. Therefore, the smallest C/M ratio used in the tests was selected to be 5 dB. Additionally, testing was conducted at C/M values of 8 and 11 dB.

The r.m.s. multipath spreads used in most of the test runs were 10 Hz, 100 Hz and 1 kHz. A 10 Hz Doppler spread is a lower bound on this parameter when the multipath reflection occurs on the sea-surface. In that case 10 Hz of Doppler spread is attributable to sea-surface motion. A 100 Hz Doppler spread is typical of that expected on satellite-to-subsonic aircraft links, while the 1 kHz spread is typical of SST type aircraft.

The simulator can also generate relative Doppler shifts and various relative multipath delays. These were found to have little or no effect on the performance of the modems tested. Therefore, all the results presented are for no relative Doppler shift and 5 μsec. relative delay.

The simulator is designed to generate multipath delay spreads up to 10 μsec through use of the 5 tap multiplier units. However, all modem tests reported here were carried out using a single tap. This is possible because the modems tested were all of relatively narrow bandwidth; therefore multiple tap simulation was found to have negligible additional effect.

4. DESCRIPTION OF TEST MODEMS

A variety of modems have been tested using both the channel simulator and the NASA ATS-6 satellite. The results of testing with three different modems are included in this report. Two of the test modems have the capability of simultaneous voice and data transmission on a single constant envelope carrier. These hybrid modems were also capable of transmitting either voice or data separately.

Although both hybrid modems employed quadrature modulation techniques to provide simultaneous voice and data capability, their voice modulation techniques were different. One hybrid modem (Hybrid I) employed true phase modulation for voice transmission, with the maximum phase deviation limited so as to preserve the carrier amplitude. The other hybrid modem (Hybrid II) used suppressed carrier pulse duration modulation (SCPDM) for voice transmission. Both hybrid modems employed differentially encoded coherent phase shift keying (DECPSK) for data transmission, i.e., their data demodulation circuitry utilized true coherent demodulation.

The third modem transmitted data only, and employed differentially coherent phase shift keying (DPSK), i.e., the reference signal for data demodulation is a delayed replica of the incoming signal.

The tests reported herein used a bit rate of 1200 bits per second, with no error correction.

---

*This discussion emphasizes sea-surface reflections rather than terrain reflections since they are larger and therefore more troublesome. The results can be applied to terrain reflections given an estimate of the relative multipath power.

5. TEST CONFIGURATION

The channel simulator test set-up block diagram is shown in Figure 2. The modulator input consists of a 2047 bit maximum-length shift register sequence generated by a Hewlett-Packard error analyzer. The modulator output is passed through the simulator where multipath distortion effects are produced and a controlled amount of bandpass gaussian noise is added to the process, so as to generate the desired carrier-to-noise ratio. The simulator output is then applied to the modem demodulator. The modem data output is analyzed by the same error analyzer which counts the bit errors. Signal and noise power were monitored by a precision power meter. The noise power was measured with a calibrated bandpass filter in cascade so that, knowing the filter noise bandwidth, noise power density could be calculated.

A similar test was performed through the NASA ATS-6 satellite. The test modems received data which was generated at a NASA ground station and transponded by the satellite to a downlink freuqency of 1550 MHz. The test modems were flown aboard an FAA aircraft in a variety of flight patterns. A narrow beamwidth antenna aimed at the satellite received the direct-path signal, uncontaminated by surface reflections. A second steerable antenna was oriented such that its output was the received signal after surface reflection. The two antenna outputs were then combined so as to generate a signal with the desired ratio of direct-to-reflected signal. The test conditions of interest are summarized in Table 1.

The test sequence generator and bit error rate analyzer used in the satellite tests was identical to that used for the laboratory channel simulation tests.

TABLE 1. TEST CONDITIONS

Parameter	Aircraft-Satellite Model (L-Band)
Doppler Spread	
10 Hz	very slow aircraft or surface ship
100 Hz	typical subsonic aircraft
1000 Hz	supersonic aircraft
Carrier-to-Multipath Ratio	
5 dB	strong multipath, little or no antenna discrimination
8 dB	typical low elevation multipath, hemispherical coverage antenna
11 dB	typical low elevation multipath, sophisticated antenna
Relative Multipath Delay	
5 μsec	low elevation angle
55 μsec	moderate elevation angle, moderate altitude
Relative Multipath Doppler	
0 Hz	aircraft in level flight
10 Hz	aircraft climbing or descending 16 ft/sec

6. DISCUSSION OF TEST RESULTS

The test results are shown in Figures 3, 4, and 5. On these figures, the solid lines represent modem performance using the channel simulator; the solid triangles, squares and circles are discrete data points taken in field tests with the ATS-6 satellite.

First, consider a comparison of performance of Hybrid I to Hybrid II (Figures 3 and 4). Although these modems provide similar performance in the presence of gaussian noise, the addition of multipath components results in noticeable performance differences. In general, Hybrid I provides better performance in the presence of multipath than Hybrid II. For example, with C/M = 11 dB, Hybrid II requires C/N_0 to be higher by roughly 1.4 dB to provide the same performance as Hybrid I averaged over the range from 40 dB-Hz to 46 dB-Hz and over D = 10 Hz, 100 Hz, and 1000 Hz. Similarly at C/M = 8 dB, Hybrid II requires C/N_0 to be 1.5 dB greater for the same average performance. With C/M = 5 dB, the performance difference is roughly 0.8 dB in favor of Hybrid I. Noting the slope of the BER curves at C/M = 5 dB, it is seen that this latter performance difference is very small when expressed in terms of BER degradation.

This observed performance trend is difficult to explain on theoretical grounds since it involves a complex interaction of the signal, noise, and multipath components with the modem's detection, carrier loop, and bit synch loop circuitry. It is likely, however, the minor differences in IF filter characteristics, loop compensation techniques, and other similar factors are responsible for the performance difference, since both systems have the same general block diagram when operated in the data-only mode.

Figure 5 represents the performance of the DPSK I modem. In the absence of multipath, this modem's performance will be inferior to that of a DEC PSK modem; however, the presence of multipath can cause more degradation in a DEC PSK modem than in a DPSK modem (Ref. 4). This is the case for the range of conditions evaluated. The DPSK I modem performance is compared to that of Hybrid I in Table 2. Table 2 shows the added C/N_O required by Hybrid I to achieve the performance of the DPSK modem (DPSK I). Since the slopes of the various performance curves are not equal, the improvement of DPSK I over Hybrid I cannot be expressed as a single number. In general, the DPSK I performance is inferior or approximately equal to that of Hybrid I where C/N_O is low and BER correspondingly high. As C/N_O is increased, the steeper slopes of the DPSK I performance curves result in better performance comparisons for the modem. The DPSK I performance advantage rapidly diverges to significant values at the higher C/N_O ratios.

TABLE 2. INCREASE IN HYBRID I C/N_O TO EQUAL DPSK I BIT ERROR RATE - AT SAME MULTIPATH LEVEL

		REQUIRED INCREASE IN C/N_O		
C/M	D(Hz)	BER = 10^{-3}	BER = 10^{-4}	BER = 5×10^{-5}
11 dB	10	0.4 dB	1.2 dB	2.0 dB
11 dB	1000	0.3 dB	0.9 dB	1.1 dB
		BER = 10^{-2}	BER = 5×10^{-3}	BER = 10^{-3}
8 dB	10	0.6 dB	1.2 dB	2.5 dB
8 dB	1000	-0.6 dB	-0.2 dB	0.1 dB
		BER = 3×10^{-2}	BER = 2×10^{-2}	BER = 10^{-2}
5 dB	10	-0.2 dB	0.2 dB	0.7 dB
5 dB	1000	0.6 dB	1.3 dB	3.3 dB

Finally, note the excellent agreement between laboratory simulation data and actual field test data through the ATS-6 satellite. The agreement is striking in the case of Hybrid II; Hybrid I shows somewhat more variation but the agreement is still excellent. DPSK I shows considerable data scatter; this is due, most likely, to the fact that each data point represents an average of a number of points within a 3 dB range of the S/I parameter. Thus the actual range of S/I shown is 4.5 dB to 13.5 dB. In addition there was somewhat less data available for this modem over the region of interest.

From this data it is apparent that the channel simulator generates an excellent replica of the actual multipath channel; consequently, data derived from laboratory tests using this simulator provide a realistic evaluation of performance data under operational conditions. Therefore comparative modem performance and selection data can be generated without resorting to complex, expensive and time-consuming field test programs.

REFERENCES

1. Bello, P. A., 1973, "Aeronautical Channel Characterization", IEEE Trans. Comm. Vol. COM-21, No. 5.

2. Boeing Commerical Airplane Div., 1973, "ATS-5 Multipath/Ranging/Digital Data L-Band Experimental Program" FAA Report No. FAA-RD-73-57.

3. DeRosa, J. K., 1970, "On the Determination of the Delay-Doppler Scattering Function for a Ground-to-Aircraft Link", Digest of the Canadian Symposium on Communication.

4. Frasco, L. A. and Goldfein, H. D., 1973, "Signal Design for Aeronautical Channels", IEEE Trans. Comm. Tech. Vol. COM-21.

5. Getchell, E. H. and Mahoney, P. F., 1976, "A Simulation to Produce Narrow-band Multipath Effects on L-Band Aircraft-to-Satellite Signals", U.S. Dept. of Transportation Report No. DOT-TSC-OST-76-44.

6. Jones, J. J., 1968, "Multichannel FSK and DPSK Reception with Three-Component Multipath", IEEE Trans. Comm. Tech. Vol. COM-16.

7. Salwen, H., 1971, "Characteristics of Satellite-to-Aircraft Links", Record of the International Commmunication Conference.

8. Salwen, H. and Duncombe, C. B., 1975, "Performance Evaluation of Data Modems for the Aeronautical Satellite Channel", IEEE Trans. Comm. Vol. COM-23, No. 7.

9. Wilson, S., 1973, "Satellite-Aircraft Digital Transmission Experimental Results at L-Band", Record of the Int'l Telemetering Conference.

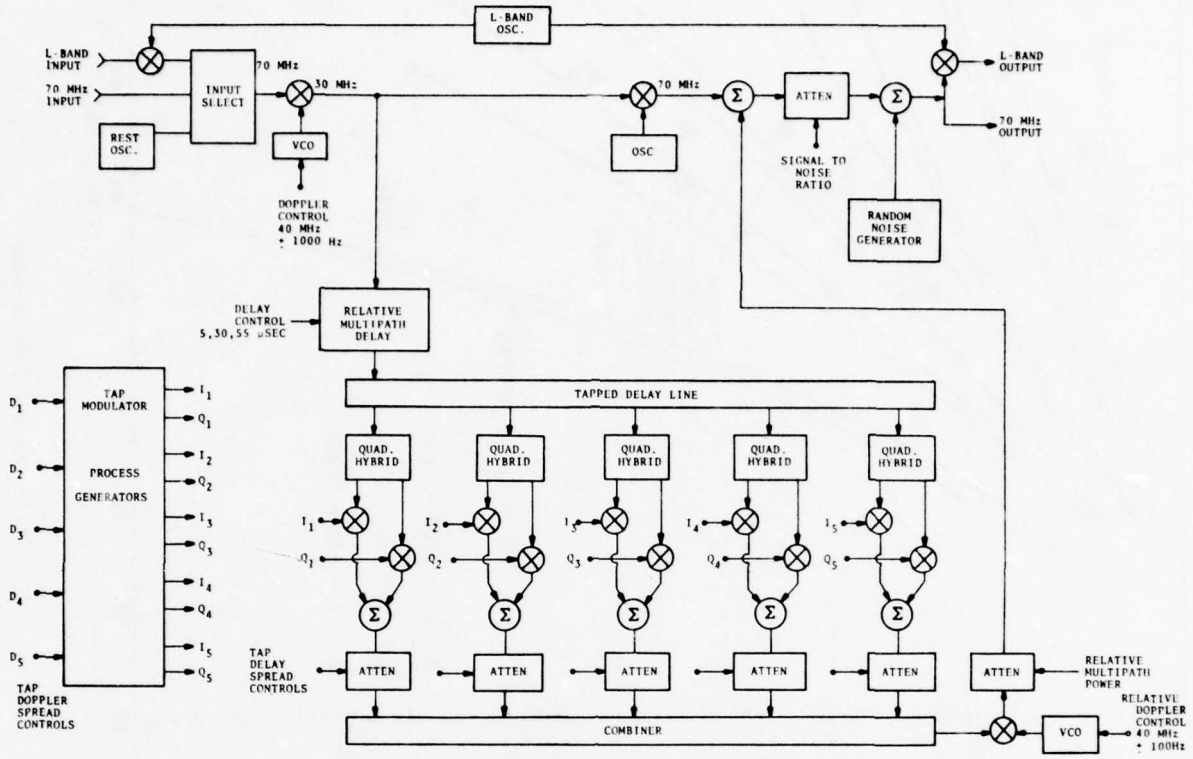


Fig.1 Channel simulator block diagram

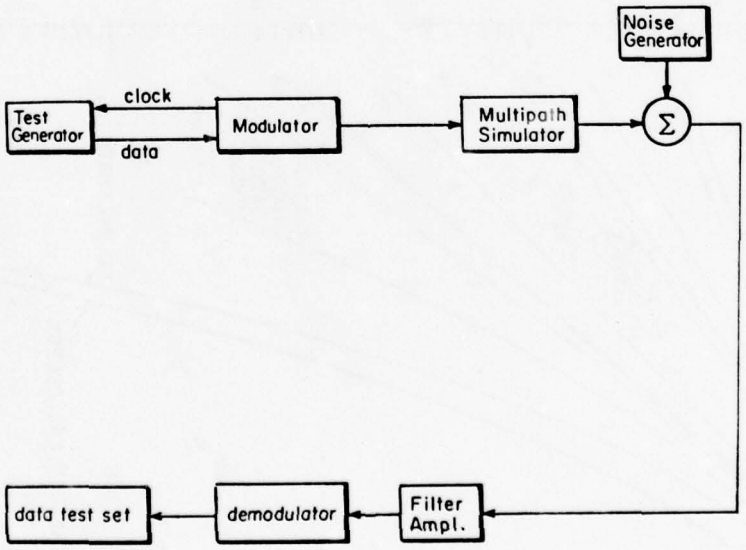


Fig.2 Test configuration

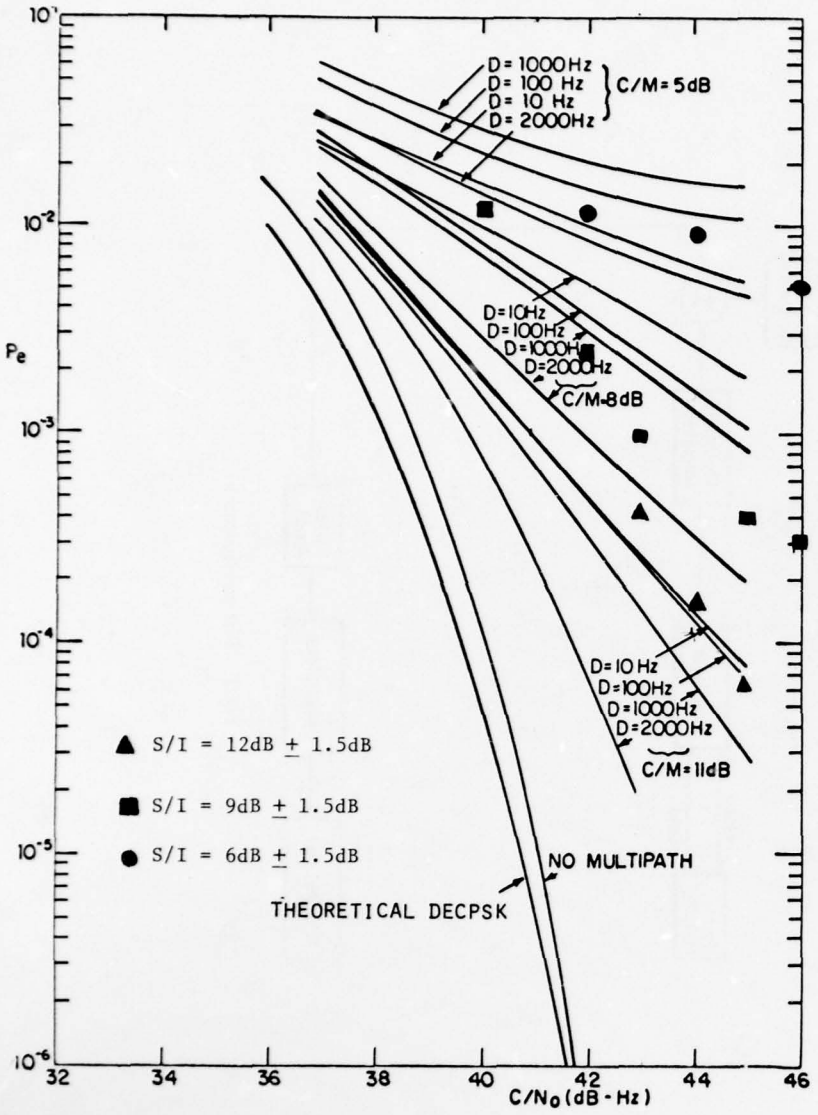


Fig.3 Hybrid I performance

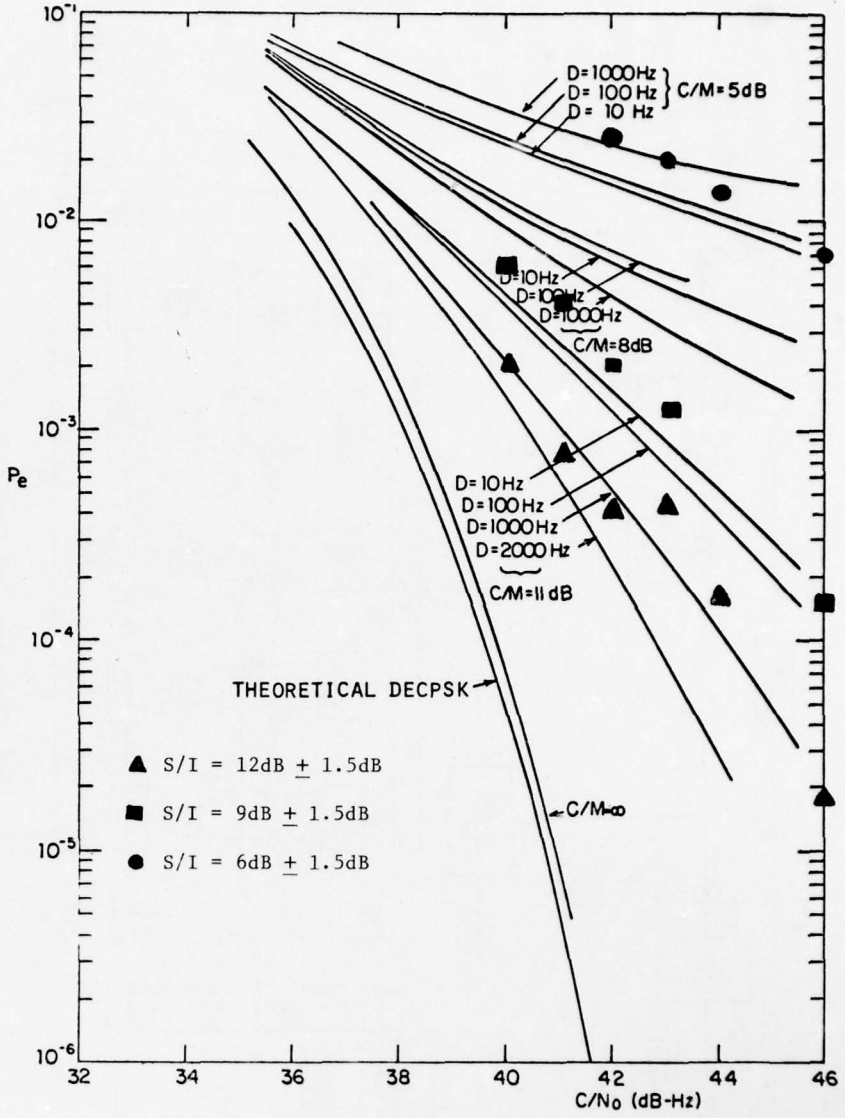


Fig.4 Hybrid II performance

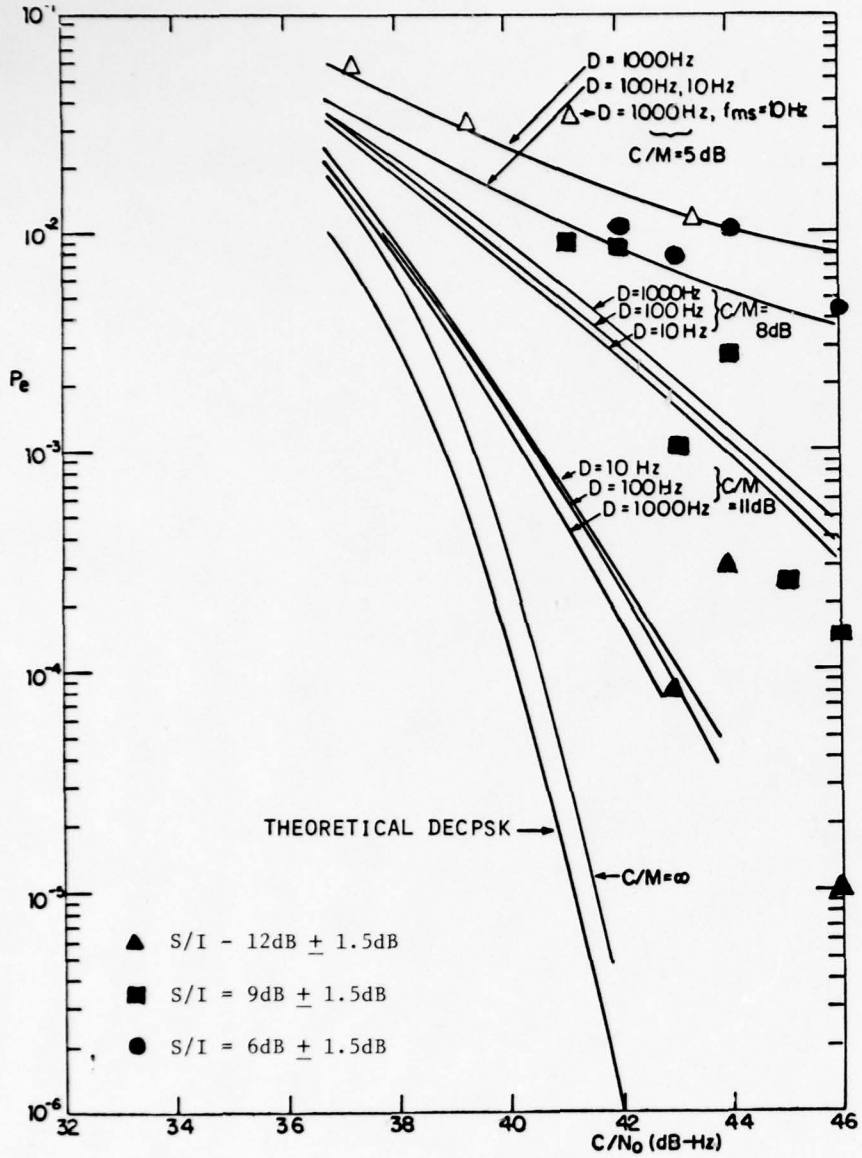


Fig.5 DPSK I performance

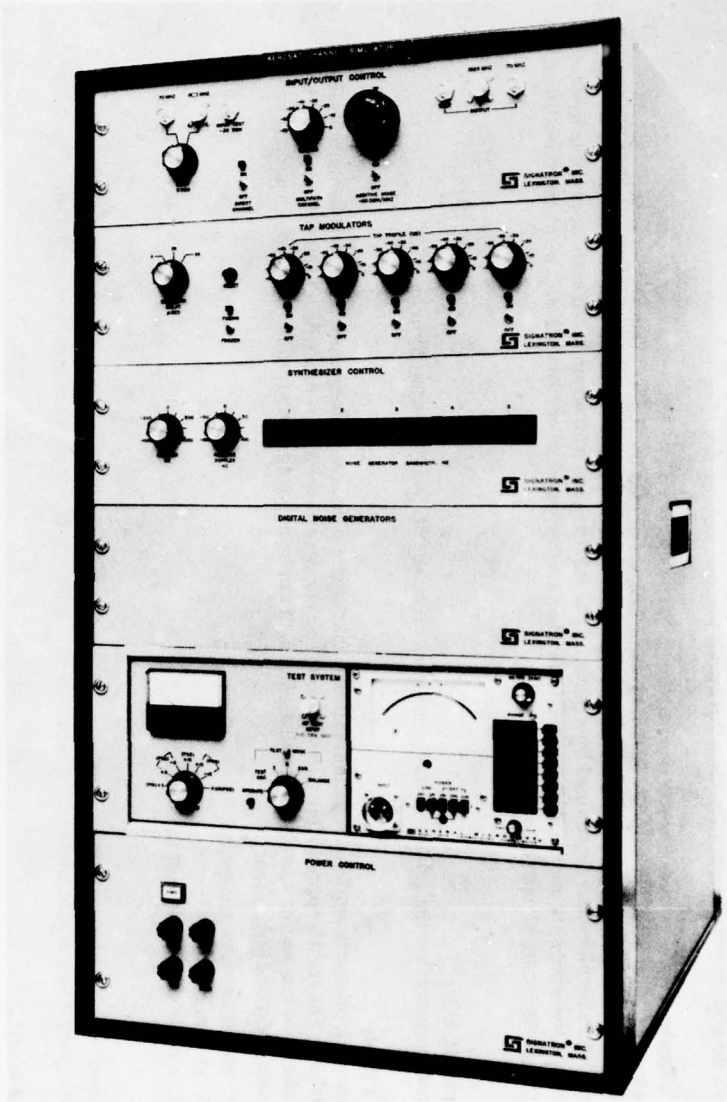


Fig.6 TSC channel simulator

DISCUSSION

J.Buchau, US

How do you objectively quantize voice intelligiblity?

Author's Reply

We use the 1000 word Phonetically Balanced Word List, developed by Harvard University for Voice Intelligibility Testing. For our tests, 400 words were randomly selected, and transmitted at 2.5 second intervals for each data point. Intelligibility was determined by a panel of trained listeners, who wrote down each word while listening to tapes of the word lists, recorded during the tests. The average panel score, for each 400 word list, constituted the intelligibility score of each data point.

L.A.Gerhardt, US

Please comment on the *applications* of your work on behalf of *Transportation Systems*.

Author's Reply

This work primarily applies to communications links where signals are transponded through a communications satellite. The results apply to *any narrow-band* (i.e., non-spread-spectrum) link which utilizes bit rates of a few kilohertz or less and/or narrow-band voice. The data can be applied to any (primarily subsonic) aircraft or ship communications link, given:

(a) model antenna pattern,
(b) system geometry,
(c) mobile performance characteristics.

J.Hagenauer, Ge

Have you got any results on the error structure of the bit errors, for instance burst and gap distributions?

Author's Reply

Yes – this data is available. It is published in the following Report:

"Boeing Commercial Aircraft Div., 1976 – Air Traffic Control Experimentation and Evaluation with the NASA ATS-6 Satellite, Report # FAA-RD-75-173, Vol.VI". This is available through NTIS. It is unclassified.

INVESTIGATION ON INFORMATION ERROR CAUSED BY TRAFFIC LOADING IN APPROACH AND LANDING SYSTEMS

Wolfgang Skupin

Institut für Nachrichtentechnik
Technische Universität Braunschweig
Schleinitzstraße 23
D-3300 Braunschweig

A Contribution of the "Sonderforschungsbereich Flugführung",
Technische Universität Braunschweig.

SUMMARY

In Approach and Landing Systems with Statistical Interrogation (ALSSI) the a/c transmit statistically distributed interrogations. These interrogations are interpreted and replied by the ground station. Because of the limited serving capacity of the ground station and of erroneous measurements with pulse coincidence the reply efficiency and the accuracy of these ALSSI depend on the number of service requiring a/c. This paper presents some methods which are suitable for the investigation of traffic loading problems.

For the evaluation of a required ALSSI serving capacity a suitable traffic model has to be developped. This model is based on the operational requirements and gives a defined standard traffic volume for an ALSSI.

For investigating the reply efficiency and the accuracy of an ALSSI several methods can be employed. By means of probability calculations the reply efficiency can be determined with comparatively low expense. For the determination of the accuracy it is useful to carry out real world system tests with a test a/c. Because of less expense the traffic load is simulated with these tests by pulse generators. More flexibility can be achieved by employing a software simulation. For this a computer model of the traffic loading has to be installed as well as a computer model of the system to be investigated. Some results achieved by employing these methods will be presented.

1. Introduction

Some of the modern Approach and Landing Systems are systems with Statistical Interrogations (ALSSI). For instance, to this group belong DME, SETAC and DLS, the German MLS. Figure 1 shows the principle of an ALSSI. In these systems the a/c transmit interrogations statistically. These interrogations are interpreted and replied by the ground station. From the replies the airborne information is achieved. Because of the limited serving capacity of the ground station and of erroneous measurements with pulse overlapping at the ground station antennas as well as at the airborne receiver antenna these ALSSI suffer from traffic loading. So the reply efficiency, the accuracy and the error rate of the data transmission in an ALSSI with data link depend on the number of a/c requiring service. The traffic loading problems increase if several channels have to be operated on the same frequency by code multiplexing because of frequency shortage.

2. Principle Traffic Loading Situation

For the case of code multiplexed channels figure 2 shows the principle traffic situation. The test a/c and a/c 1 represent those interrogators corresponding with ground station 1 (interrogation frequency f_a and reply frequency f_g with code 1). A/c 2 represents the interrogators corresponding with ground station 2 (same frequencies with code 2). These interrogations also load station 1 and the corresponding replies load the airborne receivers of the test a/c and a/c 1.

The traffic loading interference mechanisms are given in table 1. By pulse overlapping between interrogations of different sources erroneous measurements are produced and in some cases pulse drop outs originate. In the ground station further pulse drop outs arise caused by the blocking times of the transponder. One reason for installing these blocking times is the avoidance of multipath distortions. In order to suppress reflected signal components the receiver is blocked directly after detection of a pulse. Another reason for the employment of blocking times is the overload protection of the transmitter. So the minimum spacing between two replies is usually 60 µs. Moreover in most ALSSI the ground receiver is blocked during transmission of a reply in order to prevent RF interference between receiver and transmitter. All these blocking times typically amount to approximately 100 µs per reply. Further drop outs occur in case of coincidence of two or more replies. By priority criteria one reply is selected to be transmitted and the others are dropped. If there are air-to-air interferences, for instance with Radar systems like the Secondary Surveillance Radar (SSR), or in the case of code multiplexing additional distortions originate by pulse overlapping at the airborne receiver antenna. Further pulse drop outs are produced by the blocking of the airborne receiver during transmission of SSR replies. Erroneous measurements can be caused by acceptance and evaluation of non-corresponding replies. All these mechanisms effect the system reply efficiency and the accuracy. For the design of an ALSSI the required serving capacity has to be known. This capacity is given by the maximum traffic load to be handled by the system. For the evaluation of the maximum traffic load it is useful to elaborate a suitable traffic model.

3. Traffic Model

The intention of a traffic model is to define a standard traffic load, which is no more dependend on regional conditioning factors. This standard traffic load has to be the worst case loading to be handled by the ALSSI. The usual way of evaluating such a traffic model is to obtain the basic data by inquiring the air traffic volume of a region with a typical traffic situation. Starting from these data a standard traffic situation can be defined by applying the measured data to standardized traffic area units. Moreover the development of the traffic situation in the future has to be prognosticated and to be incorporated in the traffic model. Thus by considering all eventualities the maximum traffic load for an ALSSI can be evaluated and the required serving capacity is defined.

For this purpose the traffic model has basically to consider the following points:

- number and spatial allocation of the a/c, operation mode of the airborne transceivers
- number, performance and geographical sites of the airports
- frequency and channel allocation

The frequencies being available have to be assigned to the airports under consideration of applicable areas and protection areas (protection spacing between ALSSI allocations with identic channel assignment). Because of easier frequency management the complete traffic area is devided into small operation area units. To each area unit one set of frequencies is assigned. With that areas assigned to the same frequency set have to keep a minimum protection spacing in order to avoid RF interferences. Figure 3 shows an example with 4 frequency sets distributed to the traffic area under maintenance of the protection spacing. The minimum radius a is given by the system coverage and the required protection spacing. If all area units are equivalent concerning the traffic situation and the frequency availability the traffic model can be limited to one area unit only. From the number of ALSSI installations respectively the number of required channels inside one area unit and the number of frequencies being available the number of required codes can be derived directly. The evaluation of the number of airports and a/c for the traffic model is based on the operational requirements.

For the MLS contest such a traffic model was elaborated. Figure 4 gives the specifications for one operation area unit. As all area units are equally loaded the outlined traffic load is representative for any MLS installation in the complete traffic area. This traffic load was minimized by a skilful channel assignment.

4. Investigation Methods

From the traffic model a definite system traffic load results. For statements concerning the system performance of an ALSSI the influence of traffic loading has to be investigated. This can be achieved in different manners:

- Calculation of the reply efficiency by statistical methods

  This method gives merely knowledge about the reply efficiency. However, the calculations can be done easily and produce basic data quickly.

- Hardware simulation of the traffic loading by pulse generators

  This method produces real world test measurements of the system reply efficiency and of the system accuracy. However, the ALSSI has to be built up and to be operated. Parameter variation is usually difficult.

- Software simulation

  This method gives results for the system reply efficiency and the system accuracy. Parameter variation can be done easily. However, the results should be validated by additional investigations.

In the following these methods will be presented in more detail.

Calculation of the Reply Efficiency by Statistical Methods

By means of statistical calculations the pulse drop out probabilities can be determined for

- pulse overlapping at the ground station antennas
- blocking times of the ground station
- pulse overlapping at the airborne receiver antenna
- blocking times of the airborne receiver

Thus statistical data for the reply efficiency can be achieved. These data can be of basic importance for the determination of the data transmission error rate in an ALSSI with data link.

As an example for this method the principle way of calculating the reply efficiency of SETAC will be shown. Figure 5 gives the SETAC model used for the calculations. N_1 of the N_o interrogations are not disturbed by pulse overlapping and reach the ground station. N_2 of the N_1 interrogations are not cancelled by blocking times and will produce replies. N_3 of the N_2 replies are not disturbed by air-to-air interference with SSR pulses and reach the airborne receiver. Here some replies are dropped because of receiver blocking times during transmission of SSR replies. So only N_4 of the N_3 valid replies are received. The probability p_4 typically amounts to 0.98. The relation between reveived replies and transmitted interrogations N_4/N_o is the reply efficiency.

With this method reply efficiency data can be achieved quickly with comparatively low expense. This is useful for the system design. For DLS and SETAC this method was employed. The results were found to be in good congruency with those of subsequent measurements.

Hardware Simulation of the Traffic Loading by Pulse Generators

For a real world ALSSI test the system has to be operated with a test a/c. Then because of practicability reasons the traffic loading has to be simulated. This can be performed by pulse generators which transmit statistically distributed pulses. These pulses simulate the interrogations of other on-channel a/c (generator 1), off-channel a/c (generator 2) and the replies of other (off-channel) ground stations (generator 3). Figure 6 shows the block diagram of the simulation test set up in comparison with the traffic loading situation. Such a simulation test set up was employed for the DLS test measurements. So real system test data of the reply efficiency and the system accuracy could be achieved and a comparison between the loaded and the unloaded system was obtained. Typical test data is given in the figures 7 and 8. For increasing traffic load the reduction of the reply efficiency and the increase of the errors can be seen clearly.

From such measurements the capacity of an ALSSI can be determined by requiring either a minimum reply efficiency or a maximum error. From the curves the corresponding traffic loading can be read off, which gives the system capacity.

Software Simulation

For system optimization or parameter variation a real system test requires a great expense and is of low flexibility. So it is useful to perform a computer simulation. Figure 9 shows the block diagram of such a simulation. Two basic problems have to be considered because computer models have to be evaluated of the

- RF signal field (input signal of the ground station respectively of the airborne receiver)
- System to be investigated (evaluation procedures in the ground station and in the airborne equipment)

For the first point the time position of the pulses and the RF carrier amplitudes and phases have to be considered. As the different interrogators can be treated as statistical sources these quantities can be produced by random number generators. Additionally the

data- respectively the pulse-format of the system to be investigated has to be incorporated into the signal generation. These signals represent a certain traffic loading situation. Storage of them is useful as they are available for further system tests without renewed program run.

For the modelling of the ground system and the airborne equipment the data evaluation procedures have to be imitated carefully. This is especially true for the following points

- antenna patterns (significant for pulse overlapping)
- decoding and triggering
- register and memory engaging procedures
- data smoothing and prediction procedures in the airborne equipment

The results of the software simulation are data of the reply efficiency as well as data of the system accuracy. The information error (deviation between received information and correct information) can be determined directly. Moreover results of the different subsystems can be achieved easily, for instance separation of the down link errors and the up link errors. So this method is very suitable for system optimization because of its high flexibility. The obtained data, however, should be validated by supporting tests or data. Only then reliable prognostications can be achieved for system configurations not being realized yet.

5. Conclusion

All the methods presented above have been employed for investigations of different ALSSI. They have proved to be sufficient for the investigation of traffic loading problems in ALSSI. So they supply precious support for system design and system review.

Abbreviations - used in this paper:

a/c - aircraft

ALSSI - Approach and Landing System with Statistical Interrogation

DLS - DME based Landing System

DME - Distance Measuring Equipment

MLS - Microwave Landing System

SETAC - SEctor TACAN

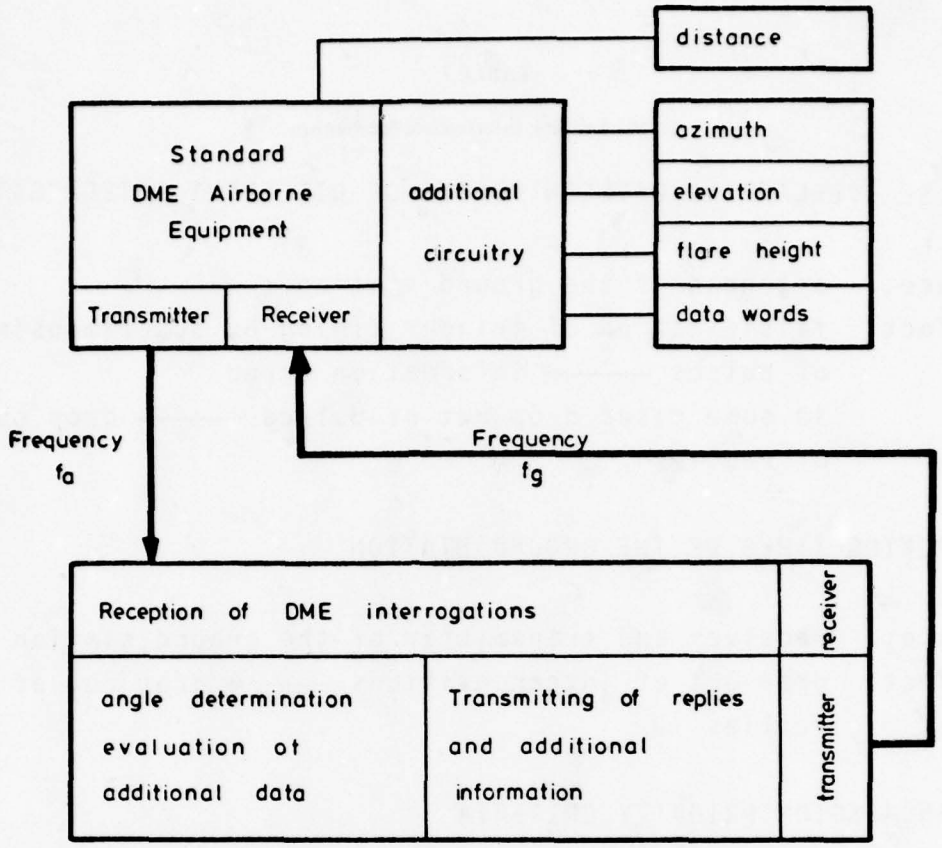


Fig.1 Signal flow for an ALSSI

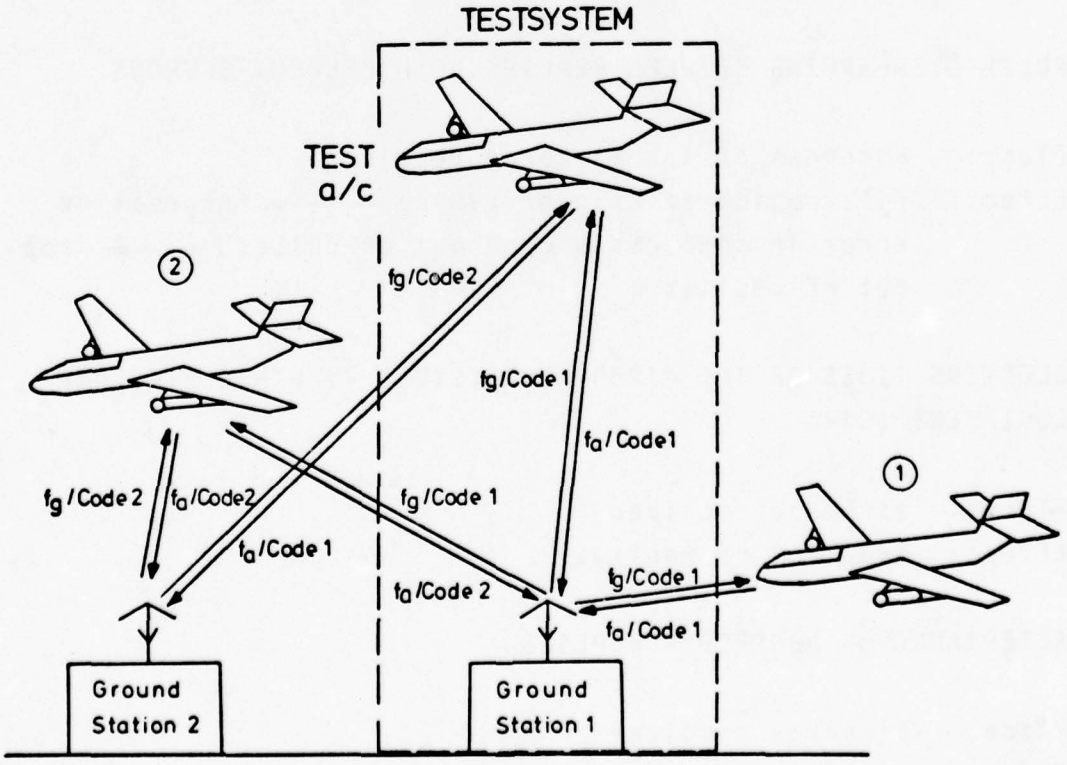


Fig.2 Principle traffic situation with code multiplex operation

TABLE 1

Traffic Loading Interference Mechanism

1. PULSE OVERLAPPING BETWEEN PULSES OF DIFFERENT INTERROGATIONS

   Place: antennas of the ground station
   Effect: falsification of trigger timing by superimposing of pulses ⟶ information error
   in some cases drop out of pulses ⟶ drop out of replies

2. BLOCKING TIMES OF THE GROUND STATION

   Place: receiver and transmitter of the ground station
   Effect: drop out of interrogrations ⟶ drop out of replies

3. TRANSMISSION PRIORITY CRITERIA

   Place: transmitter of the ground station
   Effect: drop out of replies

4. PULSE OVERLAPPING BETWEEN REPLIES OF DIFFERENT SOURCES

   Place: antennas of the airborne receiver
   Effect: falsication of trigger timing ⟶ information error in some cases drop out of pulses ⟶ drop out of replies

5. BLOCKING TIMES OF THE AIRBORNE RECEIVER BY OTHER AIRBORNE EQUIPMENT (SSR)

   Place: airborne receiver
   Effect: drop out of replies

6. ACCEPTANCE OF NONPROPER REPLIES

   Place: airborne receiver
   Effect: information error by evaluating wrong replies

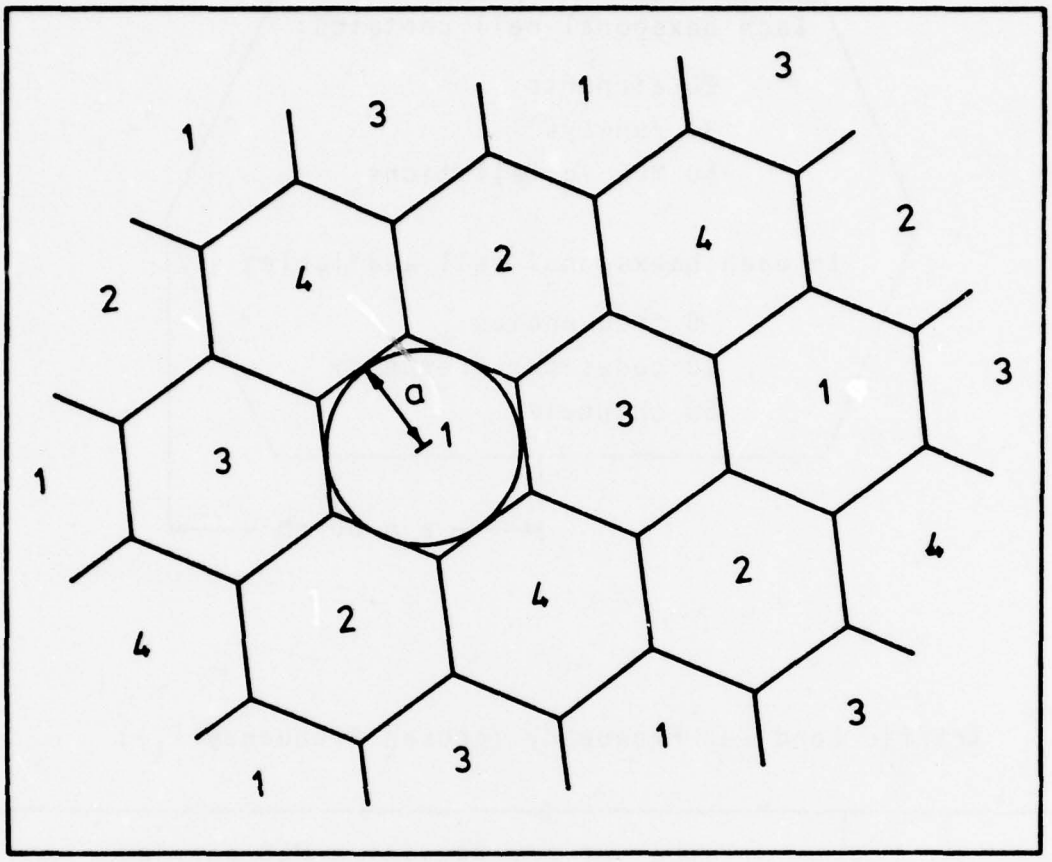
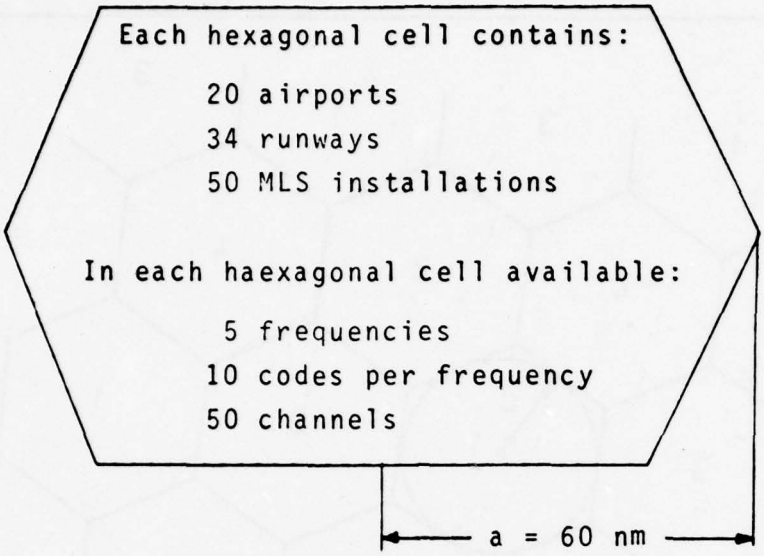


Fig.3 Schematic area unit distribution with 4 frequency groups

Specifications for a Hexagonal Cell



Traffic Load per Frequency (chosen Frequency f_1):

airport No	Code	Number of a/c requiring MLS service					
		on final approach	interm. appr.	stack	initial appr.	take off	ground
4	1	6	5	10	10	-	50
4	2	-	-	-	-	-	-
5	3	3	3	-	4	-	25
5	4	-	-	-	-	-	-
8	5	6	5	10	10	-	50
8	6	-	-	-	-	-	-
9	7	6	5	10	10	-	50
9	8	-	-	-	-	-	-
10	9	3	3	7	7	-	25
10	10	-	-	-	-	3	-

Fig.4 Specifications for any area unit of the MLS traffic model

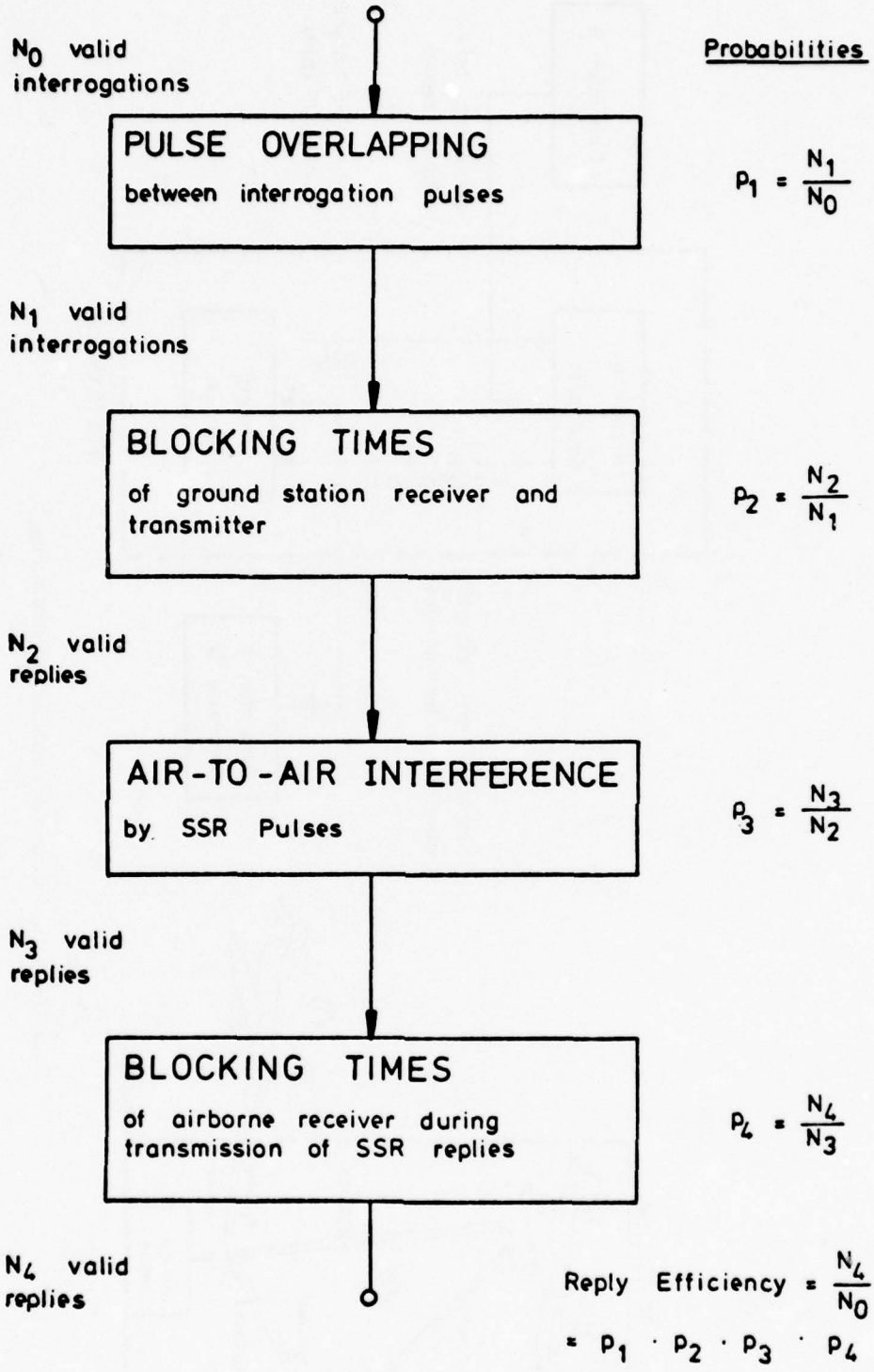


Fig.5 SETAC model for the calculation of the reply efficiency

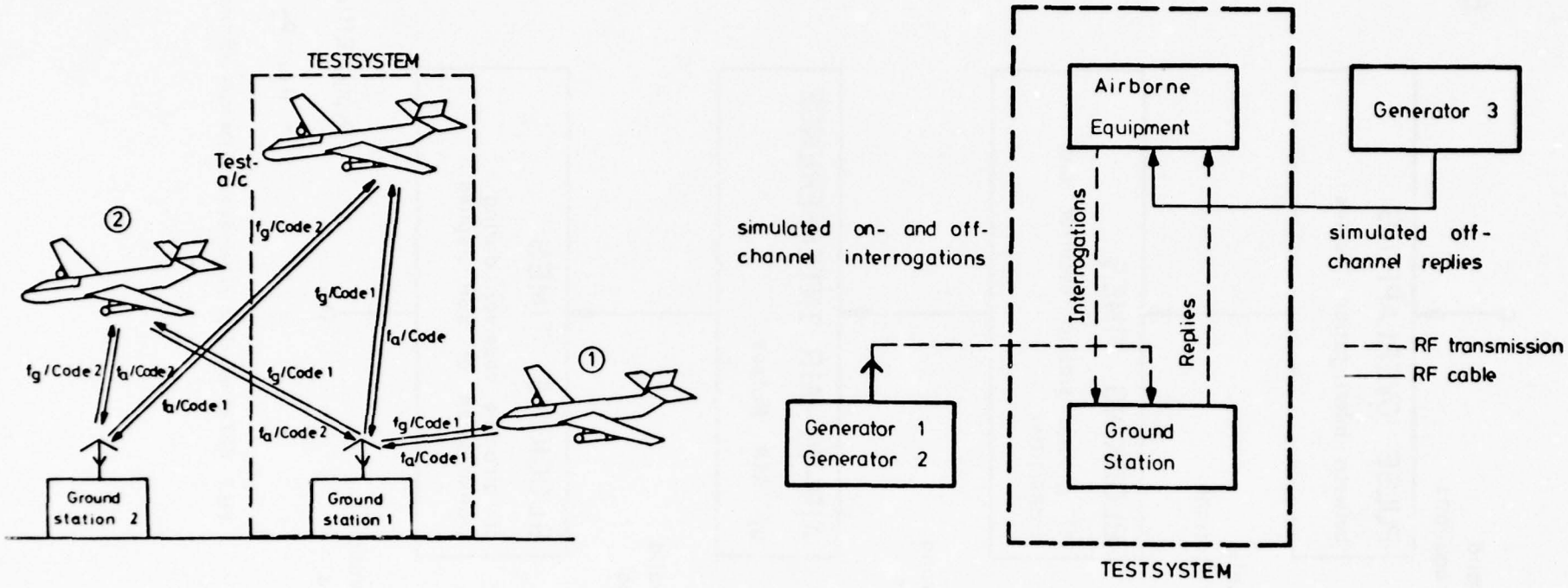


Fig.6 Test set up for hardware simulation of traffic loading

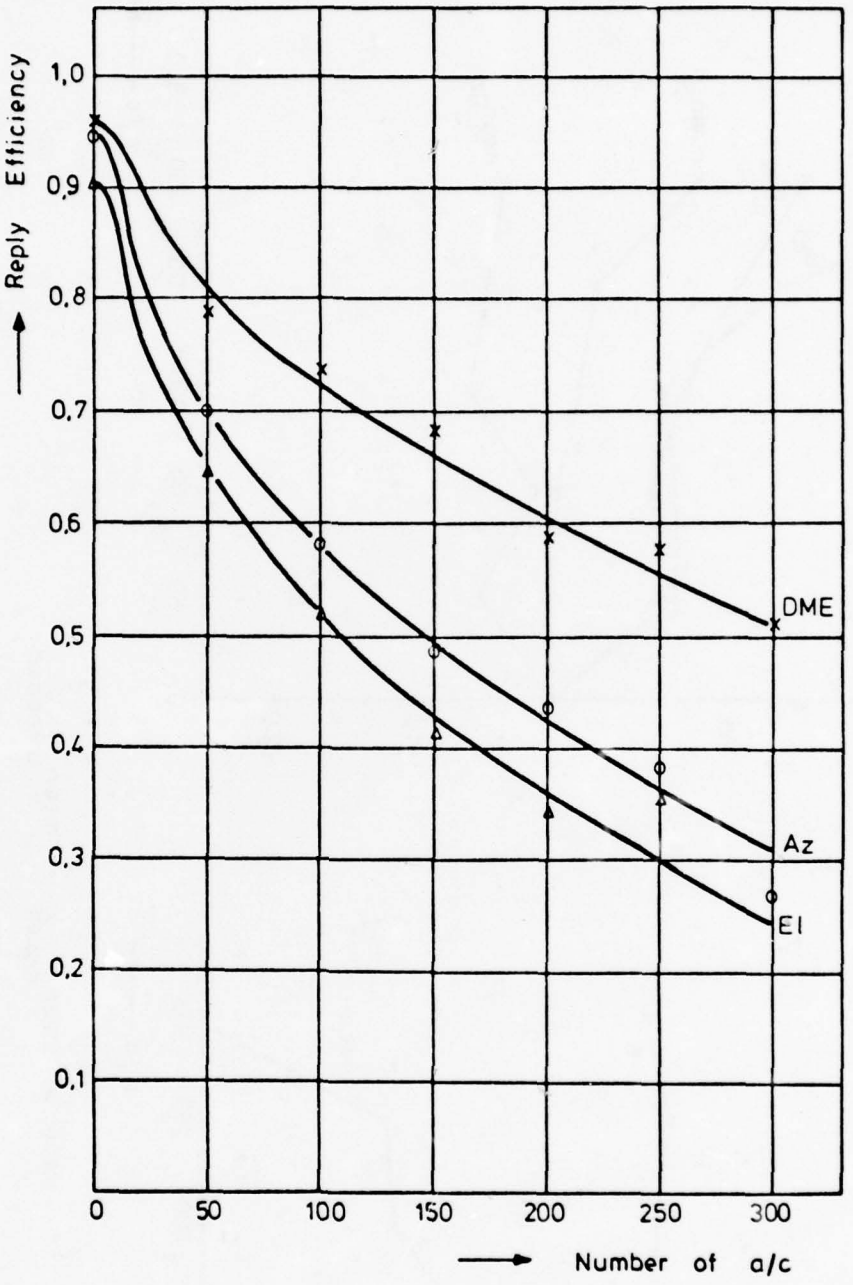
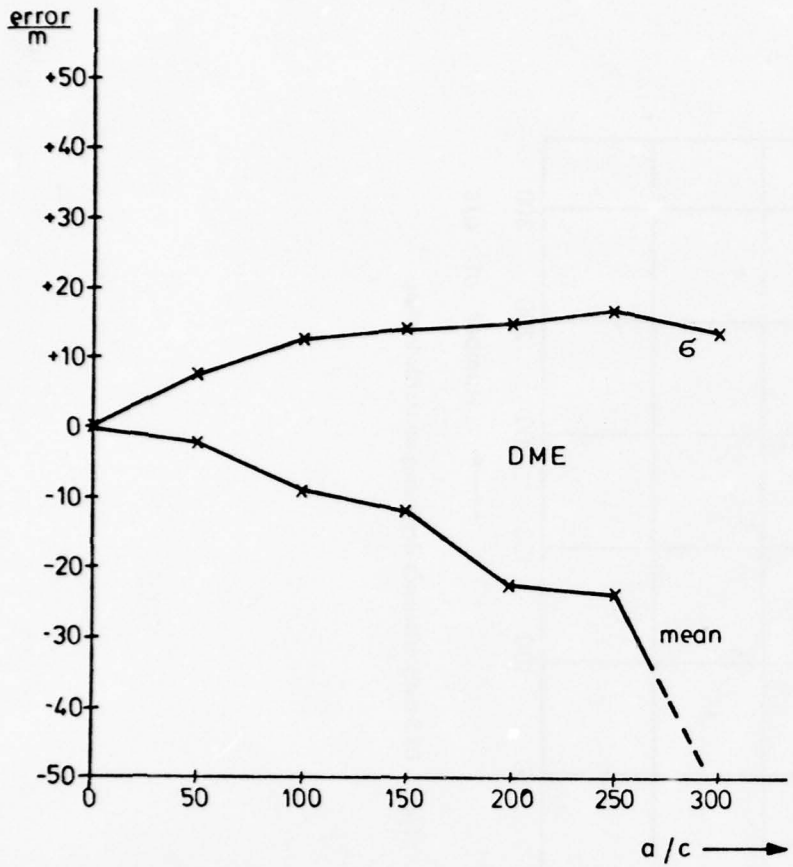
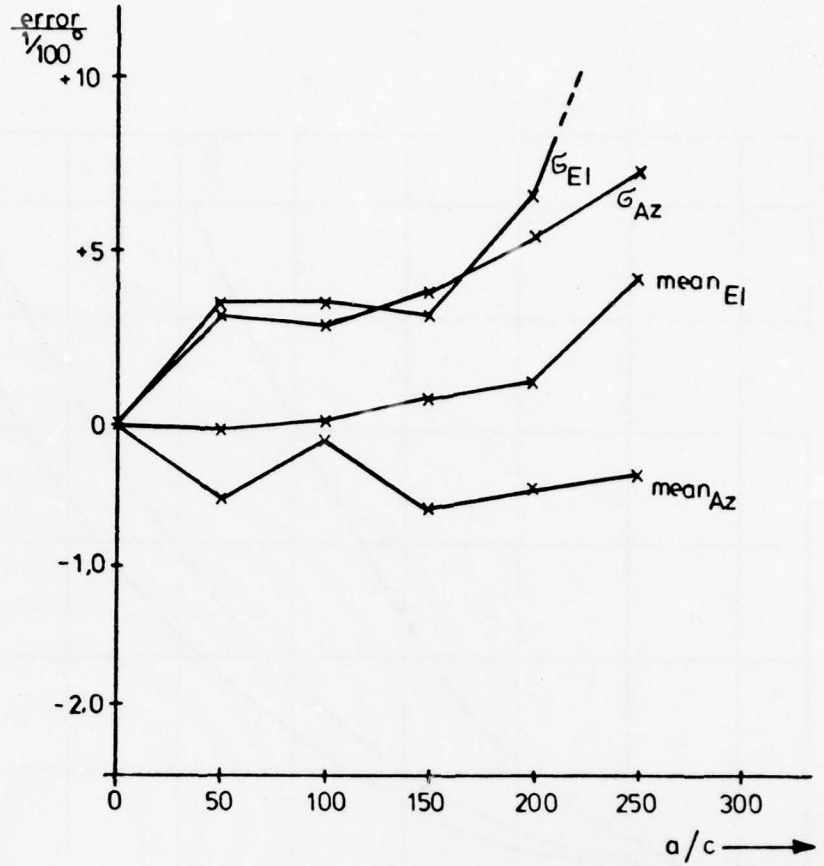


Fig. 7 DLS reply efficiency depending on traffic loading





$$\text{mean} = \text{mean}_{\text{loaded}} - \text{mean}_{\text{unloaded}}$$

$$\sigma = \sqrt{\sigma^2_{\text{loaded}} - \sigma^2_{\text{unloaded}}}$$

Fig.8 DLS accuracy depending on traffic loading

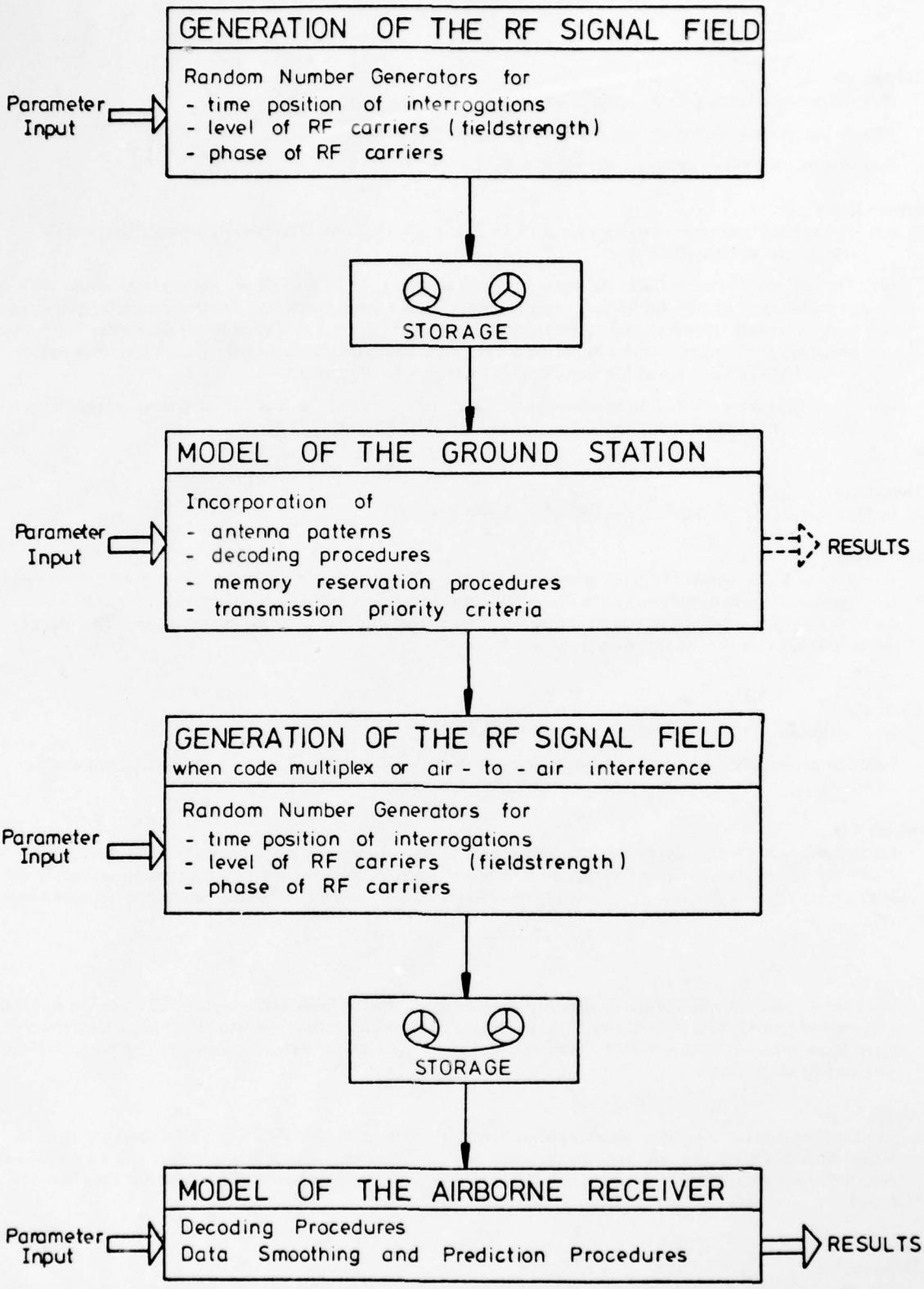


Fig.9 Block diagram of software simulation

DISCUSSION

J.Buchau, US

How often does a single aircraft, in the average, interrogate the ground system?

Why do you measure different reply efficiencies for DME, Az and El?

How are aircraft on final approach prioritized in this system?

Author's Reply

(a) The average interrogation rate per a/c is 15/sec in our investigations. This is a rate between the rates of commercial airborne DME-sets.

(b) The difference between DME- and angle-reply efficiency is given by the fact that the angle information has to pass additional circuits, for instance angle processors with a limited capacity. Another reason is given by the priority criteria. However, principally these effects could be deleted by a suitable system design. The reply efficiency for El is less than for Az, as the interrogation has to be accepted by the DLS-A-station as well as by the DLS-E-station for an El-reply, but only by DLS-A for an Az-reply.

(c) A/c in final approach have an increased interrogation rate of 50/sec. So they will have an increased number of replies. Then typical values for DLS replies/sec are: DME: 35; Az: 27; El:24.

E.Meinel, Ge

In what manner do you determine the phase of the RF carriers?

Author's Reply

The phase of RF-carriers is of importance only in the case of pulse overlapping. In the hardware simulation we take the pulse overlappings as they come. In the software simulation we generate a random number between 0 and 360 for each interrogation which represents the carrier phase related to the pulse phase of the test a/c. Thus the phase differences between the pulse-carriers are assigned.

P.Form, Ge

What is the rate of SSR-interrogations you did assume?

I ask this because the number of civil and military Airport surveillance radars and en route-radars continuously increases.

Author's Reply

Corresponding to the considerations made in the paper 12 we assume a maximum number of SSR-replies of 1200/sec. This maximum rate is fixed by the SSR transponder design, no matter how much interrogations occur. With a reply rate of 1200 SSR-replies /sec the typical reduction of the reply efficiency amounts to approximately 3%.

P.Form, Ge

What are the blocking times of the ground station receiver specified with respect to multipath? Even in a multipath environment like Salzburg Airport (Austria), we observe multipath time delays of frequently 20 μs, sometimes of up to 50 μs of considerable amplitude. Also SSR side lobe suppression technique applies blocking times of 25 μs with respect to multipath.

Author's Reply

The blocking times of the ground receivers are different in different ALSSI. However typical times are 20 μs to 60 μs. If the blocking times are very long they provide a good multipath suppression, but they effect a considerable drop out of interrogations. So a compromise between multipath resistance and traffic loading handling has to be found.

R.M.Harris, UK

Referring to the field trials (depicted in Figure 6 of the Preprint Paper) I understand that you employed Generator 2 and Generator 3 to simulate, respectively, blocking of the Ground Station 1(on Code 2) by Aircraft 2, and blocking of the Test Aircraft by replies from Ground Station 2 (to interrogations from Aircraft 2) using Code 2. These Generators were assumed to be uncorrelated, whereas in reality there will be a time relationship between the blocking of Ground Station 1 and the blocking of the Test Aircraft. The time delay between blocking of Ground Station 1 and of the Test Aircraft will be dependent on the geometrical path lengths between Aircraft and Ground Stations but in the short term this time delay will be approximately constant. I therefore question the assumption that Generators 2 and 3 properly simulate truly random blocking pulses; the probability calculation presented in Figure 5 is valid only if the blocking is truly random (uncorrelated).

Author's Reply

Actually interrogations of a/c 2 (simulated by generator 2) and replies of ground station 2 (simulated by generator 3) are correlated and so should be generators 2 and 3. However, this correlation would be of interest only if there is a synchronization or correlation between the different interrogations. But in fact all interrogations are uncorrelated and the interrogators can be treated as statistical independent sources. In the case of a random spatial distribution of the a/c this statistical independency will effect non-constant time delays between a/c interrogations and ground replies even in the short term. Because of this fact we found out that the loading of the test system with truly random pulse generators is acceptable. So the probability calculation of Figure 5 and the test set up of Figure 6 give results with good approximation to real world conditions.

M.Schilliger, Fr

Compte tenu des figures 7 et 8 de la présentation, je constate pour une charge de 250 avions une précision moyenne de 25 mètres avec un taux de réponse de la balise voisin de 55%. Par quel moyen est obtenue une telle précision?

Quelle est la cadence d'émission de la balise?

L'impulsion utilisée est-elle de forme gaussienne?

Author's Reply

This precision was achieved with a special precision DME. This PDME uses a $\cos^2$ – pulse shape and is fully ICAO compatible including the power spectrum. However, as I am not the system design engineer, I cannot give too much detailed information. The only point I can say is that this accuracy is obtained by special precision trigger techniques.

In our field test we have been employing KING and Collins commercial airborne DME sets as well and we also achieved results with a comparable accuracy.

NEW DEVICES FOR DIGITAL COMMUNICATIONS IN AVIONICS

F. I. Diamond, H. J. Bush, J. A. Graniero
Rome Air Development Center
Griffiss Air Force Base, New York 13441

SUMMARY

The rapid progress in microelectronics has been described as a revolution in electronic technology. Integrated circuits provide building blocks for complex signal processing, and new devices such as surface acoustic wave devices and charge-coupled devices offer new capabilities in signal processing. And with the advent of the microprocessor as an all-purpose programmable device, circuit designers are being offered new opportunities for ingenuity. In addition, steady progress is being accomplished in areas such as field effect transistors for both low-noise receivers and efficient high-power amplifiers; printed circuit techniques are being successfully applied to conformal antennas. With the maturing of spread spectrum and error correction techniques, and the reduction to practice of adaptive filtering in both the frequency and spatial domains, a revolution in digital communications for avionics is indeed occurring.

1. INTRODUCTION

Future trends in digital communications technology for avionics are embodied in such programs as the U.S. Army's Position Location-Reporting System (PLRS) and the U.S. Air Force/Navy Joint Tactical Information Distribution System (JTIDS). (BAHOR, H. H. et al, 1975; ELLINGSON, C. E. 1975) Both of these are multiple access systems designed to handle a large number of users. By providing commanders with such data as aircraft status, position information, weather, and target information, a powerful command and control capability is provided. Pilots gain more effectiveness by the availability of information of special interest to their missions. Similar capabilities for ships, tanks, etc., could also be achieved.

For systems such as these, a number of challenging requirements are imposed. In addition to multiple access, jam resistance and low probability of intercept (LPI) are needed because of growing concern for enemy countermeasures. (GREINKE, E. D., 1977; OTERE, R. J., 1977; MCALLISTER, N. F., 1977) Data must be capable of encryption, and in some cases, data rates greater than 100 kbs with adequate jam resistance may be required. Although aircraft generally employ onmidirectional antennas, directional antennas may be needed for some tactical situations or for jam resistance and LPI. Furthermore, sufficient radiated power is desired in order to maximize range performance and anti-jam capability.

To meet the above requirements, a host of signal processing techniques are available. Jam-resistance is obtained by spread spectrum modulation techniques--pseudo-noise phase shift keying, frequency hopping, time-hopping, or a combination of these. This, combined with error-correction coding, also minimizes self-interference. Multiple access requirements can be met through code division, frequency division or time division techniques or a hybrid combination. (CAHN, C. R., 1973) Performance can be augmented by adaptive filtering techniques--automatic notch filters to overcome narrow-band jammers, and adaptive matched filters to mitigate against such adverse effects as multipath. (MCCORD, J. M. et al 1976) Adaptive spatial filtering can either supplement the spread spectrum anti-jam protection or provide considerable protection itself by automatically forming nulls in the direction of jamming signals. (APPLEBAUM, S. P. 1976)

The practical implementation of these techniques requires the overcoming of some important hardware problems, involving considerations of bandwidth, dynamic range, and amplifier linearity. This also requires considerations such as spectral purity of oscillators and low spurious mixer outputs. In addition, circuit complexity, reliability and cost are important factors.

2. NEW DEVICES

Technological advances in microelectronics have made possible large scale integration of digital circuits that are ideally suited for signal processing, particularly techniques involving matched filters and tapped delay lines. Some typical characteristics of microelectronic circuitry as reported by Roberts are shown in Figure 1. (ROBERTS, J. B. G., 1977) Moreover, surface acoustic wave devices (SAWs) and charge-coupled devices (CCDs) are providing new or improved capabilities with significant reduction in complexity, compared to digital circuit approaches.

2.1 Surface Acoustic Wave Devices (HAYS, R. M., 1976; HICKERNELL, F. S. et al, 1977; CLAIRBORNE, L. 1977; BUSS, D. D. et al 1975)

The surface acoustic wave device is basically an analogue delay line with a propagation delay on the order of 3 μ sec/cm, with an operating frequency in the range of 10-1500 MHz, fractional bandwidths up to 50% are possible. The device essentially consists of two or more inter-digital transducers which are metal electrode-arrays deposited on a piezo electric substrate. The input transducer converts input signals to an acoustic Rayleigh wave which propagates along the surface. Output transducers as well as interdigital transducer taps can be placed anywhere along the propagation path.

The major advantage of these devices are small size, high reliability, large dynamic range (over 100 db) and linear phase response. They are used for matched filtering for spread spectrum signal processing and for spread spectrum signal generation, delay lines, filters, oscillators and frequency synthesizers.

As band-pass filters, Q's of 1-1000 and insertion losses of less than 10db (depending on bandwidth) have been designed. Some examples are shown in Figure 2. These filters are generally realized as transversal filters, but with SAW resonators, narrowband filters with bandwidths on the order of a few KHz ($Q > 10{,}000$) have been demonstrated.

SAW oscillators with high stability and simplicity of design have been produced. However, a major application has been in the field of frequency synthesizers. In one technique, the harmonics of a low frequency oscillator are generated, and SAW filters are then used to extract clean spectral lines. A second approach involves the mixing of chirp signals combined in a SAW tapped delay line.

An important application of SAWs is in spread spectrum communications. SAW matched filters have been utilized for synchronization as well as for wave generation and reception. For pseudo-noise waveforms, chip rates as high as 50 MHz with modest length codes (128 chips) have been demonstrated. Devices for programmable SAWs increase the flexibility and potential of SAWs for this application. A hybrid piezo-electric SAW with a thin film control circuit has demonstrated pseudo-noise code programmability for a 10 MHz, 128 chip code. A SAW tapped delay line whose taps are both phase and amplitude programmable has been reported. The SAW line and control circuitry are contained on a single LSI chip in monolithic form.

2.2 Charge-Coupled Devices (BUSS, D. D. et al, 1975, 1976; PATTI, J. J. et al, 1974; LIECHTI, C. A. 1976)

Charge-coupled devices (CCDs) are analog sampled-data delay lines; as such, they are ideally suited to many sampled-data signal processing functions such as analog filters and analog and digital memory. In the CCD, surface electrode potentials are used to store and manipulate charge along the semi-conductor surface. Analog signals in the form of charge are stored in a potential well underneath these electrodes. The CCD relies on clocking of the potentials to create a delay line. Discrete stages with means for tapping the delay line can be provided. Because the clock rate controls the delay time or rate of data propagation along the device, the CCD is a variable delay line. The CCD has a relatively large dynamic range (70 db). At present, operating frequencies are limited to less than 10 MHz.

Some features of CCDs are low cost, flexibility, large bandwidth and long delays (seconds), precise time control, and low power consumption. Important applications of CCDs are matched filtering and adaptive correlators. The use of a 100-stage CCD device matched to a pseudo-random noise sequence has been reported. Correlators can be realized with a pair of CCDs; a reference signal is loaded into one CCD and a second signal is clocked through a second CCD, with sequential multiplication and summation. Similar in principle to the correlator described above is the programmable transversal filter. In this case, the desired impulse response can be determined, and the reference signal accordingly adjusted.

A comparision of some SAW and CCD parameters is shown in Figure 3.

2.3 Solid State Amplifiers (TSENG, H. Q. et al, 1976; OKEAN, M. C. et al, 1977; LORENZO, D., 1978; MUNSON, R. E., 1974)

Steady advances in solid state amplifiers using bipolar transistors and GaAs field-effect transistors have been made in the past decade. Such devices are capable of low-noise amplification and high-efficiency power amplification.

At frequencies up to 2 GHz, bipolar transistor technology will support most requirements. From 2 to 4 GHz, bipolar technology will support more modest requirements. For operating frequencies in C, X, and Ku bands, high-power, high-efficiency microwave amplification using GaAs FETS has been demonstrated. Typical transistor capabilities are shown in Figure 4.

Modules have been combined at UHF and L-Band, with average power levels of several hundred watts. Some typical module characteristics are shown in Figure 5. These modules have been combined into power amplifiers. For example, Westinghouse advertses an L-Band amplifier that combines several 100 watt modules into a 1 KW amplifier in a volume of less than 0.5 cubic feet.

A variety of building blocks have evolved during the past decade for use in low-noise microwave front ends. The major emphasis has been on the parametric amplifier, the bipolar transistor and the FET amplifier. The state-of-art of low noise receivers is shown in Figure 6, but noise figures as low as 1 db have been measured on GaAs FET amplifiers. It is expected that evolutionary refinements in GaAs technology will lead to highly reliable, extremely small metal semi-conductor configurations extending to higher frequency ranges.

2.4 Microstrip Antennas (KLEIN, L., 1978; TORRERO, E. A., 1976)

Conformal, thin antennas for high velocity aircraft and missiles can be achieved with a printed circuit board antenna. Made with the same low cost photo-etch process used to make printed circuits, the so-called microstrip antenna would have a low-profile required to minimize aerodynamic effects and mechanical modification of the aircraft or missile.

A microstrip phased array can be fabricated, which incorporates the radiating system and the microwave feed system photo-etched on a printed circuit board. Solid state components can be directly added to the board to provide phase shifters, amplifiers, etc., for an integrated antenna and receiving system.

By adding pin diodes to an array of flat thin microstrip antennas, an electronically steerable beam can be achieved. (An experimental model of such an array has been built and tested, with results in agreement with theoretical predictions.) Figure 7 shows a 16 element C-band array employing a newly designed 3-bit phase shifter and quarter-wavelength microstrip elements. The phase shifter consists of switched-line phase shifters for the 90^{0} and 180^{0} bits and a loaded line phase shifter for the 45^{0} bit. This design provides the lowest loss of the various designs investigated that were compatible with the allocated area between elements. The phase shifter uses 10 chip PIN diodes and 1 chip capacitor.

The element design is a quarter-wavelength shorted patch radiator. This element was selected over the conventional patch because of its broader beamwidth and the increased area available for the phase shifter. Febrication of the complete array is very simple. The printed circuit boards are first drilled at the element shorting locations. The drilled boards are then plated to achieve plated-through holes.

Microstrip elements and feed networks (both RF and DC) are then etched and the diodes, capacitors and RF connector installed. The completed experimental model had the following characteristics:

Theoretical gain ($4\pi A/\lambda^2$)		=	17.2 dBi
Measured gain		=	15.7 dBi
Total Losses		=	1.5 dB
Bandwidth		=	200 MHz
Scan Volume,	E-Plane,	=	120°
Scan Volume,	H-Plane,	=	90°

A major limitation of the microstrip antenna is the bandwidth. The bandwidth can be substantially increased with a corresponding increase in the thickness of the array. By increasing an L-Band microstrip antenna to a thickness of 1/4 inch, bandwidths of about 100 MHz have been demonstrated. On the other hand, if a very thin antenna is desired, other methods of increasing the antenna bandwidth must be investigated. Some approaches are the use of high dielectric constants, increase in radiator inductance, or the addition of reactive components, with a corresponding decrease in antenna efficiency.

2.5 Microprocessors (BASTELO, R. A. et al, 1978; COPELAND, M. A., 1977; GE CO., 1976)

Microprocessors are a remarkably versatile new tool. They can lower the cost and increase the flexibility of electronic equipment and are ushering in a new era for digital designers.

Together with memory and peripheral circuitry, microprocessor (μP) chips form complete microcomputers. The high chip density needed for microprocessors has generally been obtained by the use of some form of MOS (metal-oxide semiconductor) technology. Recently, power-saving CMOS μP's are predominantly used when high speed is not required. Microprocessors that use bipolar technology offer the highest speeds. However, the bipolar units generally are not complete microprocessors. In most cases, several bipolar-μP "slices" must be combined to obtain the capabilities offered by a single chip.

Perhaps, the most exciting and promising application of μP's for communications is the capability of multi-level distributed processing for control of the multiple functions performed in modern communications transceivers. In multi-level distributed processing, the μP chip and periphered hardware is matched to computational complexity and execution speed required at each level of control. A hierarchical architecture is employed which implements master control of the distributed processors. Interprocessor communication can be accomplished via high speed shared memory.

The types of functions that are candidates for μP control are shown below, categorized by major transceiver function:

(a) RF Control:

(1) Perform electronic beam steering/switching

(2) Implement power control for LPI

(3) Control AGC

(b) Adaptive Spatial Filtering:

(1) Implement constraints on beam/null formation

(2) Pre-condition weights in multi-user environment

(3) Vary integration time or convergence time in adaptive feedback loops in response to changing environment.

(4) Vary threshold/decision point of desired/undesired signals

(c) Adaptive Temporal Filtering:

(1) Change bandwidth

(2) Vary convergence time

(3) Vary detection threshold

(4) Exercise notch filters

(d) Data Control:

(1) Change through-put in response to environment or priorities

(2) Vary processing gain

(3) Change code

(e) System Control:

(1) Exercise priorities

(2) Interface with peripheral equipment and other systems

(3) Perform trouble shooting and fault isolation; determine equipment status

Adaptive filters using microprocessors have been implemented for electronically programmable transversal filters for such functions as adaptive equalization and cross correlation of signals. Microprocessors have also been used for the control of multiple function phased array antennas and for adaptive pattern shaping. They are also being integrated with microstrip antennas. For example, a Motorola 6800 microprocessor with associated memory calculates the number of phase shift increments required for each element, based on directional input commands.

3. A FUTURE ADAPTIVE TRANSCEIVER (RADC, 1976)

It can be envisioned that this host of signal processing techniques and technologies can be integrated into a fully adaptive communication transceiver. The transceiver would appear as shown in Figure 8. The key characteristic would be its complete adaptability to changes - changes in the channel, data throughput requirements, A/J margin and type and direction of jamming. This adaptability is made possible by the programmable nature of the devices described above and the control provided by microprocessors.

Figure 9 breaks down the receiver into sub-assemblies, based upon cost, showing where the new devices and technologies would be applied for implementing the various functions. As can be seen, there is the possibility of a mixture of digital and analog techniques when sampled data signal processing and lowest cost is desired. Where the digital and analog implementation ratio is about 1 : 1, both are listed. Frequency hop synthesizers can be implemented by both hybrid digital and analog techniques. The adaptive array processor, depending upon the algorithm, could be one or the other. However, for programmable correlators, the analog devices, SAW's, presently, with CCD's in the new future, appear to offer the best cost/performance ratio. Presently, digital devices appear to be more favorable for error control.

REFERENCES

APPLEBAUM, S.P., 1976, "Adaptive Arrays", IEEE Transactions on Antennas and Propagation, Vol. AP-24, No. 5.

BAHOR, H. H., and VIARS, T. C., 1975, "Position Location-Reporting System (PLRS)", Proceedings of National Telecommunications Conference, New Orleans.

BUSS, D. D., ET AL, 1975, "Communications Applications of CCD Transversal Filters", Proceedings of National Telecommunications Conference, New Orleans.

BUSS, D. D., ET AL, 1976, "Charge Coupled Devices for Analogue Signal Processing", IEEE Conference Publication 144, "The Impact of New Technologies in Signal Processing", Aviemore, Scotland.

BASTELO, R. A., ET AL, 1977, "Distributed Processor Control of a Multiple Beam Adaptive Array for Telemetry, Command and Control of Airborne Vehicles (RPVs)", AIAA Symposium, Los Angeles, CA.

CAHN, C. R., 1973, "Spread Spectrum Applications and State-of-the-Art Equipments", Spread Spectrum Communications, AGARD Lecture Series No. 58.

CLAIRBORNE, L., 1977, "State-of-the Art of CCD and SAW Technologies and Potential Applications", AGARD AP Symposium on Impact of Charge Coupled Devices and Surface Acoustic Wave Devices on Signal Processing and Imagery in Advanced Systems", Ottawa, Canada.

COPELAND, M. A., 1977, "Interaction Between Microprocessors and Custom LSI", Microprocessors and Their Applications, AGARD Lecture Series No. 87.

DILORENZO, J. V., 1978, "GaAs FET Developments - Low Noise and High Power", Microwave Journal.

ELLINGSON, C. E., 1975, "Joint Tactical Information Distribution System", Proceedings of the National Telecommunications Conference, New Orleans.

GENERAL ELECTRIC CO., 1976, "Advanced Transceivers Design Study", Final Report, Contract F30602-75-C-0130, Rome Air Development Center.

GREINKE, E. D., 1977, "Management/Coordination of Data Link Development and Production in the Department of Defense", Signal Magazine.

HAYS, R. M., and HARTMANN, C. S., 1976, "Surface Acoustic Wave Devices for Communications", Proc. IEEE, Vol 64, No. 5.

HICKERNELL, F. S., and BUSH, H. J., 1977, "The Monolithic Integration of Surface Acoustic Wave and Semiconductor Circuit Elements on Silicon for Matched Filter Device Development", AGARD AP Symposium on Impact of Charge Coupled Devices and Surface Acoustic Wave Devices on Signal Processing and Imagery in Advanced Systems, Ottawa, Canada.

KLEIN, L., and SANFORD, G. G., 1978, "Recent Developments in the Design of Conformal Microstrip Phased Arrays", Proc. IEEE Conference, London.

LIECHTI, C. A., 1976, "Microwave Field-Effect Transistors - 1976", IEEE Trans. on Microwave Theory and Techniques, Vol. MTT-24, No. 6.

MCALLISTER, N. F., 1977, "Wideband Digital Transmission Systems in Ocean Surveillance", Signal Magazine.

MCCORD, J. M., and WIDROW, B., 1976, "Principles and Applications of Adaptive Filters: A Tutorial Review", IEEE Conference Publication 144, "The Impact of New Technologies in Signal Processing", Aviemore, Scotland.

MUNSON, R. E., 1974, "Conformal Microstrip Antennas and Microstrip Phased Arrays", IEEE Trans. on Antennas and Propagation, Vol. AP-22, No. 1.

OKEAN, M. C., and KELLY, A. J., 1977, "Low Noise Receiver Design Trends Using State-of-the-Art Building Blocks", IEEE Trans. on Microwave Theory and Techniques, Vol. MTT-25, No. 4.

OTERE, R. J., 1977, "The DOD Digital Data Link Commonality Program", Signal Magazine.

PATTI, J. J., and ROEDER, A. W., 1974, "An Adaptive Spread Spectrum Correlation Receiver", Proc. of IEEE Symposium on Adaptive Processes, Phoenix, AZ.

ROBERTS, J. B. G., 1977, "The Roles for CCD and SAW in Signal Processing", AGARD AP Symposium on Impacts of Charge Coupled Devices and Surface Acoustic Wave Devices on Signal Processing and Imagery in Advanced Systems, Ottawa, Canada.

TORRERO, E. A., 1976, "An Introduction to Microprocessors", Electronic Design.

TSENG, H. Q., ET AL., 1976, "Microwave Power GaAs Amplifiers", IEEE Trans. on Microwave Theory and Techniques, Vol. MTT-24, No. 12.

FIGURE 1. COMPARISON OF MAJOR LSI TECHNOLOGIES

TECHNOLOGY	CMOS	TTL LS	ECL	I^2L
DENSITY (GATES/CHIP)	100-1000	100-500	30-300	300-500
PROP. DELAY (^{n}SEC/GATE)	10-50	2-10	0.5-2	5-100
POWER (MW/GATE)	10^{-3} @ 1 KHz 1 @ 1 MHz	1-5	10-100	10^{-4}-10^{-1}
NOTE	OPTIMUM COMPROMISE	LOW PRICE	HIGH SPEED	NEW TECHNOLOGY

FIGURE 2. REPRESENTATIVE SAW BANDPASS FILTERS

FREQUENCY	34.5 MHz	328 MHz	140 MHz
3 DB BW	1.6 MHz	3 MHz	40 MHz
INSERTION LOSS	0.65 DB	5 DB	20 DB
PASSBAND RIPPLE	<0.04 DB	0.5 DB	0.1 DB

FIGURE 3. TYPICAL SAW AND CCD PARAMETERS

	SAW	CCD
STORAGE TIME	100 μ SEC	1 SEC
BANDWIDTH	600 MHz	10 MHz
TIME-BANDWIDTH PRODUCT	10^3	100-400 STAGES
CENTER FREQUENCY	10-1500 MHz	0
DYNAMIC RANGE	100 DB	70 DB

FIGURE 4. TYPICAL TRANSISTOR CAPABILITIES

FREQUENCY (GHZ)	POWER (WATTS)	COMPANY
4-5	14	BELL LABS
4-5	15	FUJITSU
8-9	2	FUJITSU
8-9	5	TI
16	1	TI

FIGURE 5. TYPICAL TRANSISTOR MODULE POWER

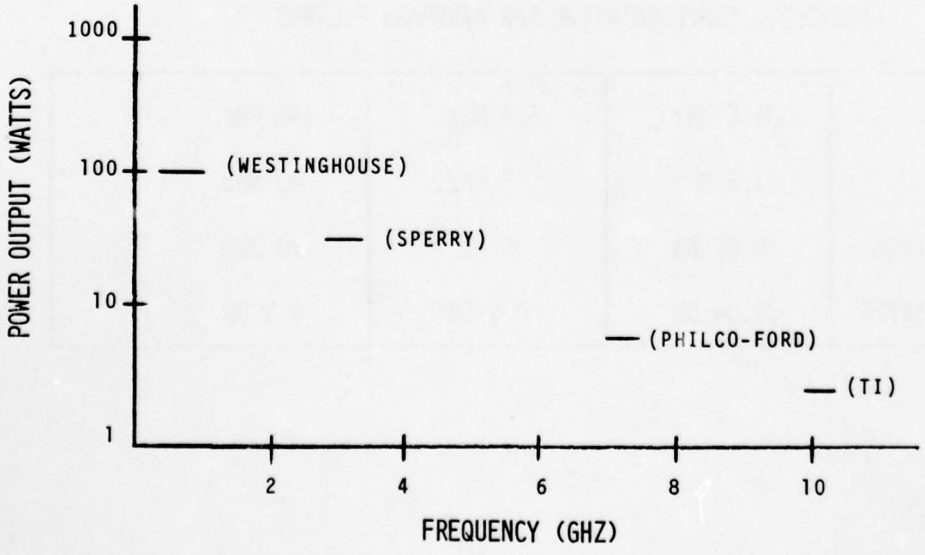


FIGURE 6. NOISE PERFORMANCE OF SOME FRONT-END AMPLIFIERS

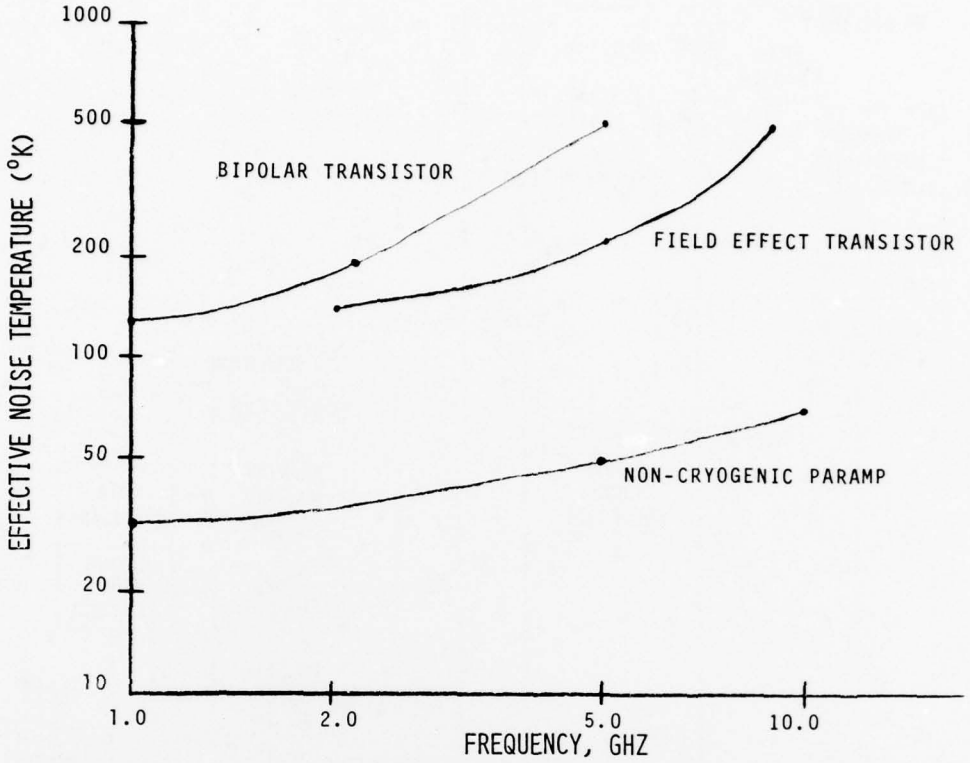


FIGURE 7. 4 x 4 C-BAND DIGITAL SCANNED ARRAY

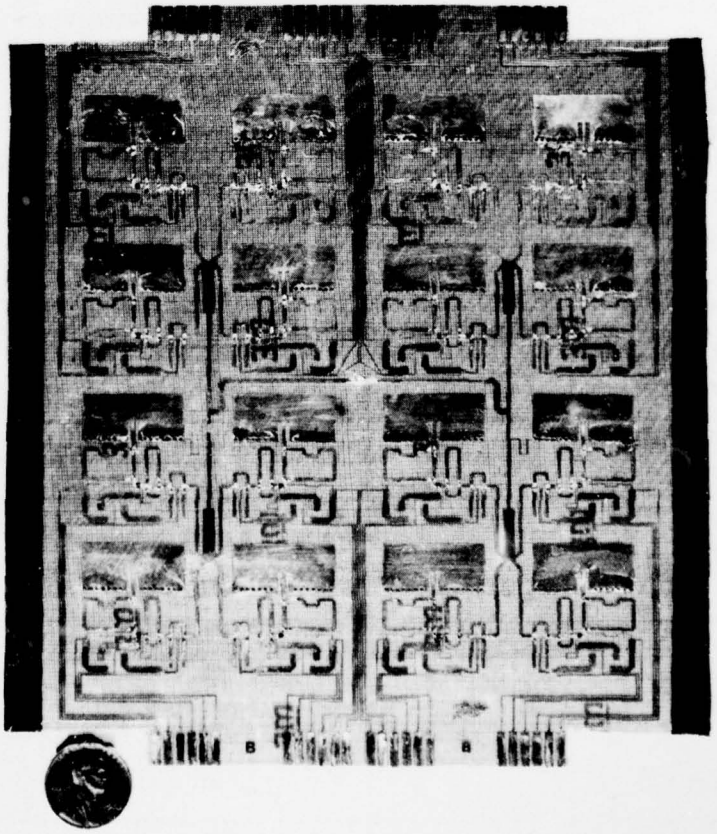


Fig. 8 TRANSCEIVER SIMPLIFIED BLOCK DIAGRAM

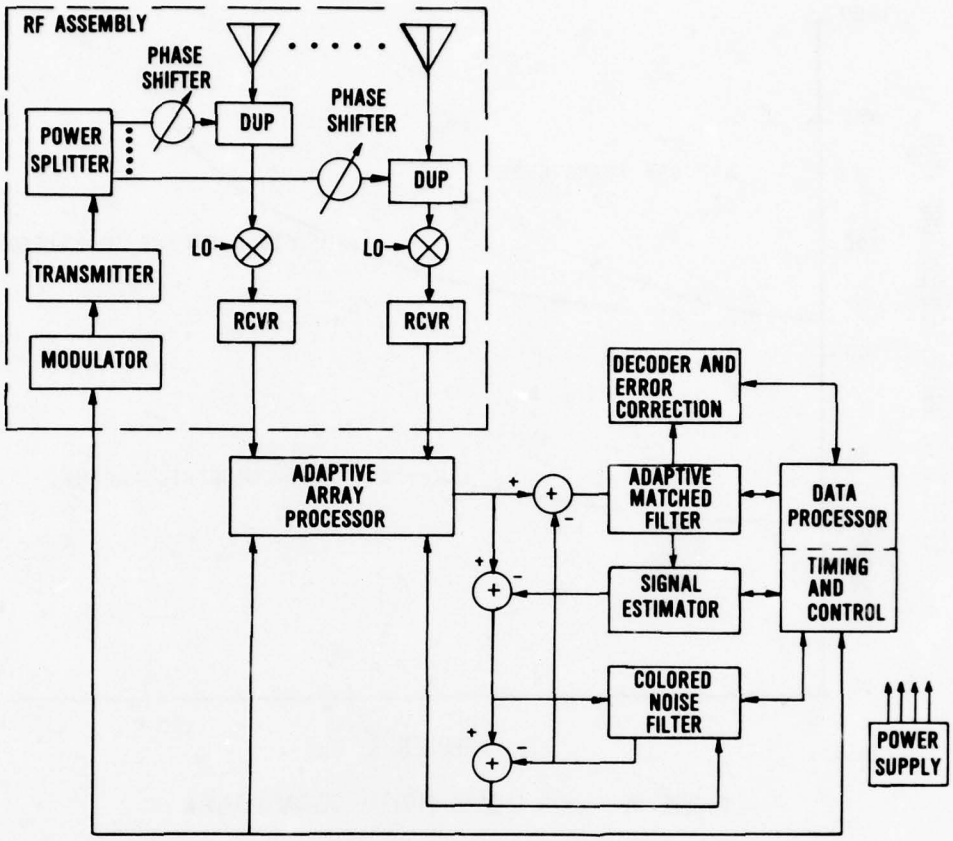


FIGURE 9. TECHNOLOGY FOR ADAPTIVE TRANSCEIVER

SUB-ASSEMBLY	DEVICES
ADAPTIVE ARRAY	
ANTENNA ELEMENTS	MICROSTRIP
PHASE SHIFTERS (RF)	PIN DIODES MICROSTRIP
ADAPTIVE PROCESSOR	DIGITAL-CMOS/SOS ANALOG-HMIC/SAW
ADAPTIVE MATCHED FILTER	
CORRELATOR	SAW, CCD
ACCUMULATOR	CCD
RECURSIVE FILTER	SAW, CCD
FREQUENCY SYNTHESIS	DIGITAL-HMIC/CMOS/SOS analog-HMIC/SAW
SIGNAL ESTIMATOR	
CORRELATOR	CCD, CMOS/SOS
LOOP	ANALOG, CMOS/SOS
COLORED NOISE FILTER	
DELAY LINE	SAW
MULTIPLIER	ANALOG @ IF
FILTER	ANALOG @ BASEBAND
TIMING AND CONTROL	MICROPROCESSOR
ERROR CONTROL	DIGITAL - SOS

DISCUSSION

B.J.Darby, UK

I should like to add two comments on applications of SAW devices to what you have already said.

Firstly I would point to the implementation of a unit for direct sequence spread spectrum acquisition using the SAW non linear convolver. This can offer approximately a valuable 1000 times increase in search rate for chip rates > 10 M chip/sec.

Secondly, you mentioned the technique of mixing two chirp signals for frequency synthesis. With SAW technology wideband fast frequency hop signals can be generated at IF. For example current state of art will allow frequency hopping over 100 MHz bandwidth into ~ 1000 contiguous frequency slots. This would correspond to a hop rate of 100 khop/sec – even at this high rate the setting time (~ 10 nsec) is a small fraction of the hop dwell time.

Both techniques have been described in recent Ultrasonics Symposium Proceedings ('75 –'77).

TRANSFORM DOMAIN PROCESSING FOR DIGITAL COMMUNICATION SYSTEMS USING SURFACE ACOUSTIC WAVE DEVICES*

L. B. Milstein
Department of Applied Physics and Information Science
University of California, San Diego
La Jolla, California

D. R. Arsenault and P. Das
Electrical and Systems Engineering Department
Rensselaer Polytechnic Institute
Troy, New York

SUMMARY

The use of surface acoustic wave (SAW) devices in communication systems has recently been receiving a reasonable amount of attention. This is because SAW devices can be used to perform accurate real-time convolutions of broadband waveforms, thus enabling them to function efficiently as matched filters, Fourier transformers, etc. In particular, they appear to have a tremendous potential in spread spectrum systems.

In the past, these devices have been shown to be capable of narrowband interference removal by Fourier transforming the received signal, passing the resulting waveform through either a notch-filter or a hard limiter, and then inverse transforming the latter signal.

In this paper, a variety of extensions of previous results will be demonstrated. One of the key limitations of processing signals in the Fourier transform domain with SAW devices is the fact that only finite segments in time of the input waveform can be transformed by the device. It will be shown that by separating the data into two parallel bit streams, this problem can be avoided.

To illustrate the interference rejection properties of the device when used as a real-time Fourier transformer, experimental results will be presented illustrating the narrowband interference rejection referred to above. Probability of error curves for a system employing a 7 bit Barker encoded binary PSK waveform embedded in additive Gaussian noise and operating both with and without the presence of a narrowband jammer will be presented. However, because of equipment limitations these latter measurements were not made with contiguous data.

The relevance of this type of technology to avionics is relatively clear. These devices are light enough and small enough to be used on board aircraft, and the ability to receive contiguous-time digital signals accurately and securely is certainly of prime importance in avionics.

1. INTRODUCTION

This paper is concerned with the detection of digital signals in the presence of additive Gaussian noise and interference. Classically, analog filtering techniques are performed by convolving the signal to be filtered by the impulse response of the filter. Recently, a new approach to analog filtering (Milstein, L.B., 1977) has been suggested, one that relies on the ability of some device, in this case, a surface acoustic wave (SAW) device, to perform a real-time Fourier transformation (and/or Fourier inversion), thereby enabling one to filter in the "frequency domain" by multiplication of appropriate Fourier transforms rather than in the "time domain" by convolution. This filtering in the frequency domain allows one the flexibility of employing filters which could not be implemented in the time domain (i.e., are unrealizable). In particular, receivers using ideal bandpass filtering and ideal notch filtering have been investigated (Das, P., 1977).

In this paper there will be a review of SAW devices in the next section, followed by a survey of some of the techniques of implementing transformations with SAW devices. The actual receiver under consideration will be presented in Section 4 and, in Section 5, experimental results will be presented showing how the receiver is capable of suppressing narrowband interference (modeled as a sine wave). Finally, Section 6 will summarize the results that have been achieved to date and indicate the directions that this technology appears to be taking in the future.

2. REVIEW OF SURFACE ACOUSTIC WAVE TECHNOLOGY

Surface acoustic waves have been well known and well studied by the seismologists since Lord Rayleigh's discovery of this mode of wave propagation in 1895. Only in the last two decades (Ultrasonics Symposium Proceedings, 1972-1976), however, has the importance of SAW in the electronic industry been realized. This is due to two main factors. First is the availability of piezoelectric substrates like lithium niobate ($LiNbO_3$), and lead zirconium titanate (PZT), and second is the easy generation and reception of SAW on a piezoelectric substrate using interdigital transducers. Thus, one can easily make a delay line with an insertion loss of a few dBs, tens of microseconds of delay, and a center frequency which varies from 10 MHz to a few gigahertz. It is to be mentioned that one can also make a delay line using bulk ultrasound (i.e., using a device wherein the wave travels through the entire volume rather than just near the surface), but there are two very important reasons why a SAW delay line is more attractive. First of all, the SAW can be very easily tapped on a piezoelectric substrate by one interdigital transducer (or a set of transducers) to make a tapped delay line. In addition, one can put independent tapping weights on the pick-off points. This makes the realization of transversal or finite impulse response (FIR) filters with prespecified characteristics (within certain limitations) very simple. Thus, using the well-known techniques of digital filter design, one can design a single mask which, employing the usual process of photolithography (well developed by the integrated circuits industry), can be used to manufacture these filters with significant reduction in cost.

SAW technology includes not only Rayleigh waves, but all the waves which can propagate on a solid surface somewhat confined near the surface. For Rayleigh waves the confinement is of the order of one wavelength (for $LiNbO_3$ at 100 MHz, $\lambda \sim 30\ \mu$), but for other waves, like Blustein-Gulayev waves, the confinement length may be much larger, say $\epsilon_r \lambda$, where ϵ_r is the effective relative dielectric constant which, for example, is approximately 30 for $LiNbO_3$. Rayleigh waves on a perfect surface of piezoelectric insulator are non-dispersive and non-dissipative, but in actual solids the dispersion and loss characteristics become of importance in the gigahertz region. For frequencies less than 500 MHz, of much more importance is the alignment of the crystallographic axes with the direction of wave propagation. This is because in an anisotropic solid, the directions of the energy and the wave vectors are collinear only in certain specified directions and for large misalignment of the vectors, significant amounts of energy travel at an angle away from the desired direction. Also, for high frequencies, diffraction loss may become significant.

An oscillator can be made using a SAW delay line and an external amplifier. The Q of this delay line oscillator is generally of the order of a few thousand. For higher Q oscillators, one uses so-called SAW resonators. The resonators are made of one or two interdigital transducers with many metal fingers or grooves to be used as reflective arrays for the planar cavity. With proper care, oscillator Qs of the order of 60,000 have been reported (Bell, D.T., 1976).

The basic advantage for the SAW devices discussed above is that they use planar technology and thus can be cheaply mass-produced. For the tapped delay line applications, one major handicap is the rather large temperature coefficient of $LiNbO_3$ or other high coupling material which has low insertion loss. If one is willing to sacrifice the efficiency, ST-cut quartz is available which has zero first order temperature coefficient. Finally, there is another class of SAW devices which use acousto-electric interaction of SAW and free carriers on a semiconductor surface in a sandwich structure of semiconductor on piezoelectric substrate (i.e., silicon on $LiNbO_3$). This is discussed below.

SAW propagating on a piezoelectric substrate (i.e., delay line) interacts with carriers in a neighboring semiconductor. This interaction takes place even though the piezoelectric and semiconducting media have their surface mechanically isolated by an air gap. The acousto-electric or space charge coupling is achieved through the electric field which accompanies the surface wave. This wave exists outside the piezoelectric substrate and can penetrate inside the semiconductor and thus induce space-charge.

Acoustic surface wave convolvers are real-time analog ultrasound signal processors using this interaction. A silicon-on-lithium niobate ($LiNbO_3$) structure (a so-called separate media structure) is the implementation most used. Devices of this type include (in addition to convolvers) correlators, match filters, Fourier transformers, ambiguity function generators, and phase comparators (Bers, A., 1974)(Ingebrigtsen, K., 1975)(Defranould, Ph., 1976)(Das, P., 1977)(Das, P., 1972)(Wang, W., 1972)(Otto, O., 1972)(Kino, G., 1976).

To illustrate the operation of this device, consider the situation in Fig. 1, where an RF signal $f(t)e^{j\omega t}$ is applied to one input transducer to generate a traveling wave. At the other end of the device, the input applied is $g(t)e^{j\omega t}$. These two waves, while traveling under the semiconductor, induce a propagating electric field and a space charge which can be represented at any point x and t inside the medium (to within a multiplicative factor) by

$$f(t - \tfrac{x}{v})e^{j(\omega t - kx)} \text{ and } g(t + \tfrac{x}{v})e^{j(\omega t + kx)}$$

where k is the propagation constant of the wave and v is its velocity. In overlapping, these waves interact and the current density inside the semiconductor consists of the fundamental and higher order harmonics. The output is proportional to the integral of the current density with respect to x. Thus, if only the second harmonic term is detected, the output is proportional to

$$\int_{-L/2}^{L/2} f(t - \tfrac{x}{v})\, g(t + \tfrac{x}{v})\, dx = v \int_{t-L/2V}^{t+L/2V} f(\tau)\, g(2t - \tau)\, d\tau \qquad (1)$$

where L is the physical length of the device. For time-limited signals, with $T = L/2V$, the above expression represents convolution, except for an output time compression factor of two; thus Fig. 1 depicts a convolver.

An important feature of these real-time analog signal processors is that they are programmable in the sense that one convolver can be used for many types of signals. This certainly is a great advantage over the tapped delay line correlators discussed earlier, as the latter are capable of responding to a fixed signal form only.

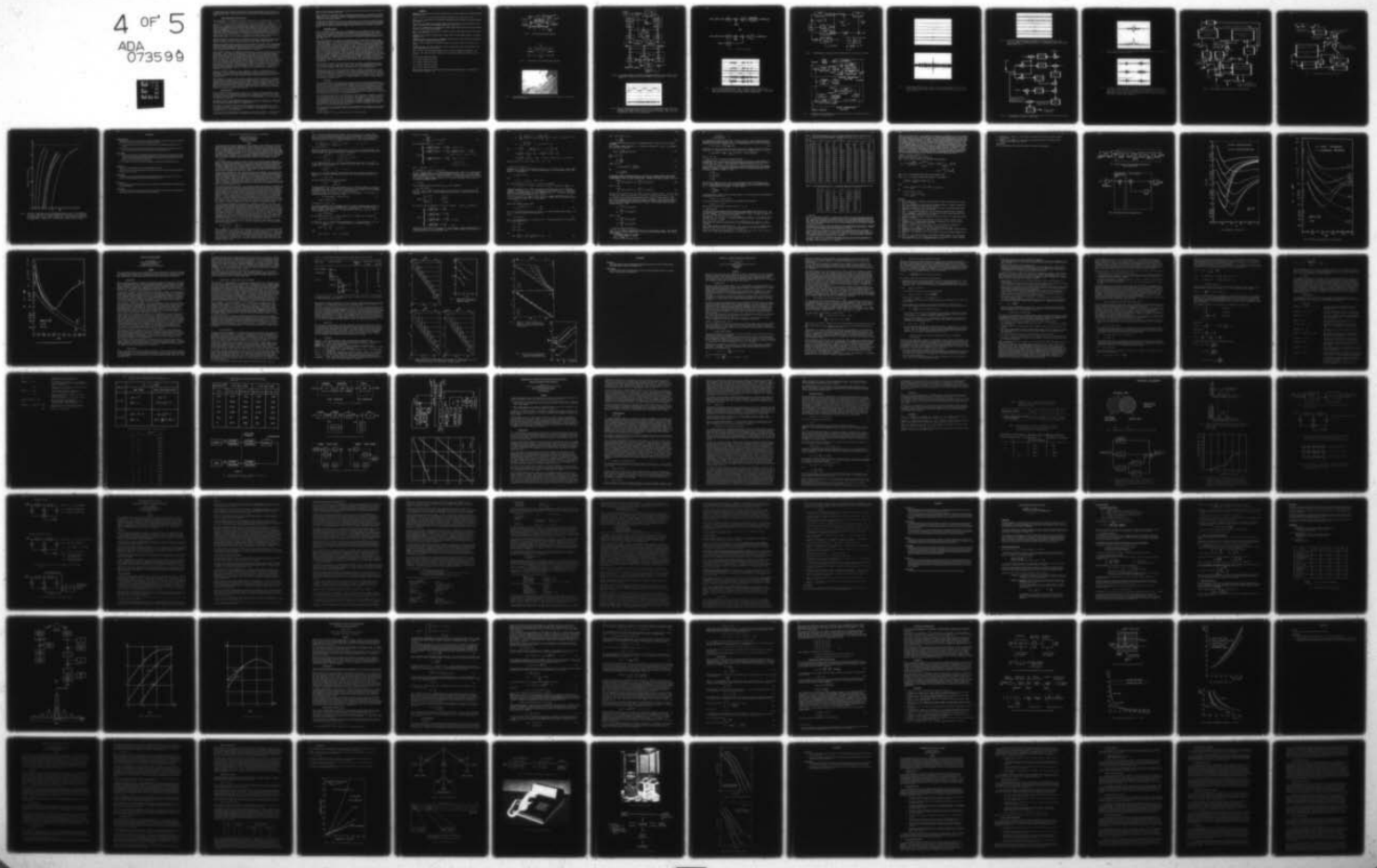
Finally, one other class of space-charge-coupled devices has been developed recently, the so-called memory or storage devices (Bers, A., 1974)(Ingebrigtsen, K., 1975)(Defranould, Ph., 1976)(Das, P., 1977), which can perform the signal processing functions mentioned earlier with a stored signal in the charge pattern of a vidicon diode array placed on a $LiNbO_3$ delay line. The storage device is shown in Fig. 2. To store the signal f(t), it is applied at input 1 as an envelope modulating a carrier at frequency ω. Another short input pulse (sometimes referred to as a "write-in" pulse) at frequency ω is applied either at input 2 (case 1) or at the output terminal (case 2) such that the waves generated by the signal and the write-in pulse interact to produce a DC current. This DC current produces a trapped periodic charge density due to the charging of capacitors associated with the diode arrays. The period of this charge is $(1/2k)$ for case 1 and $(1/k)$ for case 2.

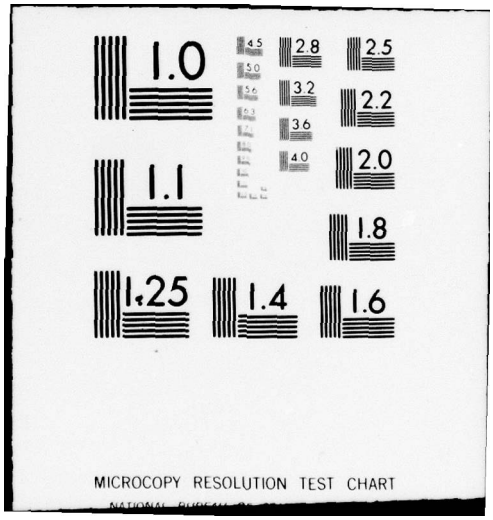
The charge pattern may remain stored from seconds to ten minutes depending on the type of diode structure and the ambient temperature. To read out the stored signal, a read-in short pulse at frequency 2ω (case 1) is applied at either terminal 3 or 4. Alternately a short pulse at frequency ω (case 2) can be applied to either terminal 1 or 2. The stored signal is recovered at the output terminal. To convolve a signal g(t) with the stored signal, the short read-in pulse is replaced by g(t) and applied at either input 4 or input 2. Alternately, to correlate, g(t) is applied at either 3 or 1.

AD-A073 599 ADVISORY GROUP FOR AEROSPACE RESEARCH AND DEVELOPMENT--ETC F/G 17/2
DIGITAL COMMUNICATIONS IN AVIONICS.(U)
JUN 79 H LUEG

UNCLASSIFIED AGARD-CP-239 NL

4 OF 5
ADA 073590




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
MICROCOPY RESOLUTION TEST CHART

In summary, research in this area has brought real-time signal processors from the drawing board to the actual application stage. Figure 3 shows a complete unit which has been fabricated and tested with the following specifications: 100 MHz convolver; 10 μsec interaction time; 25 MHz bandwidth; insertion loss 40 dB.

3. FOURIER TRANSFORMATION USING SAW DEVICES

Fourier transformation is accomplished in real-time by a SAW device in the following manner: If a signal $f(t)e^{j(\omega t+\Delta t^2)}$ (that is, a waveform f(t) modulating a linear FM or chirp waveform) is convolved with the signal $e^{j(\omega t-\Delta t^2)}$, the result of that convolution will be the Fourier transform of f(t)(Milstein, L., 1977) (Das, P., 1977). Therefore, if these two waveforms are used as the two inputs to the SAW convolver described in the previous section, the convolver output, assuming f(t) is time-limited to some value $T \leq A$, where A is the interaction time of the device (and equals the physical length of the device divided by the velocity of propagation of the SAW in the device), will be $F(\omega)$, the Fourier transform of f(t), over the range $\omega\epsilon[2\Delta T, 2\Delta A]$(Milstein, L., 1977). Alternately, rather than use the convolver, if one implements a tapped delay line with tap coefficients given by samples of the unmodulated chirp signal spaced (π/ω_c) seconds apart, where ω_c is the bandwidth of the chirp waveform, one can obtain $F(\omega)$ as the output of the delay line when $f(t)e^{j(\omega t+\Delta t^2)}$ is the input. In either case, the radian frequency variable ω will be a linear function of time, so that $F(\omega)$ will be generated at the device output in real-time.

The above technique is sufficient only if the waveform is time limited to a small enough value. When the waveform is a sequence of contiguous pulses as is typical of most digital communication systems, some means of altering the procedure is clearly necessary, and one such scheme is described below.

Basically, the technique consists of dividing the input alternately into two data streams, processing each data stream separately, and then combining the results at the end. Conceptually, it is a straightforward technique, but from an implementation point of view there were a variety of subtleties involving such things as the accurate generation of constant envelope chirps in the two parallel branches and minimization of crosstalk effects in inverse transforming that had to be addressed.

Figure 4(a) shows a detailed block diagram of the system and Fig. 4(b) shows the signals at different points in the system. A chirp clock plus a delayed version of that clock controls the system as follows: The delayed clock is used to trigger a flip-flop which in turn triggers two out-of-phase impulse generators. These impulses modulate an RF carrier and then generate two out-of-phase chirp streams after passing through the two chirp filters. A second flip-flop is triggered by the non-delayed clock and is used to gate the two chirp streams so that overlapping does not occur. The streams are changed to the opposite slope by mixing with 2ω. The signal s(t) is then made to modulate these streams. The modulated chirps are fed into two chirp filters from which emerges two out-of-phase Fourier transform streams. Mixing again with 2ω to obtain the opposite slope, summing the two streams together, and feeding this signal into another chirp filter gives the original continuous signal at the output. The out-of-phase chirp streams are then summed together and mixed with the recovered signal to eliminate the chirp carrier. Synchronization is obtained by slightly changing the chirp stream frequency, which is independent of the signal frequency. The signal can be monitored by triggering on the signal clock, whereas the chirps are monitored by triggering on the chirp clock.

Finally, as described in (Milstein, L., 1977), (Arsenault, D., 1977), two further points are worth emphasizing. The first is obvious, merely being that since one obtains a transform valid only over a finite range in frequency, one can only inverse transform over that range in frequency so that in general, one obtains at the output of an inverse transform the desired inverse convolved with a sin x/x type weighting function.

The second point is that since in either the forward transform or inverse transform case, the transformation only appears at the output of the device when the input waveform is fully contained in the device, if the nominal carrier frequency of both the upward and downward chirps are the same, the values f(0)(and immediate vicinity) and F(0)(and immediate vicinity) cannot be obtained. Therefore, appropriate time and frequency shifting procedures have to be employed (Milstein, L., 1977) and (Arsenault, D., 1977). Alternately, if F(0) is desired, it can be obtained by using different carrier frequencies for the opposite going chirps. However, this then puts the entire range of frequencies for which an accurate transform is obtained in the vicinity of baseband rather than at RF.

4. RECEIVER STRUCTURE

The general form of the receiver is shown in Fig. 5(a). It consists of a Fourier transformer, a multiplier, an inverse Fourier transformer, and a matched filter. In essence, the filtering by the transform function $H(\omega)$ is done by multiplication followed by inverse transformation rather than by convolution. This multiplication, while ostensibly being performed in the "frequency domain", is of course, accomplished by the SAW device in real-time.

Alternately, the receiver may be implemented as shown in Fig. 5(b) (Das, P., 1977)(Otto, O., 1976), wherein the matched filtering is performed by inverse transforming the product of the transforms of the filtered input waveform and the impulse response of the matched filter.

To illustrate the above ideas, Fig. 6 shows results of narrowband interference removal when s(t) is a 13-bit Barker code sequence. (Actually, the code was composed of ONEs and ZEROs rather than ± ONES.) The interference in this case was a sine wave, and it was filtered out by multiplication in frequency by a rectangular pulse (i.e., a low-pass filter).

It can be seen from the figure that the interference has been effectively eliminated. The distortion seen in the final trace is due in large part to the bandwidth of the final video filter. As an incidental

result, if traces 1 and 3 are compared with each other (also traces 4 and 6) one can see the fidelity with which the Fourier transforms can be taken.

Figure 7 shows the actual implementation used to generate the above results, and corresponds to the block diagram of Fig. 5(a). The Fourier transforms were implemented using SAW delay lines with a chirp impulse response built into the tap weights. However, the final matched filtering operation was performed using a silicon-on-lithium niobate convolver.

To demonstrate the feasibility of Fig. 5(b), the receiver shown in Fig. 8 was built and results are shown in Figs. 9 and 10. Figure 9 shows the output of a filter matched to a 255-bit PN code (again implemented with ONEs and ZEROs) when the input was that same code under interference-free conditions. When a narrow-band interferer (specifically a periodic triangular waveform) was added, the filter output is shown in Fig. 10.

5. EXPERIMENTAL RESULTS

The receiver structure shown in Fig. 5(b) can be implemented using less components as shown in Fig. 11 (Otto, O. W., 1976). The difference between this implementation and that of Fig. 8 is that no time-reversed signal is required for correlation. Also it is to be noted that the output of the receiver is dechirped. This is shown in Fig. 12 where both the correlation and dechirped correlation of 7-bit Barker code signals are shown. The performance of this receiver in the presence of high level jamming is also shown in Fig. 13.

To test the performance of this receiver, the probability of error curve was measured using the block diagram shown in Fig. 14. The signal used in the error-analysis was a $(2^{24} - 1)$ bit PN code generated by a 24-bit shift register. Each +1 bit of the signal was encoded with a 7-bit Barker code having dechirped correlation peak shown in Fig. 12, whereas for the 0-bit, a 180° phase shifter was used to obtain a negative correlation spike. This was done by inverting the output of the 7-bit Barker code generator with every +1 bit of the PN code generator and leaving it unchanged with a zero bit. The other output of the 7-bit code generator was used in the reference channel. The correlation output was applied to the threshold detecting and error-counting circuit shown in Fig. 15. The threshold level was set to zero and the output of the zero level detector was compared with the input signal. If there was an error it was counted in an 8-segment decade counter. The clock frequency for the signal was 2 KHz and the 7-bit Barker code was 23 μsec long (unfortunately, equipment limitations prevented these measurements from being performed with contiguous time data). The center frequency and bandwidth of the chirp filters were 15 MHz and 6 MHz respectively. The jamming noise was generated by using a sinusoidal oscillator. The RMS noise voltage was measured by a Dumont type 405 high frequency RMS voltmeter. This voltmeter was found to have high frequency resonances and to eliminate this, a 6 MHz low pass filter was inserted at the output of the noise generator. Since the system bandwidth was also 6 MHz, this still could be looked upon as more or less white noise.

Figure 16 shows the probability of error curves obtained for the system. Curve B was obtained using 0.145 volt RMS noise and 0.2 V peak to peak signal. Curves C and D show the degradation of the receiver in the presence of different jammer levels. Curves E and F show the improvement obtained by selective gating in the Fourier domain. Curve A shows the probability of error for an optimum receiver.

Comparing curves A and B, one finds that the present receiver is inferior to the optimum one by 3.5 dB. The antijamming capability is 0.5 dB and 2.5 dB for the low and high level jammers respectively. These results are preliminary and better performance can be expected by optimizing the system. For example, it is expected that curves E and F should be much closer to each other than shown in the figure. The reason this was not so in the present system was due to an improper gating of the jamming signal in the receiver. This resulted in spreading of the jammer in the frequency domain and thus could not be removed completely without degrading the signal itself.

6. DISCUSSION

Different implementations of a spread spectrum receiver using SAW devices as Fourier transformers have been discussed. For a particular implementation a probability of error curve was obtained. These are only preliminary results and no attempt has been made to compare them with theoretical predictions. The results presented in this paper are characteristic of the behavior of SAW devices as signal processors showing their superiority in situations that other devices, at present, might find troublesome to contend with. The fact that these devices allow access to the real-time Fourier transform of the signal as an automatic consequence of the correlation process (as has been demonstrated in this paper) allows one to employ such powerful techniques as filtering by transform gating and noise optimization or 'prewhitening' by multiplying the transform by a function related to the noise characteristics.

One very important advantage in the use of SAW devices for signal processing is the possibility of fabricating entire receivers, and the like, on a single substrate. For instance, all the chirp filters depicted in the correlation system of Fig. 11 can be fabricated as a single monolithic unit. From this one can envision dramatic decreases in the bulk of such systems. At the moment it seems plausible to state that the dynamic range of such systems, as have been discussed in this paper, is ultimately limited by other system components such as mixers since SAW devices are known to possess wide dynamic ranges. Although SAW devices may require high level inputs this does not present too much of a problem due to the present availability of excellent wideband high-gain amplifiers.

Finally, it should be mentioned that although the Fourier transform has been stressed solely in this paper, other transforms are also implementable using SAW devices (Arsenault, D., 1977), opening up new avenues for application.

*Partially supported by U. S. Army Research Office Grant No. DAAG-29-77-G-0205.

REFERENCES

ARSENAULT, D.R., P. Das, 1977, "SAW Fresnel Transform Devices and Their Applications", Ultrasonics Symposium Proceedings, p. 969.

BELL, D.T., and R. C. Li, 1976, "Surface Acoustic Wave Resonators", Proc. IEEE, Vol. 64, p. 711.

BERS, A., and J. H. Cafarella, 1974, "Surface Wave Correlator-Convolver with Memory", Ultrasonics Symposium Proceedings, p. 778.

DAS, P. and D. R. Arsenault, 1977, "SAW Space Charge Coupled Signal Processing and Transform Device with Storage", Extended Abstract, Electrochemical Society, Vol. 77-1, p. 112.

DAS, P., D. R. Arsenault and L. B. Milstein, 1977, "Adaptive Spread Spectrum Receiver using SAW Technology", DAAG-29-77-G-0205, Technical Report MA-ARO-1.

DAS, P., M. N. Araghi and W. C. Wang, 1972, "Convolution of Signals using Surface Wave Delay Lines", Appl. Phys. Letters, Vol. 21, p. 152.

DEFRANOULD, Ph., H. Gautier, C. Maerfeld and P. Tournois, 1976, "P-N Noise Memory Correlator", Ultrasonics Symposium, p. 336.

INGEBRIGTSEN, K. A. and E. Stern, 1975, "Holographic Storage of Acoustic Surface Waves with Schottky Diode Arrays", Ultrasonics Symposium Proceedings, p. 212.

KINO, G.S., 1976, "Acoustoelectric Interactions in Acoustic Surface Wave Devices", Proc. IEEE, Vol. 64, p. 724.

MILSTEIN, L.B. and P. Das, 1977, "Spread Spectrum Receiver using Surface Acoustic Wave Technology", IEEE Trans. Communications, Vol. COM-25, p. 841.

OTTO, O.W., 1976, "The Chirp Transform Signal Processor", Ultrasonics Symposium Proceedings, p. 365.

OTTO, O.W., 1972, "Real Time Fourier Transform with a Surface Wave Convolver", Electron Letters, Vol. 8, p. 623.

1972 Ultrasonics Symposium Proceedings.

1973 Ultrasonics Symposium Proceedings.

1974 Ultrasonics Symposium Proceedings.

1975 Ultrasonics Symposium Proceedings.

1976 Ultrasonics Symposium Proceedings.

WANG, W.C. and P. Das, 1972, "Surface Wave Convolver Via Space Charge Nonlinearity", Proceedings of the IEEE Ultrasonics Symposium, p. 316.

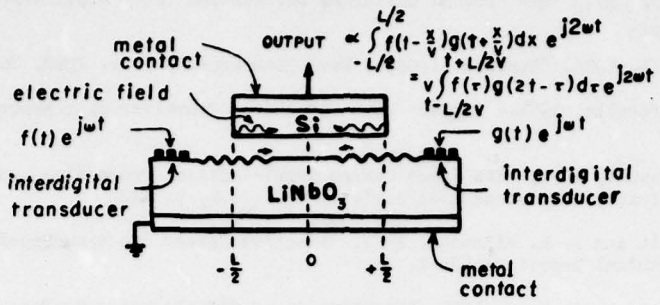


Fig. 1 The Si-on-$LiNbO_3$ Convolver Structure

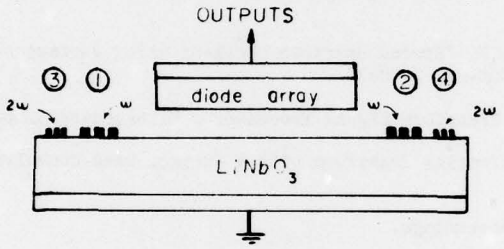


Fig. 2 The Structure of the Si-on-$LiNbO_3$ Memory Correlator



Fig. 3 External Appearance of a Si-on-$LiNbO_3$ Convolver with 100 MHz Center Frequency, 25 MHz Bandwidth and 10 μs Interaction Time

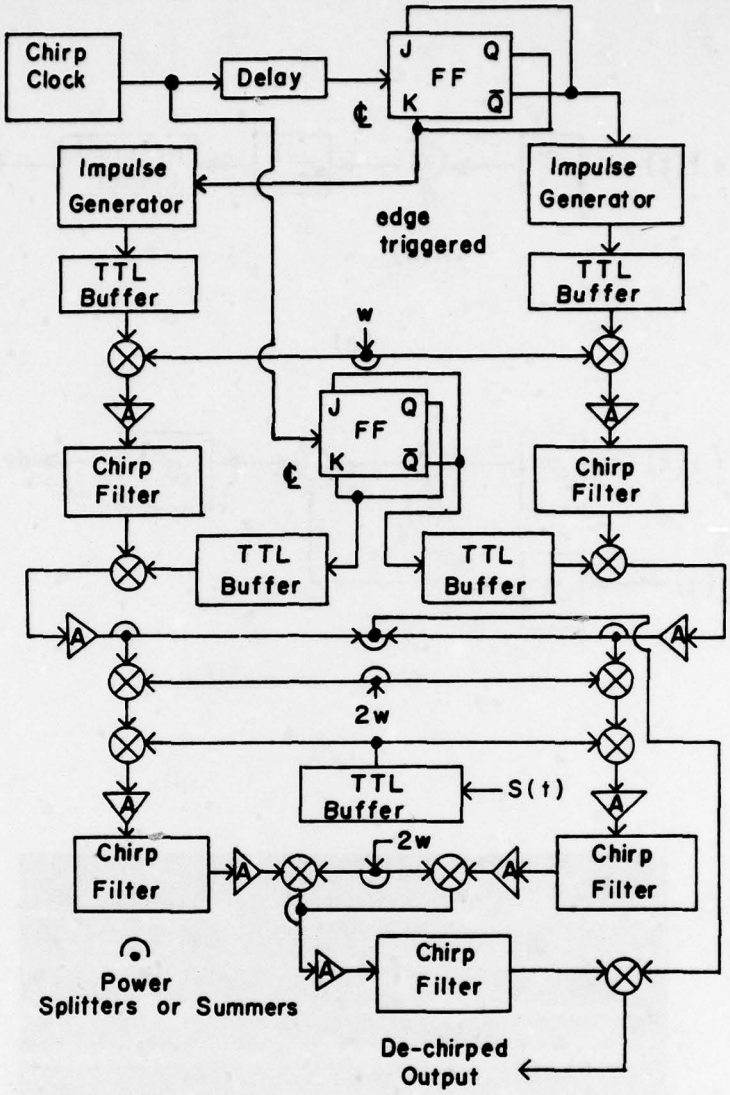


Fig. 4(a) Detailed Block Diagram of the Fourier Transformation Scheme for Continuous Signals. Gating the transform to remove unwanted frequency components is performed prior to the last chirp filter.

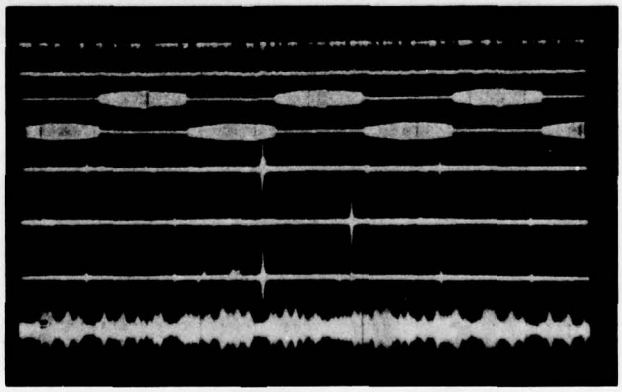


Fig. 4(b) Signals at Different Points within the Continuous Fourier Transformation System. Trace 1 is a continuous pseudo-random input where every bit is a 13 bit Barker code. Traces 2 and 3 are the alternating chirp streams. Traces 4 and 5 are the respective transforms. Trace 6 is the sum of traces 4 and 5. Trace 7 is the recovered signal.

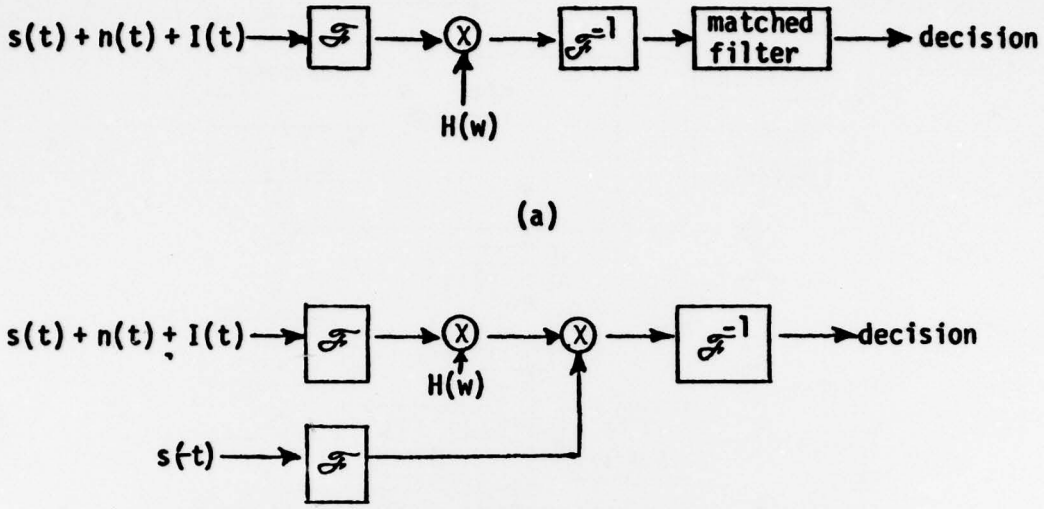


Fig. 5 Receiver Block Diagrams

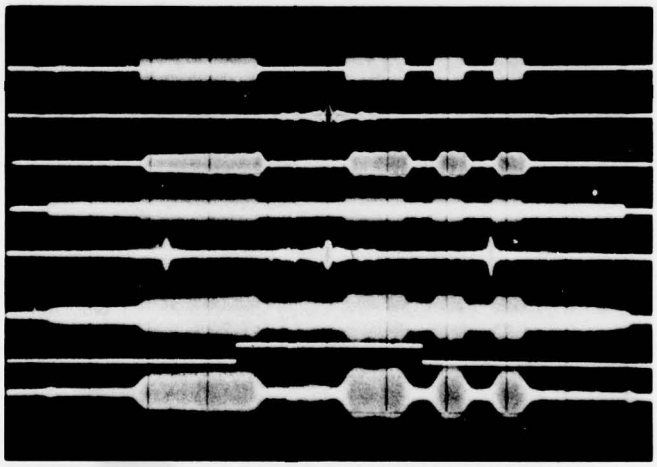


Fig. 6 Filtering of Barker Code Signal (hor. scale - 5 μsec/div.). Trace 1 - Barker Code Input; Trace 2 - Fourier Transform of Trace 1; Trace 3 - Inverse Fourier Transform of Trace 2: Trace 4 - Barker Code Plus Interference; Trace 5 - Fourier Transform of Trace 4; Trace 6 - Inverse Transform of Trace 5; Trace 7 - $H(\omega)$ - Gating Signal; Trace 8 - Filtered Signal.

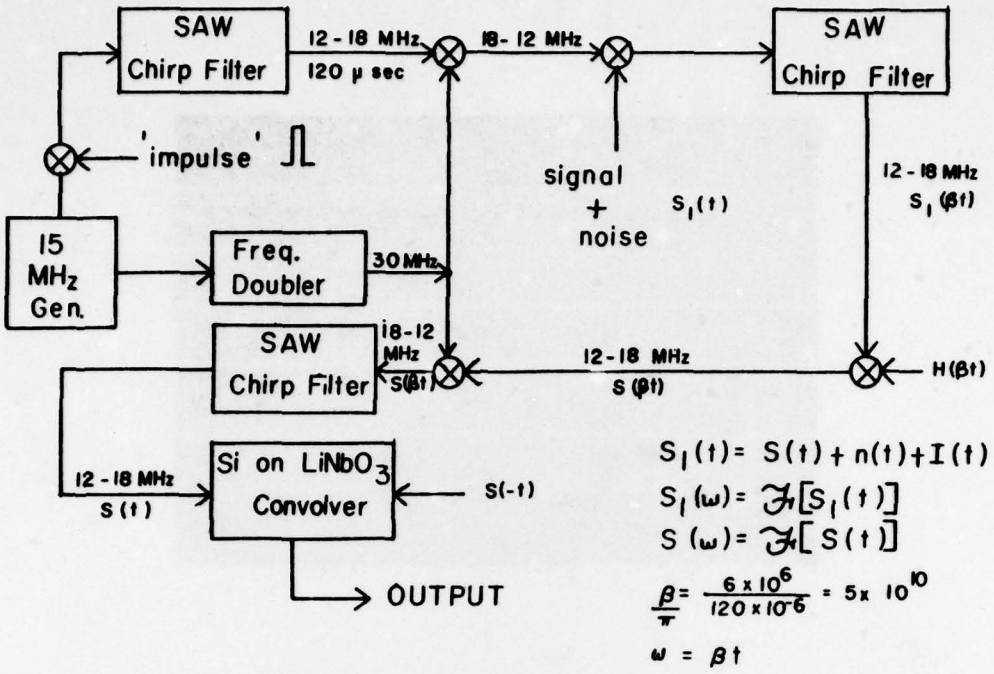


Fig. 7 Implementation of a SAW Receiver where the Matched Filter is a Si-on-$LiNbO_3$ Convolver.

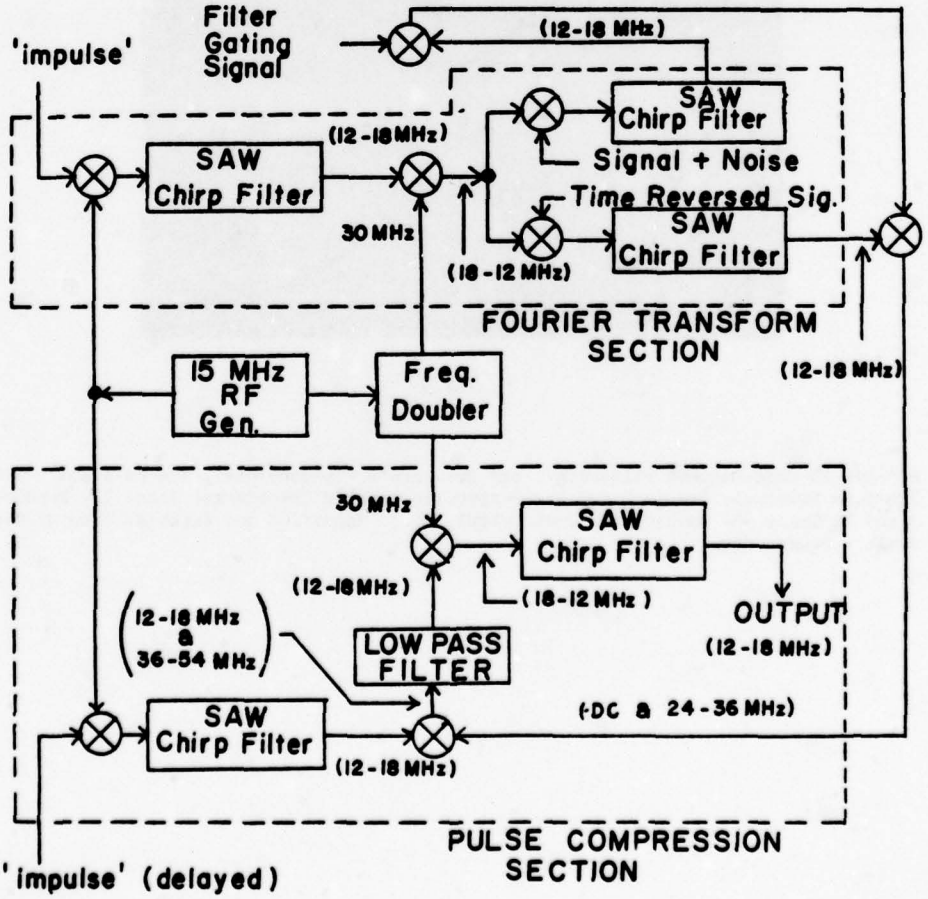
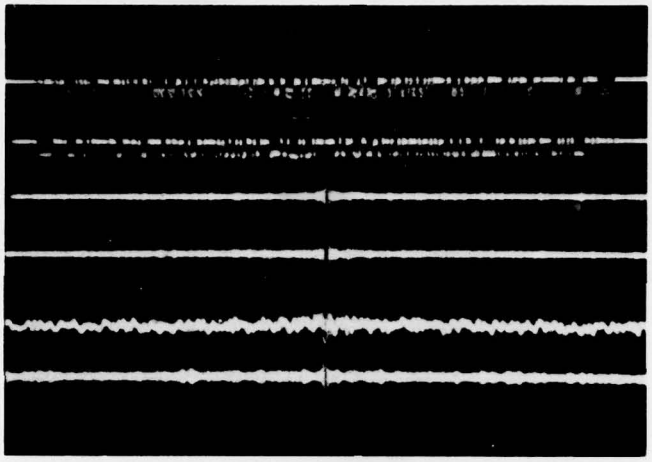
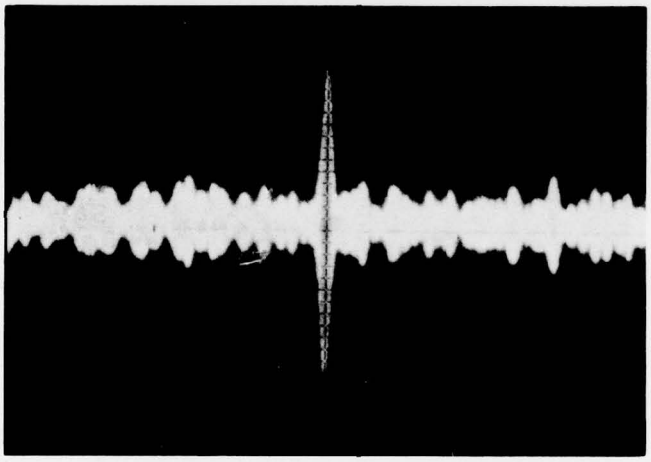


Fig. 8 Implementation of a SAW Receiver using Up-Chirp SAW Delay Lines to Obtain Matched Filtering by the Product of Transforms Technique.



(a)



(b)

Fig. 9 255-Bit PN Code-Matched Filtering. (a) Hor. scale - 5 μsec/div., Traces 1 and 2 - Code and Its Time Reversal; Traces 3 and 4 - Respective Fourier Transforms; Trace 5 - Product of Traces 3 and 4; Trace 6 - Matched-Filtered Output. (b) Magnified and Expanded View of Trace 6 (hor. scale - 2 μsec/div.

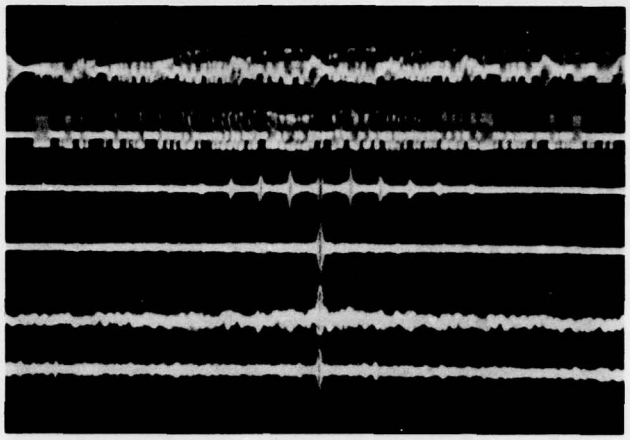


Fig. 10 255-Bit PN Code-Matched Filtering in the Presence of Triangular Interferer (hor. scale - 5 μsec/div.). Trace 1 - Signal Plus Interferer; Trace 2 - Time Reversed Signal; Traces 3 and 4 Fourier Transforms of 1 and 2 Respectively; Trace 5 - Multiplication of Traces 3 and 4; Trace 6 - Matched-Filtered Output.

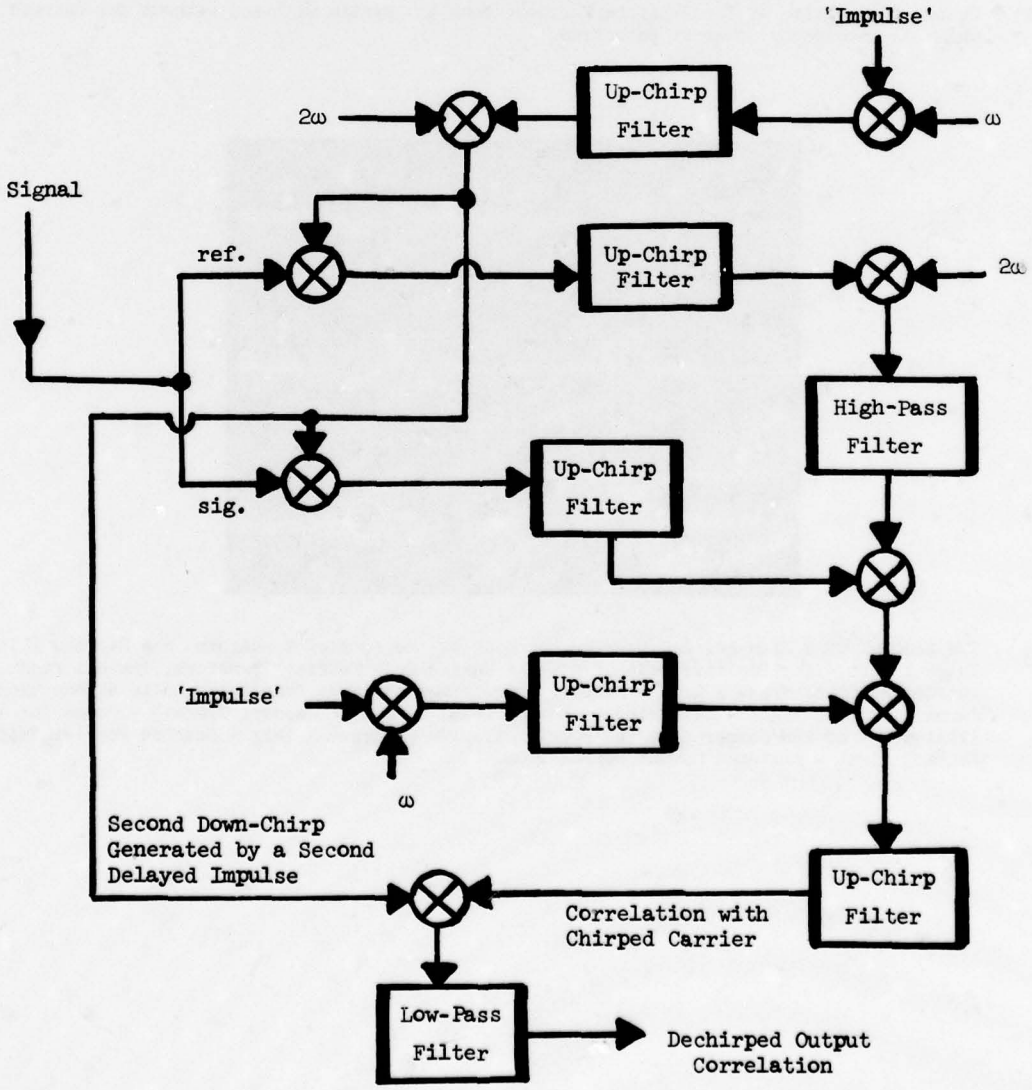


Fig. 11 Block Diagram of the Product of Transforms Matched Filter that does not Require the Generation of a Time Reversed Version of the Input Signal.

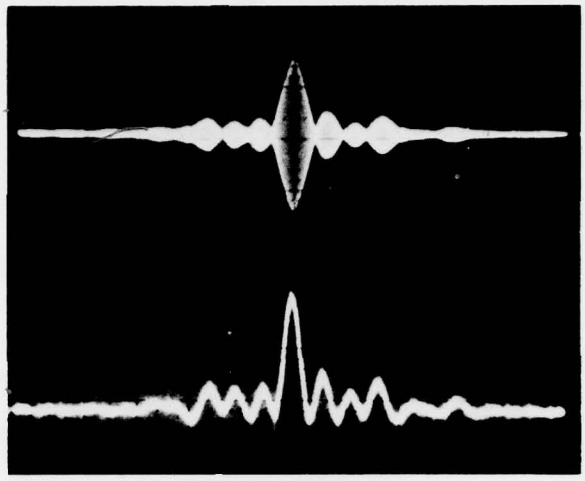


Fig. 12 Output Correlation of the 7-Bit Barker Code from the System with and without the Chirped Carrier which is Removed by Coherent Detection.

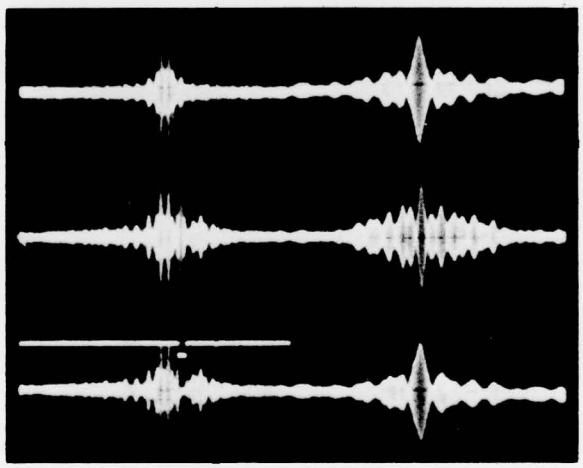


Fig. 13 The Removal of a High-Level Jammer by Notching of the Fourier Transform in a Matched Filter. Trace 1 left - The Positive Side of a 7-bit Barker Code Fourier Transform; Trace 1 right - Output Correlation; Trace 2 left - Fourier Transform with Large Component due to a Monochromatic Jammer; Trace 2 right - Distorted Output Correlation due to Jammer; Trace 3 - Notch for the Elimination of the Jammer from the Fourier Transform; Trace 3 left - Notched Fourier Transform; Trace 3 right - Filtered Output Correlation.

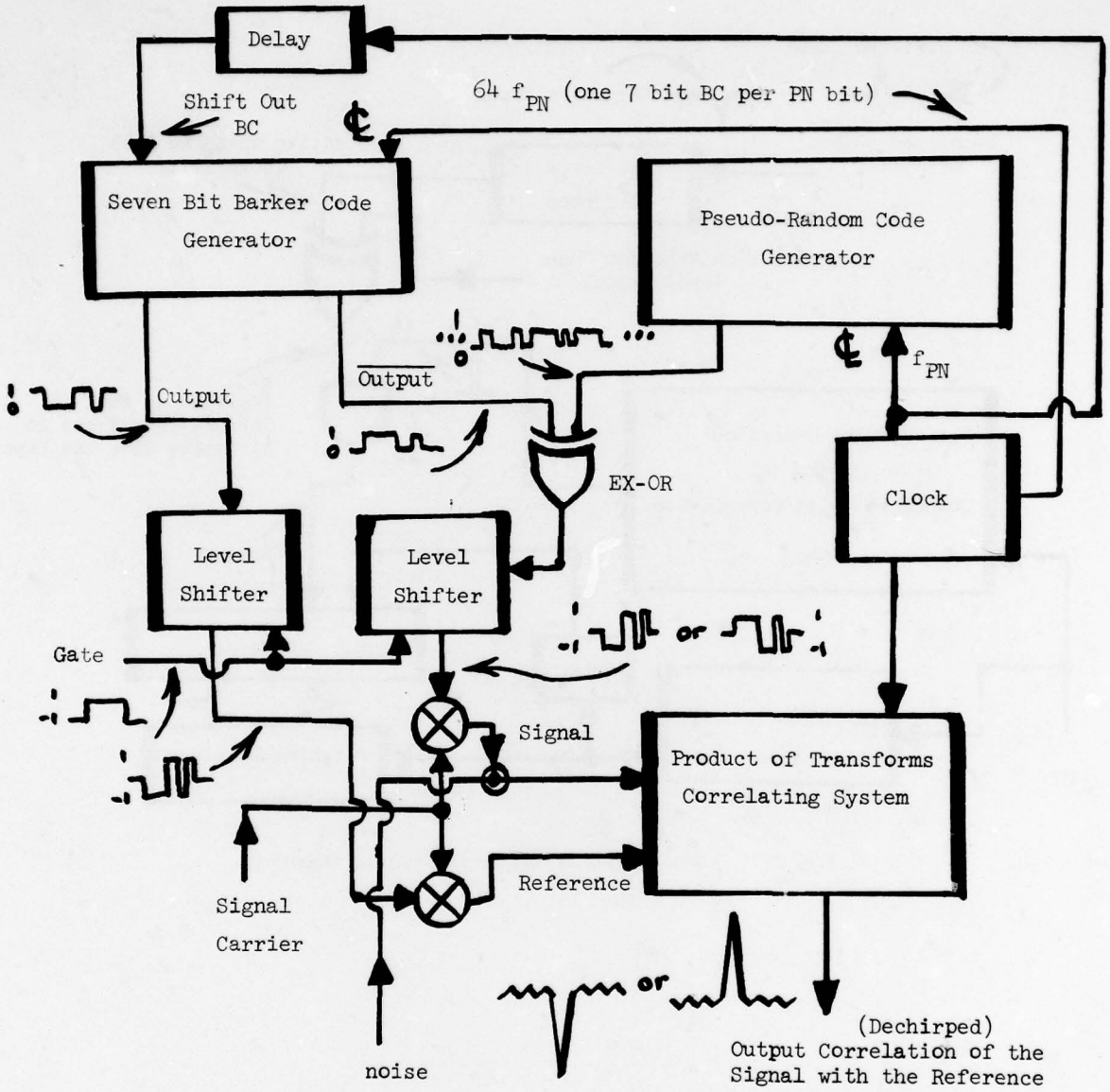


Fig. 14 Block Diagram of the Actual System used in the Error Analysis.

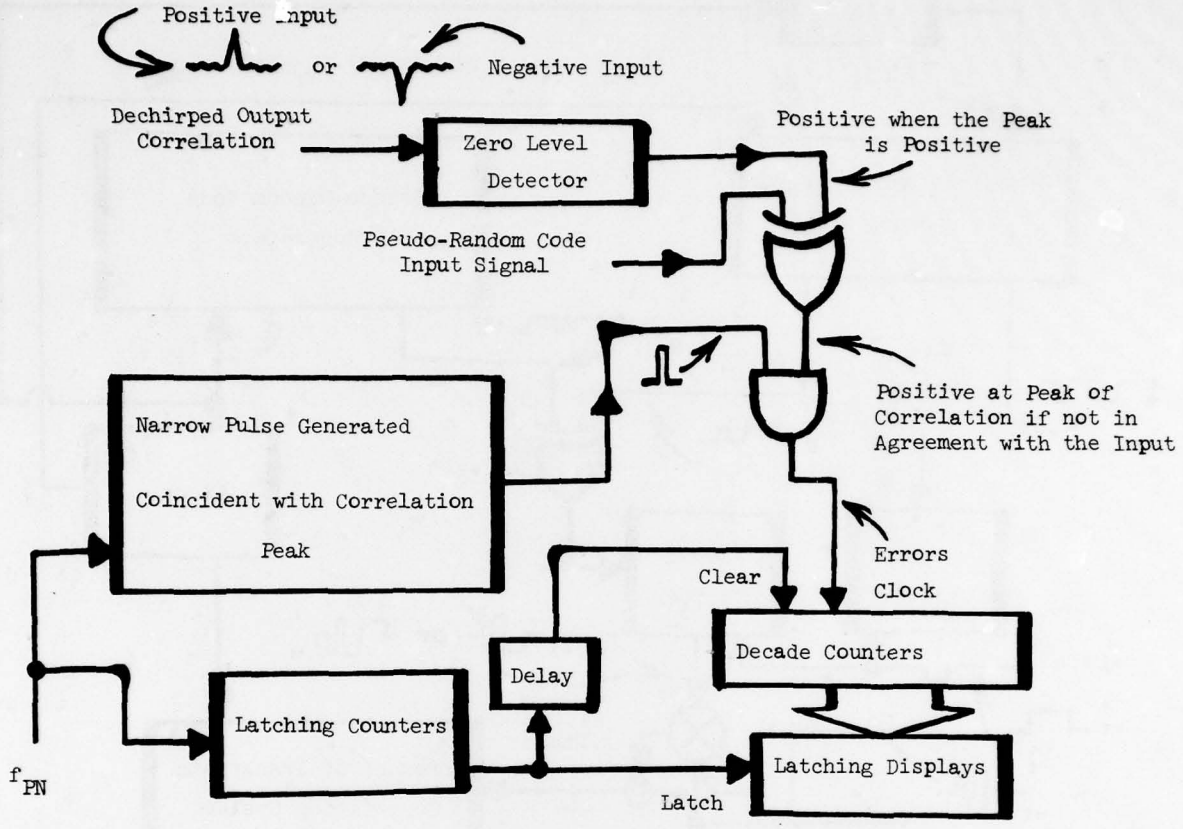


Fig. 15 Block Diagram of the Error Counting Circuitry.

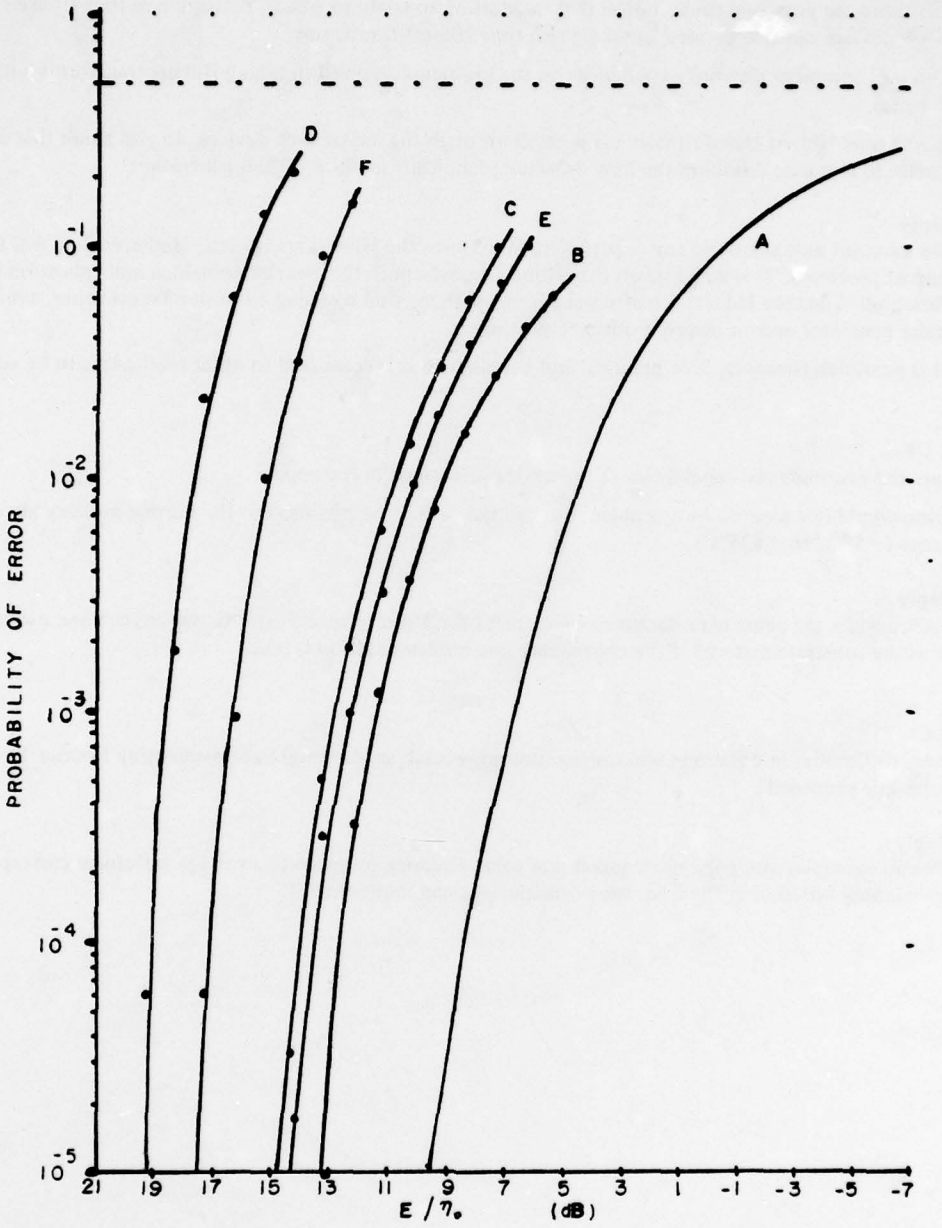


Fig. 16 Probability of Error (PE) Curve for the SAW-Implemented System. Curve A - Optimum Theoretical PE; Curve B - Experimental PE with Approximately White Gaussian Noise; Curve C - Experimental PE for Jammer Level 1.94 dB with Respect to Signal; Curve D - Experimental PE for Jammer Level 9.54 dB with Respect to Signal; Curve E - Same as Curve C, except an Attempt was made to Remove the Jamming; Curve F - Same as Curve D, except an Attempt was made to Remove the Jamming.

DISCUSSION

N.Tepedelenlioğlu, Tu

You mentioned in your oral presentation that in addition to enabling one to perform real time Fourier transformation, SAW devices can also be used in taking real time Hilbert transforms.

(i) Can you comment on your experiences on the problems involved in taking Hilbert transforms with such devices?

(ii) If real time Hilbert transformation is possible through the use of such devices, do you think that one can now realize in a practical fashion the long dreamed phase shift method of SSB generation?

Author's Reply

(i) We have not actually done any experiment to perform the Hilbert transform. However, I do not foresee any unsual problems. It is well known that Hilbert transform in the Fourier domain is multiplication by j–Signum function. Thus the Hilbert transformer can be implemented by using a Fourier Transformer, a mixer and a pulse generator and an inverse Fourier Transformer.

(ii) It is possible. However, how practical and economical it is compared to other methods is to be seen.

J.T.Martin, UK

What are the evironmental capabilities of the devices described in the paper?

Vibration would not seem to be a problem but can the devices be procured to the normal military aircraft temperature range (−55°C to +125°C).

Author's Reply

The SAW devices are being manufactured which meet the Military specifications. Of course, one uses S–T cut quartz as the substrate material if the chirp delay line implementation is used.

A.Sewards, Ca

Are there difficulties in obtaining windowing functions (such as Hanning) in implementing Fourier Transforms using the technique proposed?

Author's Reply

No. For an example, using one more mixer and some circuitry to generate a voltage waveform corresponding to the desired window function in the frequency domain, one can implement it.

AN ANALYSIS OF THE ERROR PROBABILITY OF AN ALL DIGITAL DETECTOR

Sam Reisenfeld and Kung Yao
Department of System Science
University of California
Los Angeles, CA 90024

SUMMARY

In this paper, several analytical approaches are taken to evaluate the error probability of a digital communications system with a detector implemented as a digital signal processor. Specifically, we consider a binary hypothesis problem over an AWGN channel, where the receiver consists of a low-pass filter, a sampler, a dynamic range scaling device, an A/D converter, a digitally implemented integrate-and-dump filter, and a threshold device. The digital processing for the receiver is done with fixed-point arithmetic. Exact error probability expressions are given for cases of overflow and no overflow in the integrate-and-dump filter. Approximate expressions for P_e, which depend upon approximations of the moments of the quantization error, are obtained. The P_e expressions are derived as functions of E_b/N_0, the number of bits in the integrate-and-dump filter output, the number of samples per binary symbol, and the dynamic range scaling factor. Numerical examples illustrating the above analytical approaches are given.

1. INTRODUCTION

Digital signal processing is an attractive method of detection in the design of digital communication systems. There are various advantages in the use of digital signal processing methods for detection implementations. Digital processing techniques lack some of the non-ideal characteristics of analog signal processing techniques such as gain variations, voltage offsets, electronic noise and undesirable nonlinearities. In applications such as matched filtering for spread-spectrum systems and digital receivers in satellite communication systems, the non-ideal characteristics of analog signal processing may create an intolerable level of performance degradation. Digital signal processing may result in substantially better error-rate performance over the life of the system.

Although detectors implemented by digital signal processing have many advantages, these systems have some intrinsic performance limitations. The performance degradation must be characterized analytically and bounded for an appropriate selection of parameters in a system design. Detection systems implemented by digital processing have limited dynamic range and incur error-rate performance loss due to A/D quantization noise and processor overflow. The quantization noise is minimized by using the largest available dynamic range. However, random signals operating near the full system dynamic range have a finite probability of overflow or of reaching saturation. Saturation is undesirable because it is a system nonlinearity and it results in performance degradation for a correlation type receiver operated with an AWGN channel. An amplitude scaling factor must be chosen which minimizes the error probability by a compromise between overflow and quantization error.

Some previous work has been done related to the performance of digital detectors and to the amplitude scaling problem in digital signal processing. Natali [1] investigated the SNR performance of discrete time integrate-and-dump filters. He considered the degradation due to aliasing error when various sampling rates are used with a first order pre-sampling filter, but he neglected the effect of quantization noise. Tufts and Knight [2,3] investigated the SNR of the resultant of the inner product of a vector formed by sampling and quantizing a sinusoid with additive Gaussian noise and another vector formed by sampling and quantizing a sinusoid. Lynn [4] and Cover, Freedman and Hellman [5] studied the P_e performance and structure of a class of digital detectors with the capability to overflow, referred to as finite-memory detectors. Jackson [6] approached the problem of the scaling factor in fixed-point digital filters by using the L_p norm of the input signal spectrum and the transfer functions between the input and various nodes within the filter. Jackson was able to upper bound the signal power at each node by the use of Holder's inequality. Morgan [7] studied discrete-time AGC amplifiers which automatically adjust the scale factor so that the available dynamic range of a fixed-point processor is appropriately utilized.

In this paper, we are interested in the achievable P_e performance of a detector in the form of a discrete-time integrate-and-dump filter constructed from an ideal pre-sampling filter, and A/D converter, and a digital signal processor which performs an accumulator operation. The waveform observed at the channel output is low-pass filtered by an ideal filter and is sampled at the Nyquist rate. Two cases are considered. In the first case, overflow may occur in either the A/D converter or in the digital signal processor. In the second case, overflow may occur only in the A/D converter. A relationship is established among the scaling factor, the number of bits in the A/D converter, the number of bits representing the detection statistic, the number of samples per binary symbol, the E_b/N_0, and the P_e. Approximations for P_e, based on a set of moments of the quantization noise, are given. These approximations are applicable only to a range of the scaling factor where the probability of overflow is small. An approximate expression for P_e, which is also valid over limited ranges of the parameter values based on the Central Limit Theorem, is given. Some specific numerical examples are considered.

Figure 1 shows a block diagram of the system. The signal set consists of binary, antipodal, baseband waveforms which take on the values A or -A during a T-second bit interval. The received waveforms over $t\epsilon[0,T)$ are given by,

$$r(t) = \begin{cases} -A + n(t), & \text{under } H_0 \\ A + n(t), & \text{under } H_1 \end{cases}$$

where $A = \sqrt{E_b/T}$, E_b is the received signal energy per bit, and $n(t)$ is a zero mean Gaussian process with two-sided power spectral density $N_0/2$. The transfer function of the low-pass filter, $H(f)$, has unity magnitude for $|f| \le B$ and zero magnitude elsewhere. The filtered waveform $\hat{r}(t)$ is sampled at the Nyquist rate of $f_s = 2B$ samples/second. The amplifier has a voltage scaling factor of K. The scaling factor is chosen to insure that the A/D converter and the digital signal processor have a low probability of overflow. It is well know that, for an infinite resolution receiver, the optimum detector is a correlation receiver or a matched filter. For a finite precision receiver, operated such that the probability of overflow is small, the system is essentially linear and digital matched filtering will yield reasonably good error rates.

The A/D convertor is characterized as a quantizer. The A/D output is a b+1 bit binary number, where b bits carry magnitude information and one bit carries sign information. The quantizer function Q(r) has uniformly spaced points of discontinuity and uniformly spaced output levels. In general, this quantizer function does not allow the attainment of minimum P_e for a fixed b , (Reisenfeld [8]), but the function is widely used in practice. The quantizer function Q(r) is analytically described by,

$$Q(r) = \begin{cases} 0 & , \quad 0 \le |r| < \tfrac{1}{2}2^{-b} \\ \mathrm{sgn}(r)[m2^{-b}] & , \quad (m-\tfrac{1}{2})2^{-b} \le |r| < (m+\tfrac{1}{2})2^{-b} \\ \mathrm{sgn}(r)[1-2^{-b}], & , \quad (1-\tfrac{1}{2}2^{-b}) \le |r| \end{cases} \qquad \text{for } 1 \le m \le 2^b-1$$

where $\mathrm{sgn}(r) = r/|r|$.

The digital integrate-and-dump filter is shown in Figure 2. The filter uses fixed-point arithmetic. The adder in the filter has a b_a+1 bit binary output. The output is in the sign and magnitude representation. The adder implements the saturation addition operator $\boxplus$ defined by

$$Z = X \boxplus Y = \begin{cases} -[1-2^{-b_A}] & , \quad -\infty < X+Y < -(1-\tfrac{1}{2}2^{-b_A}) \\ X+Y & , \quad -(1-\tfrac{1}{2}2^{-b_A}) < X+Y < (1-\tfrac{1}{2}2^{-b_A}) \\ [1-2^{-b_A}] & , \quad (1-\tfrac{1}{2}2^{-b_A}) < X+Y < \infty \end{cases} .$$

An A/D overflow occurs when $|r| > (1-\tfrac{1}{2}2^{-b})$ and an adder overflow occurs when $|X+Y| > (1-\tfrac{1}{2}2^{-b_A})$. In each T-second interval, L samples are processed for detection, where L= 2BT . The sequence at the sampler output $\{r_i\}$ is defined by

$$r_i = \begin{cases} -A+n_i & , \quad H_0 \\ A+n_i & , \quad H_1 \end{cases} , \qquad 1 \le i \le L ,$$

where $\{n_i\}$ is a zero mean, independent, identically distributed Gaussian sequence with variance N_0B . Figure 2 shows an equivalent noise model of the detector. The A/D converter noise sequence $\{\varepsilon_i\}$ is defined by

$$\varepsilon_i = Q(Kr_i) - Kr_i .$$

The digital comparator performs the following operation:

$$y_L \begin{cases} \ge 0 & , \quad \text{decide } H_1 \\ < 0 & , \quad \text{decide } H_0 \end{cases} .$$

The problem considered in this paper is the evaluation of P_e as a function of E_b/N_0, b, b_A, AK, and L. It is assumed that H_0 and H_1 have equal *a priori* probabilities. L is determined by the pre-sampling bandwidth and the data rate. In an application of the result, the numbers of bits b_A and b are chosen large enough to achieve a specified P_e . AK is chosen to minimize P_e .

The detector performance may be characterized conveniently by the degradation D in dB relative to a matched filter performance at a particular P_e , where

$$D = 10 \log_{10}(2E_b/N_o) - 20 \log_{10}[\mathrm{erfc}^{-1}(P_e)] \tag{1}$$

$$\mathrm{erfc}(x) = 1/\sqrt{2\pi} \int_x^\infty e^{-t^2/2}\, dt .$$

2. EVALUATION OF P_e ; CASE I: $b=b_A$

For the system implemented with equal wordlengths for the A/D output and the adder output, both the A/D and the adder may overflow. The P_e performance of this system may be obtained through Markov chain analysis [9],[4],[5]. The detection statistic at time ℓ is the value of y_ℓ , $1 \le \ell \le L$. Since $b_A=b$, $y_\ell \in \{B_i : 1 \le i \le N_{AD}\}$, where $N_{AD} = 2^{b+1}-1$. The state probability vector $\underline{p}_{\ell,m}$ is an $N_{AD} \times 1$ vector, with i-th element $p_{\ell,m}(i)$ defined as,

$$p_{\ell,m}(i) = \Pr\{y_\ell = B_i / H_m\} . \tag{2}$$

The state transition matrix $\underline{C}_m$ is an $N_{AD} \times N_{AD}$ matrix with i,j-th element $C_m(i,j)$ defined to be

$$C_m(i,j) = \Pr\{y_{\ell+1} = B_i / y_\ell = B_j, H_m\} . \tag{3}$$

The adder output $\{y_\ell\}$ is a first order Markov process because y_ℓ is totally determined by $y_{\ell-1}$ and r_ℓ . Since $y_0 = \underbrace{0_\Delta 0000\ldots0}$, where Δ is the binary point and y_0 corresponds to $B_i\big|_{i=2^b}$,

$$p_{0,m}(i) = \begin{cases} 1 & , \quad i = 2^b \\ 0 & , \quad i \ne 2^b \end{cases} , \qquad 1 \le i \le 2^{b+1}-1 ,$$

since

$$\underline{p}_{\ell+1,m} = \underline{C}_m \underline{p}_{\ell,m} , \qquad \text{then} \qquad \underline{p}_{L,m} = \underline{C}_m^L \underline{p}_{0,m} .$$

Then the P_e is given by

$$P_e = \sum_{i=2^{b_A}+1}^{2^{b_A+1}-1} p_{L,0}(i) + \tfrac{1}{2}\, p_{L,0}(2^{b_A}) \ . \tag{5}$$

The state transition matrix $\underline{C}_m$ is given by:

$$C_m(i,j) = \begin{cases} 1 - \operatorname{erfc}\left\{\sqrt{\frac{2\rho}{L}}\left[\frac{(i-j)+\frac{1}{2})\Delta}{AK} - (-1)^{m+1}\right]\right\}, & i = max\{1, j - \frac{1}{\Delta} + 1\} \\ \operatorname{erfc}\left\{\sqrt{\frac{2\rho}{L}}\left[\frac{(i-j)-\frac{1}{2})\Delta}{AK} - (-1)^{m+1}\right]\right\} - \operatorname{erfc}\left\{\sqrt{\frac{2\rho}{L}}\left[\frac{(i-j)+\frac{1}{2})\Delta}{AK} - (-1)^{m+1}\right]\right\}, & max\{1, j - \frac{1}{\Delta} + 1\} < i < min\{\frac{2}{\Delta} - 1, j + \frac{1}{\Delta} - 1\} \\ \operatorname{erfc}\left\{\sqrt{\frac{2\rho}{L}}\left[\frac{(i-j)-\frac{1}{2})\Delta}{AK} - (-1)^{m+1}\right]\right\}, & i = min\{\frac{2}{\Delta} - 1, j + \frac{1}{\Delta} - 1\} \\ 0, & i < max\{1, j - \frac{1}{\Delta} + 1\} \\ 0, & i > min\{\frac{2}{\Delta} - 1, j + \frac{1}{\Delta} - 1\}, \end{cases}$$

where $\rho = E_b/N_0$ and $\Delta = 2^{-b}$.

3. *EVALUATION OF P_e ; CASE II: No Adder Overflow*

If $b_A = b + \lceil \log_2(L) \rceil$ where $\lceil x \rceil$ is the smallest integer greater than or equal to x , no adder overflow may occur. The magnitude of the A/D output is right-shifted $\lceil \log_2(L) \rceil$ bits. In this case, the performance degradation is caused entirely by A/D quantization noise and A/D overflow. The state probability vector of the A/D output is $\underline{p}_{1,m}$. Since each $q_\ell = Q[Kr_\ell]$ is a deterministic function of Kr_ℓ and the Kr_ℓ are mutually independent random variables, the q_ℓ are mutually independent random variables [10]. Furthermore, without adder overflow,

$$y_L = \sum_{\ell=1}^{L} q_\ell \ .$$

Then, the probability vector at the decision time $\underline{p}_{L,m}$ may be computed by

$$\underline{p}_{L,m} = \mathscr{S}(L-1) \underset{L-1}{\circledast} \underline{p}_{1,m} \ , \tag{6}$$

where $\underset{L-1}{\circledast}$ is the L-1 order convolution operator which is defined to be the operand convoluted with itself L-1 times and $\mathscr{S}(k)$ is the shift operator that maps an $(N+k)\times 1$ vector into and $N\times 1$ vector by the rule

$$\mathscr{S}(k)\underline{B} = \begin{bmatrix} B_{k+1} \\ \vdots \\ B_{N+k-1} \\ B_{N+k} \end{bmatrix} \updownarrow N \ , \qquad \underline{B} = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_{N+k} \end{bmatrix} \ .$$

Eq.(5) may be used for the evaluation of P_e with $B_A = b + [\log_2(L)]$. Then $\underline{p}_{1,m}$ is obtained by,

$$p_{1,m}(i) = \begin{cases} 1 - \operatorname{erfc}\left\{\sqrt{\frac{2\rho}{L}}\left[\frac{3/2\Delta - 1}{AK} - (-1)^{m+1}\right]\right\}, & i=1 \\ \operatorname{erfc}\left\{\sqrt{\frac{2\rho}{L}}\left[\frac{\Delta(i-1/2)-1}{AK} - (-1)^{m+1}\right]\right\} & \\ - \operatorname{erfc}\left\{\sqrt{\frac{2\rho}{L}}\left[\frac{\Delta(i+1/2)-1}{AK} - (-1)^{m+1}\right]\right\}, & 2 \le i \le N_{AD}-1 \\ \operatorname{erfc}\left\{\sqrt{\frac{2\rho}{L}}\left[\frac{1-3/2\Delta}{AK} - (-1)^{m+1}\right]\right\}, & i = N_{AD} \ . \end{cases} \tag{7}$$

Although Eqs.(6) and (7) may be used to evaluate P_e , there is another equivalent representation for P_e . Since the joint density of $\{q_\ell\}$ is multinomial, P_e may be obtained by summing over those terms of the distribution which result in a detection error. Then,

$$P_e = L!\left\{\sum_{\mathscr{L}_1}\prod_{i=1}^{N_{AD}}\left[\frac{p_{1,0}^{k_i}(i)}{k_i!}\right] + \frac{1}{2}\sum_{\mathscr{L}_2}\prod_{i=1}^{N_{AD}}\left[\frac{p_{1,0}^{k_i}(i)}{k_i!}\right]\right\} \quad \text{for } N_{AD} \text{ odd, or } N_{AD} \text{ even and } L \text{ even,}$$

$$P_e = L!\sum_{\mathscr{L}_3}\prod_{i=1}^{N_{AD}}\frac{p_{1,0}^{k_i}(i)}{k_i!} \quad , \text{ for } N_{AD} \text{ even and } L \text{ odd,}$$

where $G = \frac{L(N_{AD}+1)}{2}$, $\mathscr{A} = \left\{\{k_i\}_{i=1}^{N_{AD}}: 0 \le k_i \le L\ ,\ k_i \in \mathscr{C}\ ,\ \sum_{i=1}^{N_{AD}} k_i = L\right\}$,

$\mathscr{L}_1 = \left\{\{k_i\}_1^{N_{AD}} \in \mathscr{A}: \sum_{i=1}^{N_{AD}} ik_i > G\right\}$, $\mathscr{L}_2 = \left\{\{k_i\}_1^{N_{AD}} \in \mathscr{A}: \sum_{i=1}^{N_{AD}} ik_i = G\right\}$,

$\mathscr{L}_3 = \left\{\{k_i\}_1^{N_{AD}} \in \mathscr{A}: \sum_{i=1}^{N_{AD}} ik_i \ge G + \frac{1}{2}\right\}$, $\mathscr{C}$ is the set of integers, and k_i is the number of samples for which the A/D output is equal to B_i , $i=1,2,\ldots,N_{AD}$.

4. MOMENT SPACE APPROXIMATIONS OF P_e

Approximations of P_e and D may be computed by considering $\{\varepsilon_\ell\}$ as an additive noise sequence. Moments of $\sum_{\ell=1}^{L}\varepsilon_\ell$ may be computed and P_e may be approximated through the use of these moments. The decision variable y_L is represented by

$$y_L = \sum_{\ell=1}^{L} q_\ell = X_L + N_C + N$$

where

$$N = \sum_{\ell=1}^{L}\varepsilon_\ell \ , \quad N_C = \sum_{\ell=1}^{L} Kn_\ell \ , \quad \text{and} \quad X_L = AKL \ .$$

Then, under the assumption that N_C and N are statistically independent,

$$P_e = \tfrac{1}{2}\Pr\{y_L \ge 0/H_0\} + \tfrac{1}{2}\Pr\{y_L < 0/H_1\} = \Pr\{N_C > X_L - N\} = \Pr\{N_C > X_L + N\} = E_N\{\operatorname{erfc}[\sqrt{2\rho}\ (1 + \tfrac{N}{AKL})]\}. \quad (8)$$

Although the evaluation of Eq.(8) is analytically difficult, approximations on P_e may be computed by evaluating $\{M_i = E(N/AKL)^i\ , i=1,2,\ldots,2n\}$. Furthermore, if AK is sufficiently small, b is sufficiently large, and ρ is sufficiently large, then a set of approximate moments $\{\hat{M}_i: i=1,2,\ldots,2n\}$ may be evaluated with low computational complexity such that $M_i \simeq \hat{M}_i$ for $i=1,2,\ldots,2n$. A criterion for bounding the region of validity of the approximation is given in Section 6.

If $\Pr\{|Kr_\ell| > 1-(1/2)(2^{-b})\}$ is negligibly small, $(|N|)/AKL$ is essentially upperbounded by $D_M = \Delta/2AK$. If N_{AD} is sufficiently large so that Δ is small relative to $AK\sqrt{L/2\rho}$, then ε_i is essentially uniformly distributed over $[-\Delta/2,\Delta/2]$ (see [11],[12]). Define

$$m_i = E[\varepsilon_\ell^{\,i}] \ . \quad (9)$$

Under the two assumptions mentioned above,

$$m_i \simeq \hat{m}_i = \begin{cases} (\Delta/2)^i(1/(i+1)) \ , & i \text{ even} \\ 0 \ , & i \text{ odd.} \end{cases} \quad (10)$$

Each ε_ℓ is a deterministic function of Kr_ℓ and the Kr_ℓ are mutual independent random variables. Then $\{\varepsilon_\ell\}$ is a sequence of independent random variables [10].

Define,

$$M_i = E[(N/AKL)^i] \ . \quad (11)$$

Let $\hat{M}_i$ be an approximation for M_i , obtained in Eq.(11). The following algorithm for the evaluation of M_i is given in Prabhu [13] .

$$S_k = \left(\sum_{\ell=1}^{k}\varepsilon_\ell\right)$$

$$\theta_i(k) = E[S_k^{\,i}]$$

$$\alpha_i(k) = E[\varepsilon_k^{\,i}] = m_i \ .$$

Then

$$\theta_i(k) = E\left[\left(\sum_{\ell=1}^{k-1}\varepsilon_\ell + \varepsilon_k\right)^i\right] = \sum_{j=0}^{i}\binom{i}{j}\theta_j(k-1)\alpha_{i-j}(k) \ , \quad 1<k \ . \quad (12)$$

Since $\alpha_{i-j}(k) = E[\epsilon_k^{i-j}] = m_{i-j}$,

$$M_i = \frac{\theta_i(L)}{(AKL)^i} \tag{13}$$

is obtainable from iterative use of Eq.(12). The analytical form of m_i defined in Eq.(9) is given in the Appendix.

When the probability density function of ϵ_ℓ is assumed to be uniform over $[-\Delta/2,\Delta/2]$, the odd moments of ϵ_ℓ are zero. In this case,

$$\theta_{2i}(k) = \sum_{j=0}^{i} \binom{2i}{2j} \theta_{2j}(k-1)\alpha_{2i-2j}(k) \quad , \tag{14}$$

where $\alpha_{i-j}(k) = \hat{m}_{i-k}$.

Then,

$$\hat{M}_i = \frac{\theta_i(L)}{(AKL)^i} \quad . \tag{15}$$

Define $\{C_i\}$ by ,

$$C_i = E\{[(1+ M/AKL)^2]^i\} = \sum_{j=0}^{2i} \binom{2i}{j} M_j \ , \tag{16}$$

and

$$\hat{C}_i = \sum_{j=0}^{i} \binom{2i}{2j} \hat{M}_{2j} \ . \tag{17}$$

Yao and Biglieri [14] have presented bounds on P_e based upon the four general forms of the principal representation of Krein in the theory of approximations. Under the assumption that $|N|/AKL$ in Eq.(8) is upper bounded by $D_M<1$ and N/AKL has finite first $2n$ moments, it is shown in [14] that, for n odd,

$$\sum_{j=1}^{(n+1)/2} \rho_j^{(A)} \operatorname{erfc}(\xi_j^{(A)}) \le P_e \le \sum_{j=1}^{(n+3)/2} \rho_j^{(B)} \operatorname{erfc}(\xi_j^{(B)}) \quad , \tag{18}$$

and for n even,

$$\sum_{j=1}^{(n+2)/2} \rho_j^{(D)} \operatorname{erfc}(\xi_j^{(D)}) \le P_e \le \sum_{j=1}^{(n+2)/2} \rho_j^{(C)} \operatorname{erfc}(\xi_j^{(C)}) \quad . \tag{19}$$

The upper and lower bounds in Eqs.(18) and (19) are in the form of quadrature formulas. The sets of weights $\{\rho_j^R: R=A,B,C,D\}$ and the sets of abscissae $\{\xi_j^R: R=A,B,C,D\}$ are given in [14] in terms of $\{C_i\}_1^n$. As n is increased, the lower and upper bound expressions in Eqs.(18) and (19) become arbitrarily tight.

Eqs.(18) and (19) cannot be used directly because $|N|/AKL$ is unbounded. However, under the assumption that ϵ_i is uniformly distributed, $|N|/AKL$ is bounded and Eqs.(18) and (19) can be used with $\{\rho_j^R: R=A,B,C,D\}$ and $\{\xi_j^R: R=A,B,C,D\}$ computed in terms of $\{\hat{C}_i\}_{i=1}^n$. Since the uniform density assumption is an approximation, for n odd,

$$P_e \approx \sum_{j-1}^{(n+1)/2} \rho_j^{(A)} \operatorname{erfc}(\xi_j^{(A)}) \ ,$$

$$P_e \approx \sum_{j=1}^{(n+3)/2} \rho_j^{(B)} \operatorname{erfc}(\xi_j^{(B)}) \ ,$$

and for n even,

$$P_e \approx \sum_{j=1}^{(n+2)/2} \rho_j^{(D)} \operatorname{erfc}(\xi_j^{(D)}) \ ,$$

$$P_e \approx \sum_{j=1}^{(n+2)/2} \rho_j^{(C)} \operatorname{erfc}(\xi_j^{(C)}) \ .$$

5. CENTRAL LIMIT THEOREM APPROXIMATION TO P_e

Since $\{\epsilon_\ell\}$ is a sequence of independent and identically distributed random variables, N/AKL converges in distribution to a Gaussian random variable $L\to\infty$ because of the Central Limit Theorem. Under the following assumptions:

(i) Negligible probability of overflow,
(ii) Δ is small compared to $AK\sqrt{L/2\rho}$,
(iii) L is large,
(iv) N_{AD} is sufficiently large so that ,

$$\frac{Cov(\varepsilon_\ell, Kr_\ell)}{Var^{1/2}(\varepsilon_\ell)\,Var^{1/2}(Kr_\ell)} \simeq 0 \quad ,$$

ε_ℓ is approximately uniformly distributed over $[-\Delta/2,\Delta/2]$ and $\sum_{\ell=1}^{L} \varepsilon_\ell$ has an approximately Gaussian distribution with zero mean and variance $L\Delta^2/12$. The additive closure property of Gaussian random variables may be used to obtain an approximate expression for degradation when (i)-(iv) hold. The expression is:

$$D \simeq \hat{D} = 10 \log_{10}\left(1 + \frac{2\rho}{L}\frac{\Delta^2}{12}\frac{1}{(AK)^2}\right) \text{ dB} . \tag{20}$$

If conditions (iii) and (iv) are valid, Eqs.(8) and (13) may be used to obtain an approximate P_e expression. Through (iii), it is assumed that the central limit approximation may be applied to N/AKL. The result is,

$$P_e \simeq E_N\{\text{erfc}[\sqrt{2\rho}\,(1 + \frac{N}{AKL})]\} \simeq \frac{1}{\sqrt{2\pi}\,\sigma_x}\int_{-\infty}^{\infty} \text{erfc}[\sqrt{2\rho}\,(1 + X)]\exp\left[-\frac{(X-M_1)^2}{\sigma_x^2}\right] dX , \tag{21}$$

where $\sigma_x^2 = M_2 - M_1^2$.

6. REGIONS OF VALIDITY FOR APPROXIMATE PERFORMANCE EXPRESSIONS

The approximations used in Eqs.(18),(19) and (20) are good over limited regions of system parameters. It is assumed that $\{\hat{C}_i\}$ are used to obtain the weights and abscissae in Eqs.(18) and (19). The approximations are applicable only when ε_ℓ has an approximately uniform distribution over $[-\Delta/2,\Delta/2]$.

As AK increases, the probability density function of ε_ℓ becomes increasingly skewed. This phenomenon may be used as a basis for a criterion for defining the regions of validity for the previously mentioned equations. The skewness of a probability density function $P(x)$ may be measured quantitatively [15,16] by the moment coefficient of skewness γ_1 where ,

$$\mu_1 = E[x] ,$$
$$\nu_2 = E[(x-\mu_1)^2] ,$$
$$\nu_3 = E[(x-\mu_1)^3] ,$$
$$\gamma_1 = \frac{\nu_3}{\nu_2^{3/2}} .$$

The coefficient of skewness is zero of $P(x-\mu_1)$ is an even function, becomes increasingly positive as $P(x-\mu_1)$ is increasingly skewed to the right, and becomes negative as $P(x-\mu_1)$ is increasingly skewed to the left. Since a measure of skewness about the origin is required for the probability density function of ε_ℓ , a modified coefficient of skewness Γ_i may be defined by ,

$$\Gamma_i = \frac{m_{2i+1}}{\left(m_{2i}\right)^{\frac{2i+1}{2i}}} \quad , \quad i=1,2,3,4,\ldots .$$

For Eqs.(18) and (19), $\{m_i\}_1^{2n}$ are used to determine bounds on P_e and the region where the approximations are applicable may be specified by ,

$$\{AK,\ N_{AD},\ L,\ \rho :\ |\Gamma_i| \le C ,\ 1 \le i \le n\} .$$

For Eq.(20), the region where the approximations are applicable may be specified by

$$\{AK,\ N_{AD},\ L,\ \rho :\ |\Gamma_1| \le C\} ,$$

where, by empirical determination, a reasonable value of C is 0.1 .

7. NUMERICAL EXAMPLES

In order to illustrate the concepts discussed in Sections 2-6, some numerical results are given here. The error probability expressions depend upon the following independent variables: the SNR(dB) = 10 log (E_b/N_0) the number of samples L per binary decision; the number of bits b in the A/D output; the number of bits b_A in the adder output; and the scaling factor AK .

The degradation D in dB given in Eq.(5) is plotted against AK in Figure 3. For this family of curves, SNR = 8dB , L=8 , and $b \in \{1,2,3,4,5\}$. Both cases of adder overflow, with $b_A=b$, and no adder overflow, with $b_A = b + \lceil \log_2(L) \rceil$, are shown.

Figure 4 shows the degradation in D in dB vs. AK for SNR=8dB , L=8 , and $b \in \{1,2,3,4\}$ in the no-adder-overflow case. The actual performance from Eq.(5) and the approximate performance from Eq.(20) are also shown.

Figure 5 shows the moment-space approximations D_2^L , D_2^U , D_5^L , D_5^U and D in the no-adder-overflow case for SNR = 8dB, L=8 , and b=4 . D_n^L and D_n^U denote the smaller and larger approximations on D , respectively, when $\{\hat{C}_i\}_1^n$ are used to evaluate the approximations.

Table I shows the moment-space approximations on D for SNR = 8dB , L=8 , and b=4. Table II shows D_2^L, D , D_2^U , $|\Gamma_1|$, and $|\Gamma_2|$ for SNR = 10dB , L=6 , b=4 and $0.1 \le AK \le 1.0$.

TABLE I: Moment Space Approximations to Detector Performance Degradation Relative to Matched Filtering using the Uniform Density Assumption. Parameters: SNR = 8dB , b = 4 , L = 8

AK	D, dB								
	D_2^L	D_3^L	D_4^L	D_5^L	D	D_5^U	D_4^U	D_3^U	D_2^U
.05	.894(-1)	.572	.635	.751	.966	1.259	1.476	2.757	3.253
.10	.611(-1)	.193	.204	.214	.267	.229	.247	.506	.724
.15	.367(-1)	.922(-1)	.953(-1)	.971(-1)	.121	.986(-1)	.101	.162	.248
.20	.274(-1)	.532(-1)	.544(-1)	.549(-1)	.685(-1)	.552(-1)	.558(-1)	.757(-1)	.114
.25	.199(-1)	.344(-1)	.349(-1)	.351(-1)	.439(-1)	.352(-1)	.354(-1)	.436(-1)	.638(-1)
.30	.150(-1)	.240(-1)	.242(-1)	.243(-1)	.304(-1)	.243(-1)	.244(-1)	.284(-1)	.402(-1)
.35	.117(-1)	.176(-1)	.177(-1)	.178(-1)	.225(-1)	.178(-1)	.178(-1)	.200(-1)	.275(-1)
.40	.932(-2)	.134(-1)	.135(-1)	.135(-1)	.183(-1)	.135(-1)	.136(-1)	.148(-1)	.198(-1)
.45	.759(-2)	.105(-1)	.106(-1)	.106(-1)	.180(-1)	.106(-1)	.106(-1)	.114(-1)	.150(-1)
.50	.628(-2)	.848(-2)	.852(-2)	.852(-2)	.226(-1)	.852(-1)	.853(-2)	.905(-2)	.116(-1)
.55	.527(-2)	.695(-2)	.697(-2)	.698(-2)	.332(-1)	.698(-2)	.698(-2)	.734(-2)	.928(-2)
.60	.447(-2)	.578(-2)	.580(-2)	.580(-2)	.499(-1)	.580(-2)	.581(-2)	.606(-2)	.756(-2)
.65	.383(-2)	.487(-2)	.489(-2)	.489(-2)	.723(-1)	.489(-2)	.489(-2)	.507(-2)	.625(-2)
.70	.331(-2)	.415(-2)	.414(-2)	.416(-2)	.997(-1)	.416(-2)	.416(-2)	.430(-2)	.525(-2)
.75	.288(-2)	.357(-2)	.358(-2)	.358(-2)	.131	.358(-2)	.358(-2)	.368(-2)	.445(-2)
.80	.252(-2)	.309(-2)	.310(-2)	.310(-2)	.165	.310(-2)	.310(-2)	.318(-2)	.381(-2)
.85	.221(-2)	.270(-2)	.270(-2)	.270(-2)	.202	.270(-2)	.270(-2)	.276(-2)	.329(-2)
.90	.195(-2)	.236(-2)	.237(-2)	.237(-2)	.239	.237(-2)	.237(-2)	.242(-2)	.286(-2)
.95	.173(-2)	.208(-2)	.209(-2)	.209(-2)	.277	.209(-2)	.209(-2)	.213(-2)	.251(-2)
1.00	.154(-2)	.184(-2)	.184(-2)	.186(-2)	.316	.185(-2)	.185(-2)	.188(-2)	.220(-2)

TABLE II: Moment Space Approximations to D and Modified Coefficients of Skewness: SNR=10dB, L=6, b=4

AK	D_2^L	D	D_2^U	$\lvert\Gamma_1\rvert$	$\lvert\Gamma_2\rvert$
0.1	.921(-1)	.558	1.452	7.948(-3)	5.988(-3)
0.2	.458(-1)	.152	.303	1.551(-8)	1.198(-8)
0.3	.270(-1)	.669(-1)	.106	5.544(-12)	3.659(-9)
0.4	.178(-1)	.391(-1)	.508(-1)	5.042(-5)	7.314(-4)
0.5	.126(-1)	.253(-1)	.294(-1)	1.194(-1)	2.697
0.6	.941(-2)	.192(-1)	.191(-1)	3.325	4.444
0.7	.730(-2)	.213(-1)	.134(-1)	4.202	3.001
0.8	.584(-2)	.351(-1)	.988(-2)	3.080	2.351
0.9	.478(-2)	.620(-1)	.762(-2)	2.407	2.009
1.0	.400(-2)	.101	.606(-2)	2.033	1.807

8. CONCLUSION

The error rate performance analysis of a receiver which does detection by digital processing has been presented. For the case of overflow in the detection filter, the P_e was obtain through Markov chain analysis. Overflow may be eliminated in the detection filter by using a sufficiently long binary word at the adder output. With no filter overflow, the P_e expression was given in the form of a multiple convolution of the A/D output probability vector with itself. In the above approaches, the resulting error probability expressions are exact.

Other approaches to the error-rate analysis, which uses various approximations, were presented. Approximate P_e expressions were obtained in terms of approximations to the moments of the quantization error. Approximate P_e expressions dependent upon Gaussian approximations were given.

For small values of the scaling factor AK , the detection performance was essentially independent of whether detection filter overflow is possible or not, because the probability of overflow is small. For larger AK, significantly better P_e performance may be realized by incurring the slightly increased receiver complexity necessary to prevent overflow. The value of AK which best compromises between overflow and quantization error yields the best P_e performance.

For a given channel and set of performance requirements, the results may be used for the selection of wordlengths. Since AK will usually be set by an automatic gain control loop, the sensitivity of P_e with respect to small variations of AK is important.

Finally, it is of some interest to give comparisons of advantages and disadvantages of the various approaches used in the evaluation of P_e for the digital detector. The Markov chain analysis (Eqs. 2-5) of the state of the adder output yields exact error rate results. However, the number of multiplications needed to evaluate this P_e increases exponentially with respect to the adder output wordlength and linearly with respect to the number of samples per decision. For systems with small adder wordlength, direct evaluation of this P_e is feasible but clearly, for large adder wordlength, direct evaluation is prohibitively costly. The convolution approach (Eqs.5-7) also yields exact P_e. Now, the number of multiplications needed to evaluate P_e increases exponentially with respect to the A/D output wordlength and quadratically with respect to the number of samples per decision. In the quadrature approach (Eqs.8-19), the computational effort needed to evaluate all the relevant moments, as well as the quadrature formulas, are modest. In the Central Limit Theorem approaches, the degradation expression given in Eq.(20) is explicit, while the approximation to P_e given in Eq.(21) is also simple to evaluate. Thus, several approaches are possible to the evaluation of error probability. Exact performances can be obtained generally with high computational cost, while approximate performances can be obtained with more modest computational effort.

APPENDIX: A Functional Description of m_i

The following two lemmas may be used to evaluate m_i.

LEMMA 1: For the quantizer function $Q(x)$ described in the Introduction,

$$m_n = (-1)^n \sum_{k=0}^{n} \binom{n}{k} \left\{ \sum_{i=1}^{N_{AD}} (1-i\Delta)^{n-k} \int_{\Delta(i-1/2)-1}^{\Delta(i+1/2)-1} x^k p_x(x)dx + (1-N_{AD}\Delta)^{n-k} \int_{\Delta(N_{AD}-1/2)-1}^{\infty} x^k p_x(x)dx \right.$$

$$\left. + (1-\Delta)^{n-k} \int_{-\infty}^{-(1-\Delta/2)} x^k p_x(x)dx \right\}$$

where $p_x(x)$ is the probability density function of the quantizer input.

LEMMA 2: If x is a Gaussian random variable with mean μ and variance σ^2,

$$\int_a^b x^n p(x)dx = \sum_{k=0}^{n} \binom{n}{k} \sigma^k \mu^{n-k} I\left(\frac{a-\mu}{\sigma}, \frac{b-\mu}{\sigma}, k \right),$$

where

$$I(A,B,k) = \frac{1}{\sqrt{2\pi}} \left(A^{k-1} e^{-A^2/2} - B^{k-1} e^{-B^2/2} \right) + (k-1)I(A,B,k-2),$$

and

$$I(A,B,0) = \text{erfc}(A) - \text{erfc}(B)$$

$$I(A,B,1) = \frac{1}{\sqrt{2\pi}} \left(e^{-A^2/2} - e^{-B^2/2} \right).$$

REFERENCES

[1] NATALI, Francis D., 1969, "Comparison of Analog and Digital Integrate and Dump Filters," *Proc.IEEE, Oct. 1969,* 1766-1768.

[2] TUFTS, D.W., and KNIGHT, W.C., 1969, "Statistics of the Inner Product of a Quantized, Random Vector and a Constant Vector," *Proc.IEEE (Letters),* 57, 821-822.

[3] TUFTS, D.W., and KNIGHT, W.C., 1970, "How Linear Prefiltering Prior to Quantization Affects Signal Detectability," *Proc.IEEE (Letters),* April, 599-600.

[4] LYNN, P., 1971, "Finite Memory Detectors," Ph.D. Dissertation, Polytechnic Institute of Brooklyn.

[5] COVER, T., FREEDMAN, M., and HELLMAN, M., 1976, "Optimal Finite Memory Learning Algorithms for the Finite Sample Problem," *Inform. and Contr.,* 30, 49-85.

[6] JACKSON, L.B., 1969, "An Analysis of Roundoff Noise in Digital Filters," Ph.D. Dissertation, Stevens Institute of Technology.

[7] MORGAN, D., 1975, "On Discrete Time AGC Amplifiers," *IEEE Trans. Circuits and Systems,* CAS-22,

[8] REISENFELD, S., 1977, "Digital Detection with Quantizer Observations," *Proc. 1977 Intl.Symp. on Information Theory,* Cornell University, Ithaca, NY.

[9] KEMENY, J.G. and SNELL, J.L.,1960, *Finite Markov Chains,* D. Van Nostrand Co., Princeton, NJ.

[10] CHUNG, K.L., 1974, *A Course in Probability Theory,* Academic Press, NY., p. 50.

[11] OPPENHEIM, A.V., and WEINSTEIN, C.J., 1972, "Effects of Finite Register Length in Digital Filtering and the Fast Fourier Transform," *Proc. IEEE,* 60, 957-976.

[12] OPPENHEIM, A.V., and SCHAFER, R.W., 1975, *Digital Signal Processing,* Prentice-Hall, Englewood Cliffs, NJ.

[13] PRABHU, V.K., 1971, "Some Considerations of Error Bounds in Digital Systems," *BSTJ,* 50, 3127-3138.

[14] YAO, K., and BIGLIERI, E., 1977, "Moment Inequalities for Error Probabilities in Digital Communication Systems," *NTC'77 Conf.Rec.,* Los Angeles, CA, 26:4-1 - 26:4-4.

[15] BURINGTON, R.S., and MAY, D.C., 1970, *Handbook of Probability and Statistics with Tables*, McGraw-Hill, New York.

[16] JOHNSON, N.L., and LEONE, F.C., 1964, *Statistics and Experimental Design in Engineering and the Physical Sciences*, Wiley & Sons, New York.

ACKNOWLEDGMENT

This work was supported by the Electronics Program of the Office of Naval Research.

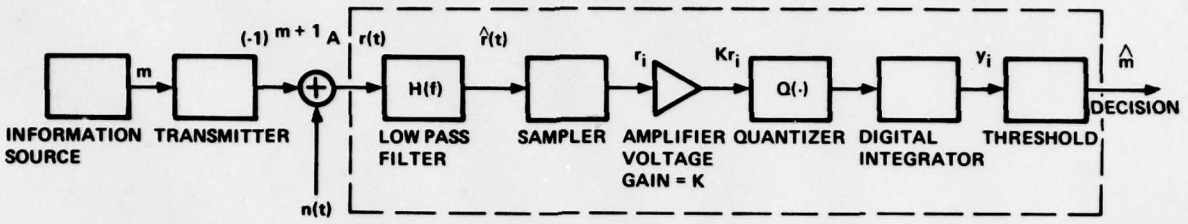


Fig.1 The block diagram of the communication system implemented with an all digital receiver

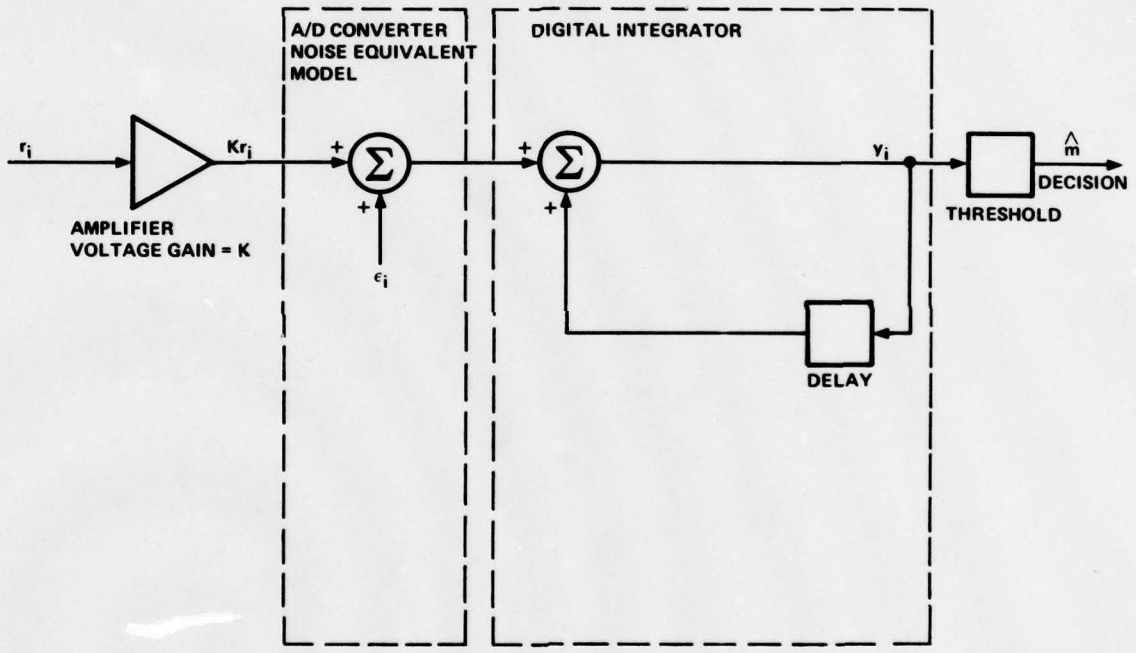


Fig.2 The equivalent noise model of the digital receiver

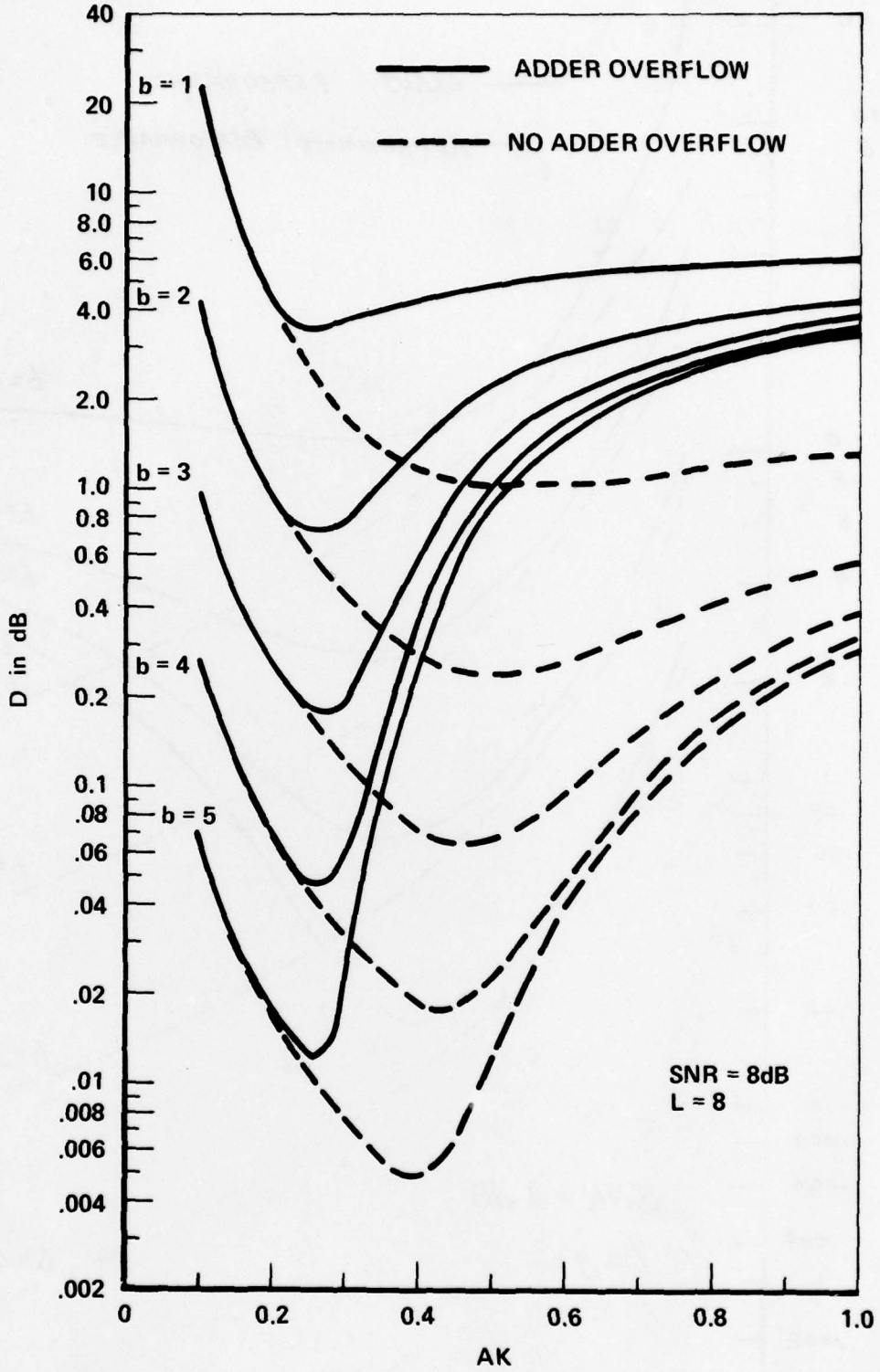


Fig.3 Degradation D in dB versus AK

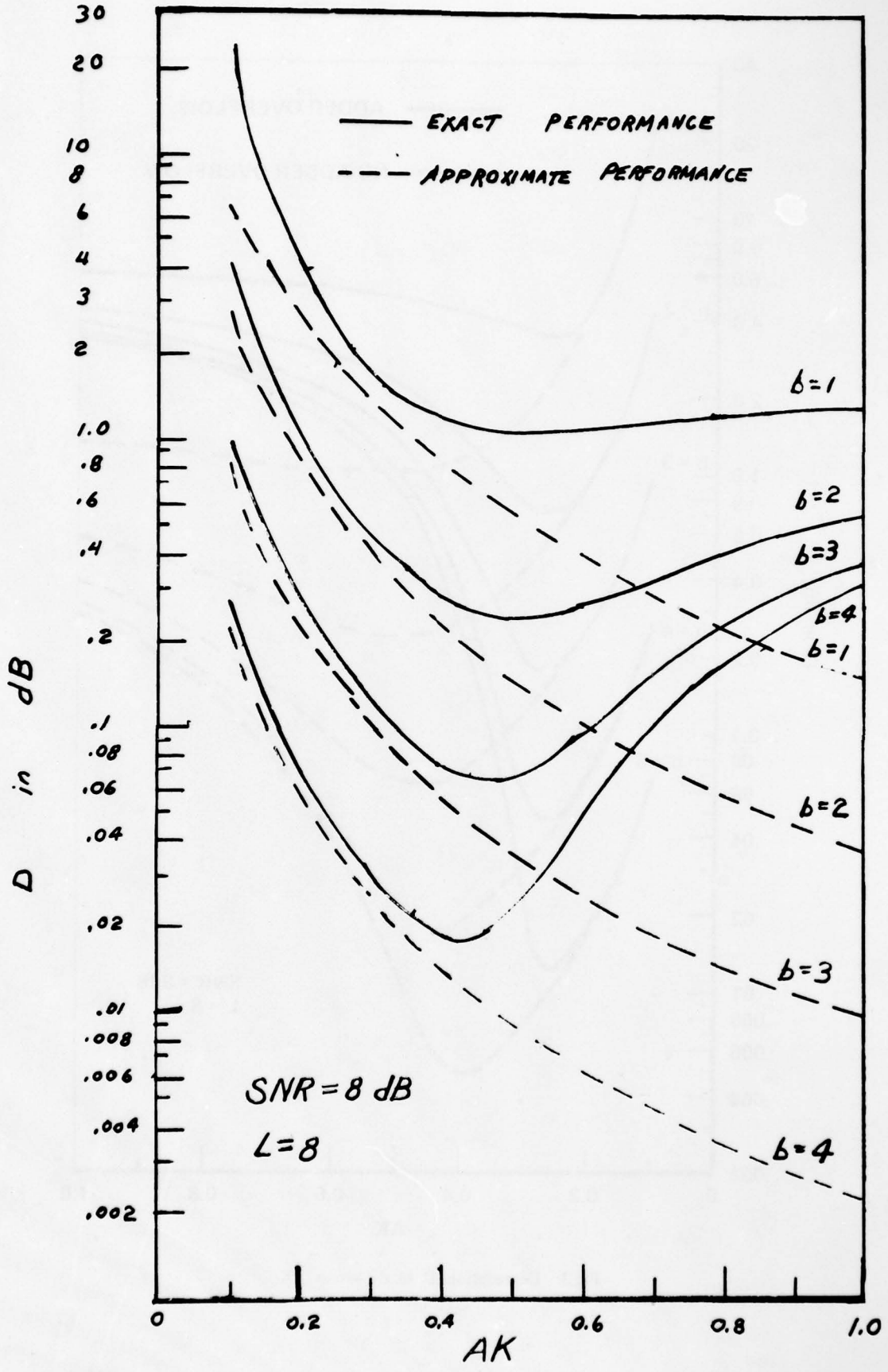


Fig.4 Central limit theorem approximation to detector performance

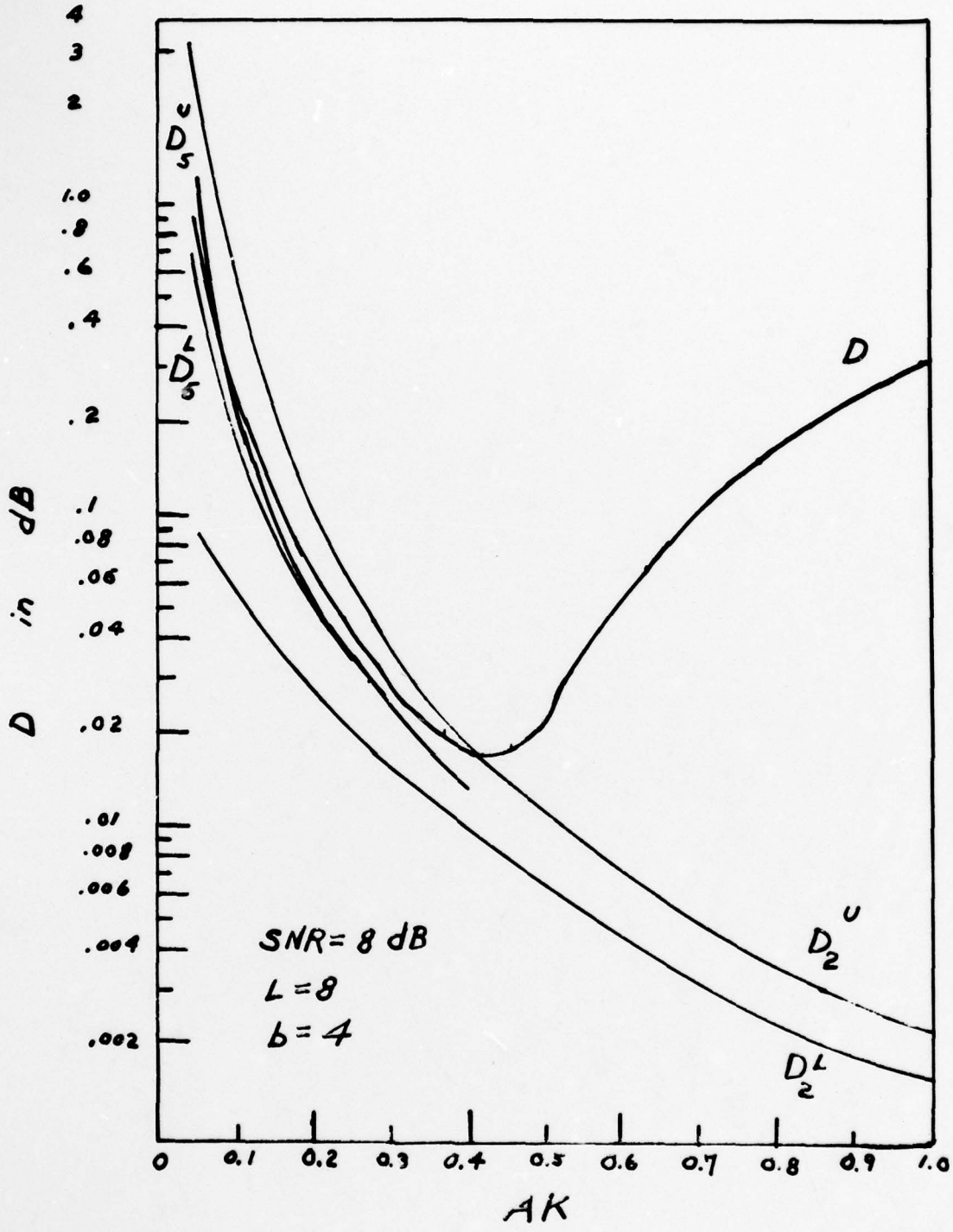


Fig.5 Moment space approximation to detector performance

ASPECTS OF SOURCE ENCODING

Dietrich Wolf
Institut für Angewandte Physik
Universität Frankfurt
Robert-Mayer-Straße 2-4
D-6 Frankfurt am Main, FR Germany

SUMMARY

After some general remarks on the concepts of source encoding some illustrative examples of actual interest are presented. In particular, recent results on source models of bandpass-limited speech signals, on optimum quantization of non-Gaussian random sources, on data compression encoding schemes for speech and television signals are discussed.

1. INTRODUCTION

Man or in general biological, physical, and technical systems are sources of information. They mutually communicate by transmitting and receiving physical representations of the source information which, e.g., can be sequences of numbers, letters, words or patterns, sounds or images, or mechanical or electrical waveforms. Frequently the physical representations of source outputs are redundant or even contain parts irrelevant to the possible user. In order to find representations which incorporate significant information only, one can appropriately encode the source information.

A mathematical treatment of source encoding firstly requires a model of the information source. Source models in information theory are random process models, the simplest type of which is a discrete memoryless source. This source generates a sequence of statistically independent symbols randomly selected from a predetermined alphabet according to some fixed set of symbol probabilities. If the probabilistic description is independent of time the source is classified to be stationary. The stationary discrete memoryless source may be used as a first order approximation of actual sources, which in general show statistical dependences between successive symbols.

Source encoding assigns the symbols of the source alphabet or blocks of them to corresponding symbols of some code alphabet. In practice the encoding rules are chosen in such a way that the source sequence can be reconstructed from the encoded sequence and that the average number of code symbols per source symbol is minimized. As is well known, employing a binary code one finds the minimum average number of binary symbols of a discrete memoryless source to be the source entropy H (in bits per source symbol). If the discrete source possesses a memory and is ergodic that minimum average number of binary symbols is determined by the conditional entropy H_n.

Important information sources often are found to be nondiscrete but continuous where the source output is a random process. Obviously it is impossible to encode a continuous source output into a sequence of discrete (binary) symbols from which the source information can be reconstructed exactly. In this case the reconstruction of the source output in principle can be achieved only approximately within certain limits of distortion. Then the problem arises to find a source encoding scheme which yields a minimum average number of code symbols and satisfies a given fidelity criterion. This fidelity criterion depends upon the statistical properties of the source and on a measure of the distortion. Usually continuous source outputs can be encoded by taking samples of the signal, quantizing the samples, and encoding the quantized data into a sequence of code symbols. This procedure always implies a loss of information and hence the introduction of unavoidable distortion along with the quantization.

Up to now the fundamental problem in source encoding remains to develop appropriate feasible models of the information sources such as speech, data, and picture signals and to find reasonable measures of distortion. Although remarkable progress in treating this complex problem has been achieved in the last years it will be still a long way to a sufficient solution. The complexity is due to the mutual interferences of various physical, physiological, and psychological aspects of perception, recognition, and processing of information in the human mind.

The concepts of source encoding shall be illustrated in this paper by some selected actual examples of speech and picture encoding. In particular, recent results on source models of bandpass-limited speech signals in conventional telephone channels, on optimum quantization of non-Gaussian random sources, and on data compression encoding schemes for speech and television signals will be discussed.

2. SOURCE MODELS

Recent investigations on telephone speech-waves (300 ... 3400 Hz) have led to a source model of speech which allows an analytical treatment of many practical, important problems of speech encoding, e.g. linear predictive coding, optimum quantization including statistical dependences, linear and some nonlinear transform coding schemes (WOLF, 1977; WOLF and BREHM, 1977).

This model describes speech signals by the product of two mutually statistically independent Gaussian processes and it approximates an elliptically invariant stochastic process. From this model the joint amplitude distribution densities $p_{\underline{\xi}}(x)$ of n amplitudes $\xi_i=\xi(t_i)$ where $\underline{\xi}$ denotes the column vector with the n components ξ_i $\underline{\xi}$ extracted from the random process $\xi(t)$, have been derived. Using these expressions numerical calculations of the conditional entropies of optimally logarithmic-quantized model signals presently are performed. First results (•) are shown in Fig. 1 compared to the measured data (o) of COHEN (MUSMANN, 1977). A good agreement can be stated. The curves indicate that the conditional entropy H_n (H_o is the decision content) is reduced roughly by about 3 bits per symbol if a memory of the source is taken into account.

This approach to a source model by multiplying two random processes - or in a similar way by multiplying two random sequences - has successfully been extended to large classes of continuous or discrete sources with and without memory. Thus, models characterized by a probability density function (pd) which corresponds to gamma, Laplacian, or Bessel functions K_0, resp., have been established.

3. QUANTIZATION - CODING WITH FIDELITY CRITERION

Quantization, i.e. the conversion of a continuous valued variable into a discrete valued one, is a typical problem of source encoding with a fidelity criterion. A common fidelity criterion is the maximum tolerable value for the mean-squared-error between reconstructed value and source value. The minimum number R_{min} of code symbols (or the minimum number of equivalent binary symbols) per source symbol, subject to the condition that the average distortion does not exceed the given limit D as defined by the fidelity criterion, is determined by rate distortion theory (GALLAGER, 1968; BERGER, 1970). The function $R_{min}(D)$, the so-called rate distortion function, represents the strict theoretical bound which is an absolute constraint to any encoding scheme. An arbitrary encoding scheme can be characterized by a certain point in the R_{min}-D plane. The distance of this point from the rate distortion function then may be considered to be a measure of the effectiveness of the source encoding scheme. Mostly calculations of functions $R_{min}(D)$ are quite difficult since the sources in communication systems generally are not memoryless and moreover are governed by non-Gaussian statistics. Therefore, for non-Gaussian sources analytical results are known only if the sources are memoryless (except for few particular cases).

Fig. 2 presents for three source models - with Laplacian pd (L), gamma pd (Γ), and Bessel function K_0 pd (K_0) of source output - and for various 2^n levels quantization schemes the mean number R in bits per sample as a function of distortion D. In Fig. 2 the open circles, squares, and triangles indicate the different optimum quantization algorithms (Δ uniform, ∇ quasi-uniform, $\square$ nonuniform logarithmic, o according to MAX) while the corresponding full symbols relate to those values obtained if the quantizer outputs are further encoded by variable-length Huffman code (ZIMMERMANN, 1977). The results demonstrate uniform quantization combined with Huffman coding to yield the best performance. Doing so no more than 0.3 bits per sample are required in addition to R_{min} for equal distortion D.

It may be noticed that in practice estimating the advantage of a quantization scheme additional criteria like the sensitivity to mismatch of source statistics and source output power level or the expense of the coder hardware have to be taken into consideration. The sensitivity of Max quantizers which are optimized to Gaussian or K_0 pd, resp., to mismatch in source pd which is assumed to be a Gaussian pd (G), gamma pd (Γ), K_0 pd (K_0), Laplacian pd (L) or uniform pd (U) are illustrated in Fig. 3. The curves show that a Gaussian pd matched quantizer is very sensitive leading to considerably increased distortion while the K_0 pd matched quantizer is insensitive to the choice of another source statistics.

4. DATA COMPRESSION

In order to eliminate redundant or irrelevant data of the source output the source encoding scheme has to be arranged for achieving data compression. Often data compression is required if the given capacity of a channel is exceeded by the rate of information to be transmitted. Since in this case distortion is unavoidable one has to take care that the significant information reaches the destination. The criteria of what is significant or what is irrelevant in general depend on subjective, perceptual conditions and thus data compression algorithms mostly cannot be based on strict theoretical analysis.

One important application of data compression is found in digital speech communication systems. Many different methods for speech waveform encoding have been proposed during the last years. Meanwhile some of them like pulse code modulation (PCM) or delta modulation (DM) have been introduced into public communication networks. More recently adaptive algorithms or transform coding schemes which, too, may be adaptive are discussed which promise an appreciable reduction of the transmission rate R for a given distortion.

Fig. 4 gives a survey of performance limits classified by the signal-to-noise ratio (SNR) as have been observed experimentally by employing various encoding algorithms (NOLL, 1975; JAYANT, 1976; ZELINSKI and NOLL, 1977). This comparison includes, besides DM and PCM, adaptive and non-adaptive differential pulse code modulation (ADPCM, DPCM) and adaptive and non-adaptive transform coding (ATC, TC). All of these make use of an additional quantization with forward estimation (AQF). The results show clearly that adaptive transform coding provides for the lowest bit rate at a given distortion or at a given bit rate for the best SNR value.

It shall be mentioned that these new data compression encoding strategies as well as

Table 1. Picture Coding Schemes achieving minimum quality class 4 of CCIR

Signal Type	Encoding Scheme	Sampling Frequency MHz	Quantization bits/pel	Bit Rate Mbits/sec
Video-Telephone	PCM	2	8	16
	DPCM + SQ	2	4	8
TV monochrome	PCM	10	8	80
	DPCM	10	6	60
	DPCM 2d	10	5	50
	DPCM 2d + SQ	10	4	30 †
Color TV	Composite: PCM	13.3	8	107
	DPCM	8.8	6	53
	DPCM 2d	13.3	5	67
	Component: U_Y DPCM 2d + SQ	10	4	
	U_{R-Y} DPCM	2	2	34.4
	U_{B-Y} DPCM	2	2	
	U_Y DPCM 2d + SQ	8.8	4	
	U_{R-Y} DPCM	2.2	2	34 †a
	U_{B-Y} DPCM	2.2	2	

SQ switched quantizer; 2d two-dimensional spatial linear prediction; † using blankings
a extensive quality tests running

those based on vocoder principles still suffer from their expensive technical realization and some lack in processing velocity. It may be expected that further developments in IC-technology will overcome those difficulties in the near future.

Today data compression techniques also are studied in the field of image processing. Here one aim is to reduce the bit rate of color TV signals to as low as 34 Mbits/sec while retaining the conventional CCIR recommended quality class. Two different encoding techniques are under investigation (MUSMANN, 1978). The component technique separately encodes luminance (U_Y) and chrominance (U_{R-Y} and U_{B-Y}) signals each, while the composite technique encodes the complete color TV signal. As can be seen from Table 1 which summarizes characteristic data and results of picture coding methods satisfying the CCIR quality class 4 component coding leads to a bit rate close to 34 Mbits/sec. The bit rate of 34 Mbits/sec will be compatible with the recommended third stage of PCM hierarchy and will allow transmission over the standard radio relay network or satellite communication channels.

5. CONCLUSION

The few examples discussed here have shown that many interesting and important results have been obtained and various ingenious strategies have been developed in the field of source encoding and data compression. However, yet many problems are unsolved and remain a challenge to the scientists in this field. It is again impressive to see how effective biological systems can extract and condense relevant and delete irrelevant information even in the presence of severe distortion. A clue to the solution of the fundamental problems of source encoding and data compression may be provided by a better insight into the perceptual and information processing principles in nature.

6. REFERENCES

BERGER, T., 1970, "Rate Distortion Theory", Prentice Hall, Englewood Cliffs.

GALLAGER, R.G., 1968, "Information Theory and Reliable Communication", Wiley, New York.

JAYANT, N.S. Ed., 1976, "Waveform Quantization and Coding", IEEE Press, New York.

MUSMANN, H.G., 1977, private communication.

MUSMANN, H.G., 1978, "Digital Coding of Television Signals", to be published.

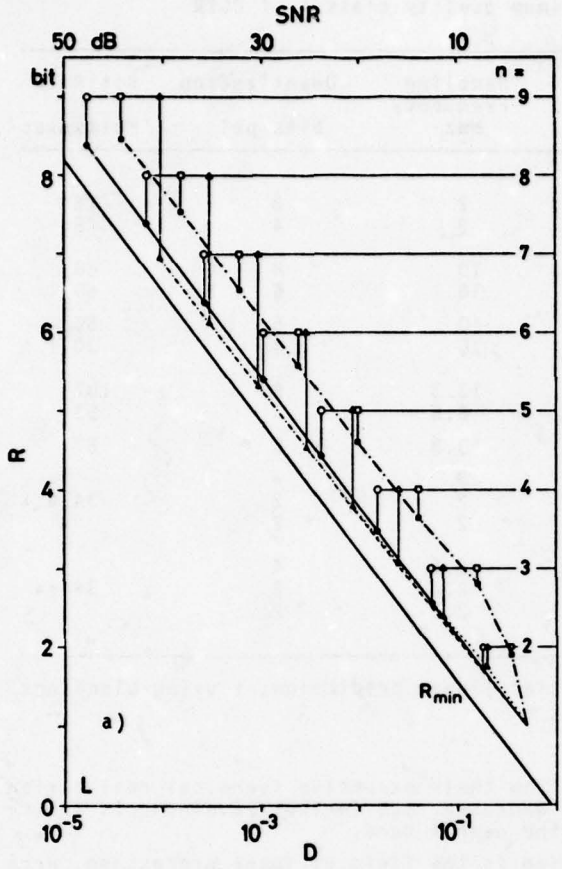
NOLL, P., 1975, "A comparative Study of Various Quantization Schemes for Speech Encoding" Bell System Techn. J. 54, 1597-1614.

WOLF, D., 1977, "Analytische Beschreibung von Sprachsignalen", Archiv elektr.Obertragung 31, 392-398.

WOLF, D. and H. BREHM, 1977, "Mathematical Treatment of Speech Signals", IEEE Symposium on Information Theory, Ithaca N.Y.; IEEE-cat.No. 77 CH 1277-3 IT, p.105.

ZELINSKI, R. and P. Noll, 1977, "Adaptive Transform Coding of Speech Signals", IEEE-Trans. on Acoustics, Speech and Signal Processing, ASSP-25, 299-309.

ZIMMERMANN, G., 1977, "Optimale Quantisierung stochastischer Modellprozesse für Sprache", Dipl. Thesis, Inst. Appl. Phys.Univ. Frankfurt.



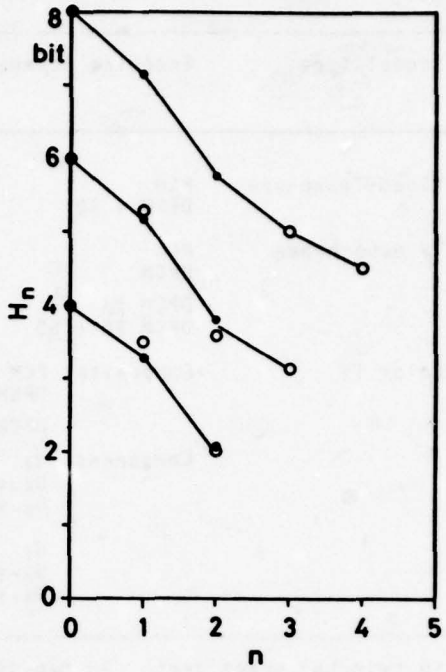
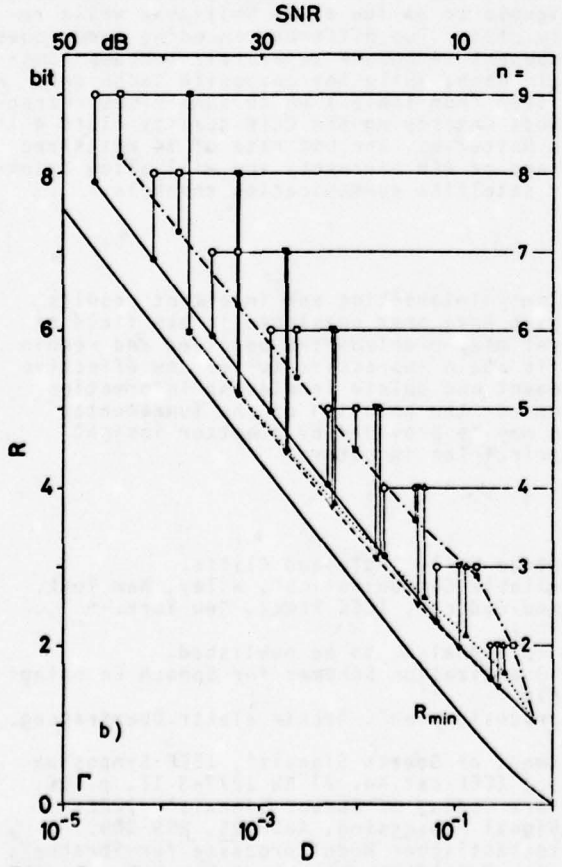


Fig. 1. Conditional entropy H_n vs. number n of statistically dependent symbols



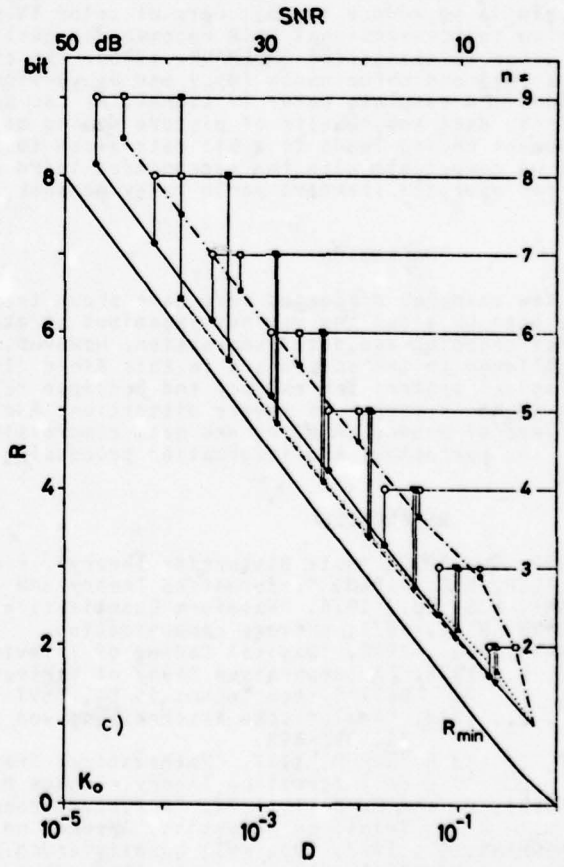


Fig. 2. Rate distortion function $R_{min}(D)$ and average number R of code symbols per source symbol achieved by 2^n levels quantization schemes vs. distortion D.
a) Laplacian pd source, b) Γ pd source, c) K_o pd source

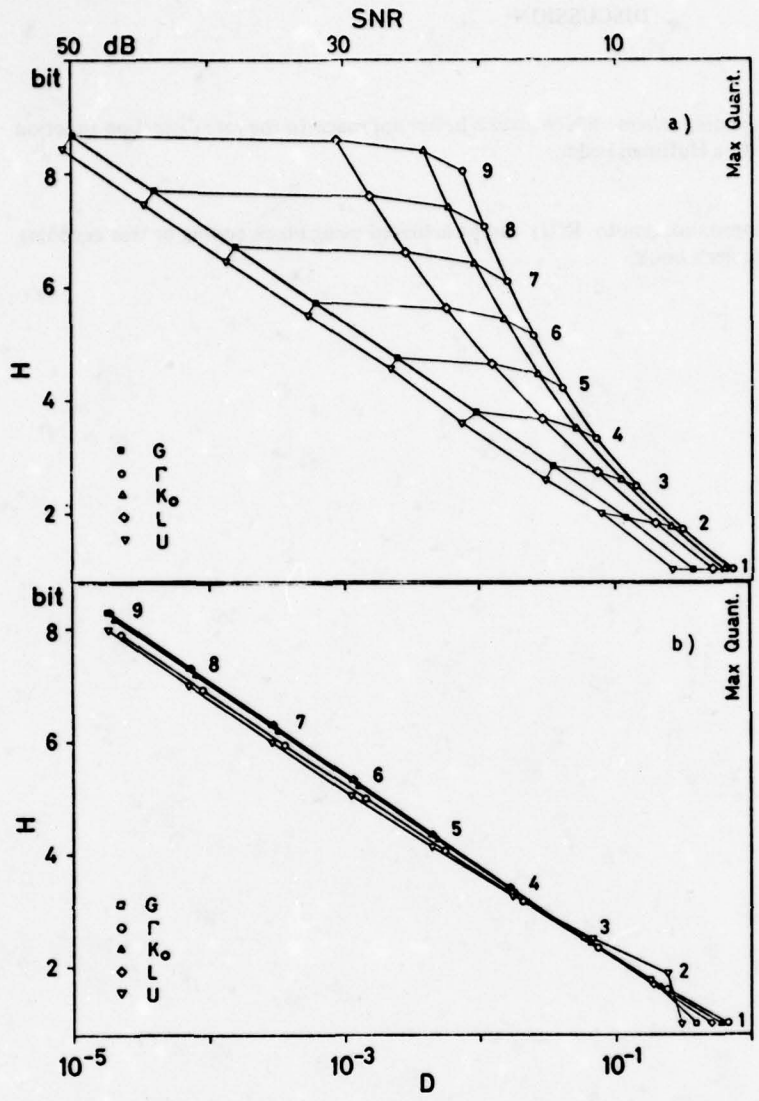


Fig. 3. Entropy of signals quantized in 2^n levels, n = 1...9, vs. distortion D. Quantizer matched to Gaussian pd (a) or K_0 pd (b)

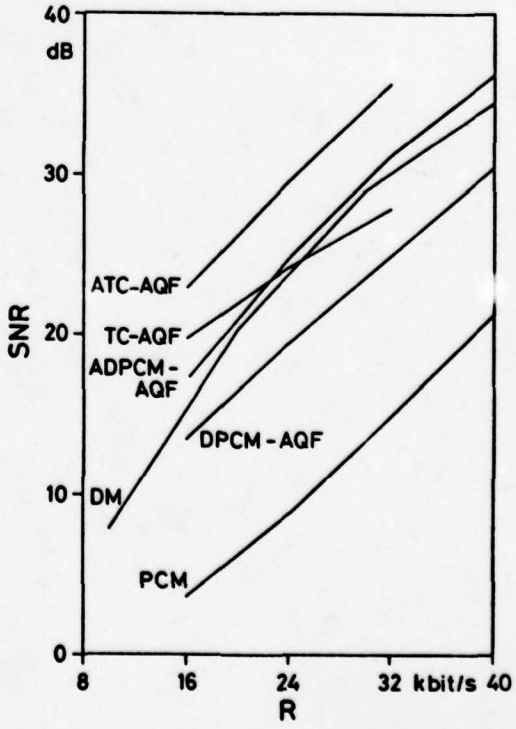


Fig. 4. SNR vs. bit rate R (bits/sec) for speech data compression schemes (as explained in text)

DISCUSSION

H.J.Matt, Ge

Is there any possibility to leave an encoding scheme which gives a better approach to the rate distortion function than the Max quantizer combined with a Huffman coder.

Author's Reply

Yes, it has been shown that a close approximation to R(D) can be achieved using block coding or tree encoding schemes. For further details see T.Berger's book.

PROBLEMS IN COMBINING SOURCE AND CHANNEL CODING

Hans Jürgen Matt
c/o Dr Wolf
Institut für Angewandte Physik der Johann Wolfgang Goethe Universität
Robert Mayer Str.
6000 Frankfurt
Germany

ABSTRACT

Shannon's theory about the noisy channel and coding of a discrete source with a fidelity criterion allows to separate the problem of information transmission into two independent disciplines: Source- and Channel Coding. For many practical problems however there remains the question of how could both coding schemes be adapted together and how much effort should be spent on the realization of the codecs. These questions will be discussed in this paper by the example of DPCM coded video signals.

1. INTRODUCTION

Shannon's noisy coding theorem (Shannon, C.E., 1949) states, that the information of a discrete stationary source with entropy H can be encoded for transmission over a discrete memoryless channel with capacity C at an arbitrarily small error probability (or an arbitrarily small equivocation) iff $H < C$.
Complementory to this theorem, Shannon's source coding theorem (Shannon, C.E., 1959) states, that the symbols of a discrete memoryless source can be encoded with respect to a fidelity criterion (i.e. with an average distortion D) at a minimum rate not less than R(D).
From these two theorems the information transmission theorem follows, concluding that the information of a discrete memoryless source can be reproduced with a certain fidelity (distortion D) at the output of any discrete memoryless channel of capacity C, provided $C > R(D)$.

The first of these three theorems imposes only one theoretical restriction on the source letters for reliable transmission over a disturbed channel, i.e. the entropy must remain below some specific value C given by the channel capacity. The channel encoding of the information to be transmitted can be considered independent from the properties of the source or the user. The second theorem refers to the properties of the source-user pair only, arguing that there exists a certain minimum rate R(D) which can by no means be reduced by which the information of a source can be represented with an average distortion D. This theorem establishes no further relation between the source and the transmission channel.
The third theorem concludes that there exists the possibility to transmit the information of a source with fidelity D over any channel with capacity C, iff $C > R(D)$ and that there is no way to do so iff $C < R(D)$.
Thus the fundamental theorems of information theory allow to separate the problem of reliable transmission of information into two independent disciplines:

SOURCE- and CHANNEL-CODING (Fig. 1)

It is the purpose of this paper, to show that for practical applic tions where no perfect coding is possible (e.g. due to finite word length) it looks advantageous to adapt both coding schemes to each other. Moreover there is a possibility for joined optimization of both coding schemes, if the "concept of a cost-distortion" function (similar to rate-distortion theory) is introduced.

2. SOURCE CODING SCHEMES

Source coding schemes involve methods for reducing redundancy and methods to reduce both, redundant and irrelevant information.
Since the majority of these schemes are based on discrete or digital information processing whereas many sources have an analogous output an A/D conversion is first necessary. The A/D conversion reduces the infinite entropy of the source signals to the finite entropy of the letters, thus removing some information. If the user accepts the quality offered by the letters then the removed information is said to be irrelevant.

Commonly the A/D conversion is realized in two steps; the analogous signal is sampled according to the Nyquist theorems and then the time discrete samples are quantized. In both steps a systematic error is introduced into the original signal, which for the sampler can be calculated to be

$$(2.1.) \qquad E_s(t) = f(t) - \sum_{i=-\infty}^{+\infty} h(t - iT)\, f(iT)$$

and for the quantizer to be

$$(2.2.) \qquad E_Q(t) = \sum_{i=-\infty}^{+\infty} h\,(t - iT)\,[f(iT) - q(iT)]$$

where f(t) denotes the original signal, f(iT) the samples, q(iT) the quantized samples and h(t) the impulse response of the reconstruction filter (Fig. 2). Both error signals can be made small enough at the expense of increasing bit rate and cost for implementation.
Max, J., 1960, gave an algorithm to minimize the mean square error of a quantizer for a given input signal probability distribution and a number N of quantizing levels. The mean square error is most commonly used as an easy to handle distortion measure, since more adequate distortion measures to source-user pairs are for many practical applications not available.

For source encoding of the digitized signals in order to eliminate redundancy a number of methods from Shannon, Fano, 1949, Huffman, D.A., 1952, Davisson, L.D., 1973, Ohnsorge, H., 1973, Ancheta, T.C., 1976, and run-length coding are available. The schemes of Shannon, Fano and Huffman make use of the known statistics of the source letters leading to variable-length comma-free codewords whereas the universal coding of Davisson and syndrome coding of Ohnsorge adapt to sources with unknown statistics. Ohnsorge's method for removing redundancy makes use of error correcting block codes. For that purpose the data sequences of a source are subdivided into blocks of length n. Each block is then regarded as the error pattern of a noisy channel and is completely determined by the syndrome vector of an appropriate error correcting code which is able to correct such patterns. The efficiency of this method increases with increasing block length, if suitable error correcting codes are available.
For further reduction of the source bit rate, coding methods for reducing information with respect to a given fidelity criterion need to be applied. Such methods may be Transformation Encoding (Anderson; Huang, 1971), Difference Pulse-Code-Modulation DPCM or methods using error correcting codes according to Berger, T., 1971, and Jelinek, F., 1969. The absolute bound for bit rate reduction is given by the rate distortion function, which represents a theoretical value (like channel capacity) unattainable for practical systems, since it would require infinite length of codewords to reach that bound even in the simple case of discrete memoryless sources, as has been shown by Berger.

2.1. Error sensitivity of source encoded signals

All the mentioned source coding schemes have in common an increased error sensitivity of the compressed data, since for data reduction always sequences of input signals need to be processed. If channel errors are introduced in compressed data, then the reconstructed sequences are affected as a whole. During reconstruction of the original signals transmission errors also may lead to limited or catastrophic error propagation.
As an example, let us consider the DPCM encoding schemes for video signals. Fig. 3 shows the block diagramm of a DPCM transmission system. The predictor is assumed to apply simple linear previous point prediction, i.e. $f(iT) = a(f((i-1)T) + q((i-1)T))$.

The output signal at the receiver is then given by

$$f(t) = \sum_{i=-\infty}^{+\infty} h(t-iT) \left\{ f(iT) + (q(iT) - e(iT)) + \sum_{k=0}^{\infty} a^k \left[d((i-k)T) - q((i-k)T) \right] \right\} \tag{2.3.}$$

where the term q(iT)-e(iT) indicates the error signal caused by the quantizer and the term $\sum_{k=0}^{\infty} a^k \left[d((i-k)T) - q((i-k)T) \right]$ represents the error signal resulting form channel errors (Essman, Wintz, 1973).

Unfortunately the channel errors may cause infinite error propagation if the prediction coefficient $a = 1$. For $a < 1$ the error propagation decreases exponentially with a^k. The case a = o corresponds to PCM which shows an error sensitivity much less compared to DPCM. The error propagation of DPCM systems depends much from the prediction factor a or the prediction scheme used. Usually the prediction scheme is optimized to obtain the highest possible picture quality assuming error free transmission. In these cases commonly a strong error propagation results.

The most effective way known up to now to reduce the error propagation in DPCM systems has been proposed by van Buul, 1976. The block diagramm of his Hybrid DPCM system is shown in Fig. 4. It operates as follows: At the DPCM coder some information about the predictor (e.g. the 4 most significant bits) is added to the data and similarly subtracted at the decoder. Since the DPCM-values are frequently very small or even zero the information transmitted during these times comes from the predictor. If an error occured during transmission, the predictor of the decoder contains some other value than the predictor of the coder. If the next DPCM-values are then transmitted correctly the predictor of the decoder will be reset approximately (e.g. in the 4 MSB) to the correct value of the coder, thus eliminating catastrophic error propagation.
Now we shall try to investigate the question what influence the parameters of a compression scheme do have on the distortion caused by transmission errors.

2.2. Distortion caused by transmission errors

The distortion caused by transmission errors depends from the characteristics of the quantizer, from the properties of mapping the representative values on to codewords, from the input probability density function and the distortion measure given by the user. Let us define a distortion measure d which assigns the distortion $d(y_k;y_i)$ to the pair $(y_k;y_i)$ of transmitted and received values. Let us further assume the probability for the letter y_k to be sent to be $P(y_k)$, the probability for e errors introduced by the channel on y_k to be $P(e)$ and the probability to receive y_i on the condition of y_k and e to be $P(y_i/y_k;e)$. Then an averadge distortion $\bar{d}$ can be calculated

$$(2.4.) \qquad \bar{d} = \sum_i \sum_k \sum_e P(y_k)\ P(e)\ P(y_i/y_k;e)\ d(y_i;y_k)$$

Some special cases of this formula will be discussed:

1. Yamaguchi and Huang, 1966, derived formulas for the averadge distortion $\bar{d}$ for linear PCM assuming the letter probability $P(y_k)$ = const., a linear quantization $/y_{i+1}-y_i/$ = const., an error probability p of a memoryless channel and a distortion measure $d(y_i;y_k) = (i-k)^2$. If the letters y_k are encoded with the n-bit long Dual code

$$(2.5.) \qquad \bar{d}_{Dual}(n) = \frac{1}{3}\ (4^n-1)p$$

if the coding is done with the n-bit long Gray code

$$(2.6.) \qquad \bar{d}_{Gray}(n) = \frac{1}{6}\ (4^n-1) - (\tfrac{1}{2}-p)\ \frac{4^n-(1-2p)^n}{4-(1-2p)}$$

Since $\bar{d}_{Gray}(n) - \bar{d}_{Dual}(n) = (1-2p)\ d_{Gray}(n-1) > 0$

is a positive number the Dual code causes less distortion compared with the Gray Code in that case.

2. Next we assume only one error to be occurred in position x^j in any letter y_k. We then may define an averadge distortion d(j) for one error occurring in position x^j.

$$(2.7.) \qquad d(j) = \sum_i \sum_k P(y_k)\ P(y_i/y_k;x^j)\ d(y_i;y_k)$$

For a linear quantizer (e.g. in PCM systems) we find always an exponential increase of d(j) with j , whether the absolut difference between a pair of letters or its square has been used for a distortion measure. Tab. 1.

For a nonlinear quantizer (e.g. in DPCM systems) some additional assumptions are necessary to get equivalent results: The input signal probability distribution is assumed to be a Laplace distribution

$$(2.8.) \qquad p(x) = \ln 10 \cdot 10^{-2|x|}$$

and the quantizing characteristic is according to Tab. 2 (Arguello et al., 1971). For the 4 bit Dual- and Gray-Code we then get the results shown in Tab. 3. The values of Tab. 3 show the same tendency as Tab. 1 though we cannot establish simple functions for d(j) in this case. It is remarkable that the Gray Code now causes less distortion on the averadge compared with the Dual Code.

2.3. Methods to reduce the effects of transmission errors on compressed video signals

Some methods for reducing the effects of transmission errors will be outlined now by the example of compressed video signals. (For comparison of these methods by subjective tests it is important to compare always the same kind of video equipment, since the visibility of transmission errors in video signals depends heavily from whether single picture- or TV-transmission is used. A photograph of a picture effected by some errors gives a visual impression of distortion which corresponds to the distortion given by a real time TV transmission with a several orders of magnitude lower error rate.)

1. Optimizing predictors with respect to channel errors:
   In DPCM systems transmission errors cause error patterns which depend from the prediction scheme used. Among several predictors with comparable picture quality those with the least visible distortion for the human eye on a single error will be preferred.
2. Limiting error propagation:
   Error propagation of DPCM codecs may be limited using some kind of leaky prediction (e.g. $a < 1$) or van Buul's Hybrid-DPCM scheme.

3. Detecting errors with the aid of remaining redundancy:
   Since DPCM systems do not remove the picture redundancy perfectly, the remaining picture redundancy may be used to detect transmission errors as has been proposed by Lippmann, R., 1973.

4. Reconstructing wrong picture elements (pels):
   Moreover,the remaining picture redundancy can be used to reduce the visibility of detected errors by reconstructing the disturbed picture elements (pels). This may be accomplished with interpolation or substitution of the wrong pels by correct pels in the neighbourhood or in the previously transmitted picture.

All these methods help to reduce the influence of transmission errors at a very limited degree; that is they are useful if the error rate is below some threshold ($p \sim 10^{-6} \ldots 10^{-7}$). For higher channel error rates it is therefore necessary to add systematic code redundancy which can be used most effectively if an adaption strategy for source- and channel codecs is applied. We propose the following strategy:

1. Low weight error patterns ($w \leqslant t$) occurring frequently should be eliminated by the channel codec.
   This matches particularly with the properties of block codes, which are able to correct up to t errors in a block of length n with Hamming distance $d > 2t$.

2. High weight error patterns ($w > t$) occurring less frequently should be detected by the channel codec, since block codes are able to detect all errors with weight $t < w \leqslant d-t-1$ and many more error patterns with high probability.
   If n,t and d are properly chosen the majority of all errors will be eliminated by the channel decoder and the remaining uncorrectable error patterns will be detected.

3. Uncorrectable error patterns should be indicated to the source decoder. The source decoder can then reconstruct the wrong picture elements.

4. The channel coding scheme should be adapted to the different bit-error sensitivity d(j) of the source letters.
   In order to achieve a position independent distortion the channel coding scheme should have a position dependent residual error probability $P_r(j)$ satisfying the equation

$$P_r(j) = \frac{\text{const.}}{d(j)} \qquad (2.9.)$$

   One way to get an approximation to equ. (2.9.) is to use interleaved channel coding with an individual code for every position x^j. However this would require individual codecs and seems not to be an economical solution.

3. A CONCEPT FOR A CHANNEL CODEC

An example for the protection of compressed video signals on modern communication channels (e.g. satellite-, PCM- or optical fibre links) will now be discussed. It is based on the considerations of chapter 2. and the following assumptions to match approximately the requirements of a DPCM codec and a given type of transmission channel:

1. The transmission channels are assumed to have a bit error rate $p \leqslant 10^{-4}$.

2. The systematic code redundancy should be fairly small, i.e. a few percent are regarded to be sufficient.

3. The block length n of the systematic code is chosen to take up a whole line of the TV system.

4. The channel codec should be able to correct random errors as well as bursts. Uncorrectable error patterns have to be detected with high probability and indicated to the source decoder.

5. The source codec (DPCM) delivers a constant number k of bits per line.

6. If a line is received uncorrectable the source decoder makes use of line substitution.

7. The amount of hardware for the codecs and the number of clocks required to correct all errors should be minimum.

These requirements lead to the concept of a tri-state channel codec using BCH block codes (Matt, H.J., 1978).
To calculate the gain of the proposed channel codec we use a (n = 1023, k = 992, d = 8) BCH-Code able to correct either random errors or one burst error up to 12 bits, as has been shown by Matt, 1978. Though the Hamming distance d = 8 of the code would allow a triple error correction, we correct only up to 2 errors per block to have sufficient error detection capability. The residual block error probabilities have been computed (Fig. 5) under a worst case assumption that the channel is memorlyless with an error probability p and that the decoder starts with burst error correction, because he gets no additional information form outside whether a burst pattern has occurred or not. If there is no correctable burst the decoder continues its search with random error correction. P_{CB} denotes the probability of an erroneous block, P_u the probability of uncorrectable blocks and P_B the probability of undetected erroneous blocks.

If we transform the result of Fig 5. on the parameters of standard TV (625 lines, 25 pictures per second) we find, that an error rate $p = 2 \cdot 10^{-5}$ without channel coding causes 1 among 66 TV lines to be disturbed on the averadge ($\hat{=}$ 10 lines per picture $\hat{=}$ 250 lines per second). By adding 3 % of systematic code redundancy we get an improvement of 1 among 10^6 lines to be uncorrectable by the channel decoder ($\hat{=}$ 1 line per minute). Since the source decoder gets knowledge of these lines he may substitute them by previous lines. Only 1 ‰ of the uncorrectable lines will not be detected by the channel decoder which causes 1 among 10^9 lines to be wrong, an event which occurs once within 18 hours.

The block diagramm of the channel decoder is given in Fig. 6. It shows three main modules:

1. A bufferstorage to store the received data block,
2. a burst error decoder including an overall syndromeregister with feedback connections (corresponding to the generator polynomial g(x) of the BCH Code) and equipped with a zero test logic for burst error trapping,
3. a random error decoder for algebraic decoding of 2 errors, including a set of registers to calculate the partial syndromes S_1, S_1^3, S_3 and the coefficients σ_1, σ_2, of the error locator polynomial and the Chien searcher.

The modules operate as follows: When a data block enters the decoder all syndromes are calculated simultaneously. The burst error decoder then tries to find a correctable burst pattern meanwhile the random error decoder calculates the coefficients σ_i. If a correctable burst pattern was found it will be subtracted from the data block when it leaves the output. If no correctable burst pattern was found the random error decoder corrects the data bits indicated by the Chien searcher. In this case the decoded random error pattern is fedback to the syndromeregister of the burst decoder to check if the correction was successful. The correction was successful if the syndromeregister contains only zeros. By this final check the uncorrectable error patterns are detected.
Thus the channel decoder operates in a tri-state mode: It corrects bursts or random errors and detects most of the uncorrectable error patterns.

4. JOINED OPTIMIZATION OF SOURCE- AND CHANNEL CODING

We now consider the question whether there is any criterion for joined optimization of the source- and channel coding schemes. The theorems of information theory led to a separation of the information transmission problem into independent disciplines.
Ohnsorge, 1977, first pointed out that the economical gain of any source- or channel codec is limited, since there is always a point from which an increase of data-compression or protection becomes more expensive than the supply of additional channel capacity.
We now extend this idea to the more general problem, whether it is possible to determine the parameters of a whole transmission system from an economical viewpoint, more precisely from some kind of cost-effectiveness calculation.
An information transmission system is characterized by two main factors, its cost and the signal quality or distortion D produced at the receiver. We therefore try to find a set of cost functions for each component of the transmission system:

1. The source coder and decoder is assumed to have a cost-function K_1 depending form the input- (IR_1) and output rate (OR_1) of the coder and from the distortion D_1 introduced by it.

(4.1.) $$K_1 = f_1(D_1, IR_1, OR_1)$$

2. Analogously the cost-function K_2 of a channel codec may be calculated depending from its input rate IR_2, the systematic code redundancy R_c, the channel error rate p and the residual error probability p_r; the latter can be expressed by an equivalent distortion D_2 depending from the error sensitivity of the source codec.

(4.2.) $$K_2 = f_2(p_r, IR_2, R_c, p) \quad \text{with } IR_2 = OR_1$$

(4.3.) $$D_2 = \varphi_2(p_r, \bar{d})$$

3. A cost-function K_3 may be calculated for the special part of the source decoder which enables it to reconstruct picture elements by interpolation or substitution. This function depends from the uneliminated redundancy R_s of the source signals and the amount of (negative) distortion D_3, due to signal quality improvements.

(4.4.) $$K_3 = f_3(D_3, R_s)$$

4. The cost-function K_4 of the channel is assumed to be a function of the channel rate R and the error probability p.

(4.5.) $$K_4 = f_4(R,p) \quad \text{with} \quad R = \frac{IR_2}{1-R_c}$$

All cost-functions represent costs per time unit. If we are successful in finding the desired cost functions or an approximation for them we then may try to solve a variational problem similar to the rate-distortion problem. We therefore call a solution of this variational problem a cost-distortion function.
The cost-distortion function $K(D^+)$ is now defined to be the minimum of the whole transmission system cost subject to the constraint that the total distortion D_t does not exceed a given limit D^+ and that other boundary conditions are satisfied:

$$(4.6.) \qquad K(D^+) = \min_{D_t \leq D^+} \sum_{i=1} K_i$$

$$(4.7.) \qquad D_t = \varphi_1(D_1, D_2, D_3) \rightarrow \phi_1 = \varphi_1(D_1, D_2, D_3) - D_t$$

$$(4.8) \qquad D_2 = \varphi_2(p_r, \bar{d}) \rightarrow \phi_2 = \varphi_2(p_r, \bar{d}) - D_2$$

. . .
. special constraints .

This variational problem can be solved by Lagrange's multiplier method, forming an augmented function J with the parameters $s_1, s_2 \ldots$; one parameter s_k for each ϕ_k.

$$(4.9.) \qquad J = \sum_{i=1}^{4} K_i + s_1 \phi_1 + s_2 \phi_2 + \ldots$$

Then partial differentiation of J yields a set of equations from which all the parameters $D_k, s_k, \ldots$ $(k = 1,2,3\ldots)$ can be determined. Thus from the solution of this variational problem we find a COST-DISTORTION FUNCTION which for a given distortion D^+ specifys the total system cost and all other parameters of the transmission system components. Among these parameters we find particulary the distortion (D_1, D_2, D_3) for the different codecs used in the system.

As an example we consider the following set of simplified cost functions:

$$(4.10.) \qquad K_1 = \frac{a_1}{D_1} \quad ; \qquad 0 < D_1 < \infty$$

$$K_2 = \frac{a_2}{D_2} \quad ; \qquad 0 < D_2 \leq b_2$$

$$K_3 = \frac{a_3}{D_3 + b_3} - \frac{a_3}{b_3} \quad ; \qquad -b_3 < D_3 \leq 0$$

$$K_4 = a_4$$

and assume

$$(4.11.) \qquad D_t = \sum_{j=1}^{3} D_j \rightarrow \phi_1 = \sum_{j=1}^{3} D_j - D^+$$

then we form

$$(4.12.) \qquad J = \sum_{i=1}^{4} K_i + s_1 \phi_1$$

$$(4.13.) \qquad \frac{\partial J}{\partial D_1} = 0 \; ; \; \frac{\partial J}{\partial D_2} = 0 \; ; \; \frac{\partial J}{\partial D_3} = 0$$

and find after some manipulations the desired results:

$$(4.14.) \qquad D_1 = (D^+ + b_3) \frac{\sqrt{a_1}}{\sum_{i=1}^{3} \sqrt{a_i}}$$

$$D_2 = (D^+ + b_3) \frac{\sqrt{a_2}}{\sum_{i=1}^{3} \sqrt{a_i}}$$

$$D_3 = -b_3 + (D^+ + b_3) \frac{\sqrt{a_3}}{\sum_{i=1}^{3} \sqrt{a_i}}$$

$$(4.15.) \qquad K(D^+) = \frac{\left(\sum_{i=1}^{3} \sqrt{a_i}\right)^2}{D_t + b_3} - \frac{a_3}{b_3} + a_4$$

Since the equations (4.13.) are only a necessary condition for a maximum or minimum of $K(D^+)$ we must check $K(D^+)$ to be a true minimum. Also if for a certain value of D^+ one or more D_i are out of range the corresponding boundary values of D_i must be inserted into equ. (4.10.) and the variation calculus repeated.

5. SUMMARY

In this paper some problems in combining source- and channel coding are discussed. By the example of DPCM coded video signals the effects of channel errors were analysed to see how these effects could be minimized. An adaption strategy for source- and channel codecs is proposed based on the fact that channel decoders are particulary suited to correct small numbers of errors occurring with high probability so that greater numbers of errors occurring less frequently are left to the source decoder for reconstruction. For protection of the compressed signals a low cost tri-state channel codec is proposed and its performance is outlined. Finally the problem of optimizing a transmission system as a whole is shown to be a variational problem relating the cost of the system to the signal quality (distortion) produced at the receiver. Its solution leads to a cost-distortion function which may include all relevant parameters of the transmission system's components.

ACKNOWLEDGEMENT

The author would like to thank his former collegues at AEG-Telefunken Dr. Haller, Dr. Ohnsorge, M. Prögler and Prof. Dr. Wendland for useful discussions and encouragement and Mr. Schäffner for his extensive work on computer programming.

6. REFERENCES

ANDERSON, G.B.; HUANG, T.S., 1971: Piecewise Fourier transformation for picture bandwidth compression. IEEE Trans. COM-19, pp. 133-140

ANCHETA, T.C. Jr., 1976: Syndrome source coding and its universal generalization. IEEE Trans. IT-22, no 4, pp. 432-436

ARGUELLO et al., 1971: The effect of channel errors in the DPCM transmission of sampled imagery. IEEE Trans. COM-19, no 6, pp. 926-933

BERGER, T., 1971: Rate distortion theory. Prentice-Hall, Inc. Englewood Cliffs, New Jersey

van BUUL, M.C.W., 1976: Hybrid D-PCM, a combination of PCM and DPCM. Intern. Pict. Coding Symp., Asilomar, Calif., USA

DAVISSON, L.D., 1973: Universal noiseless coding. IEEE Trans. IT-19, pp. 783-795

ESSMAN, J.E.; WINTZ, P.A., 1973: The effects of channel errors in DPCM systems and comparison with PCM systems. IEEE Trans. COM-21, vol 8, pp. 867-877

HUFFMAN, D.A., 1952: A method for the construction of minimum redundancy codes. Proc. IRE., vol 40, pp.1098-1101

JELINEK,F., 1969: Tree encoding of memoryless time discrete source with fidelity criterion. IEEE Trans. IT-15, pp. 584-590

LIPPMANN, R., 1973: A technique for channel error correction in DPCM picture transmission. Proc. IEEE Conf. on Comm., Seattle, Washington

MATT, H.J., 1978: On burst error correction ability of cyclic block codes. NTG-Intern. Conf. on information theory and system theory in digital communications.W.Berlin

- Zur Optimierung der Kanalcodierung und Adaption an quellcodierte Signale. Submitted to Technical University of Hannover

- Verfahren zur gesicherten Datenübertragung DPS 24 45 508/1974

MAX, J., 1960: Quantizing for minimum distortion. IRE Trans. IT-6, vol 1, pp. 7-12

OHNSORGE, H., 1973: Data compression system for the transmission of digitizing signals. IEEE Int. Conf. on Comm., Seattle, Washington

- 1977: Economical gain of source and channel encoding. Frenquenz 31, pp. 270-274

SHANNON, C.E., 1949: A mathematical theory of communication. Univ. of Illinois Press, Urbana

- 1959: Coding theorems for a discrete source with a fidelity criterion. IRE National Conv.Rec., part 4, pp. 142-163

SHANNON, FANO,R.M., 1949: The transmission of information. Technical report no 65, Research Lab. of Electronics, MIT, Cambridge, Mass.

YAMAGUCHI, Y., HUANG,T.S., 1965: Optimum binary code. MIT Quarterly Progress Report, no 78, July 15, no 79,Oct. 15

- 1966: Report no 82, July 1966

Tab.1 Distortion d(j) caused by transmission errors, depending from error position x^j in a linear PCM codec.

$d(y_i ; y_k)$	$y_{i+1} - y_i = \text{const.}$	
	Dual Code	Gray Code ; $P(y_k) = \text{const.}$
$\lvert i - k \rvert$	$d(j) = 2^j$ $\sigma(j) = 0$	$d(j) = 2^j$ $\sigma^2(j) = \frac{1}{3}(4^j - 1)$
$\lvert i - k \rvert^2$	$d(j) = 2^{2j} = 4^j$ $\sigma(j) = 0$	$d(j) = \frac{1}{3}(4^{j+1} - 1)$ $\sigma^2(j) = \frac{16}{15}(4^{j+1} - 1)(4^j - 1)$

Tab.2 Quantizer characteristic (Arguello et al.)

y_k	x_k	Codeword Nr	$P(y_k)$
- 103		1	0,010939
	- 83		
- 63		2	0,039061
	- 50		
- 37		3	0,075594
	- 30		
- 23		4	0,092664
	- 18		
- 13		5	0,097221
	- 10		
- 7		6	0,081685
	- 5		
- 3		7	0,058841
	- 2		
- 1		8	0,043995
	0		
1		9	0,043995
	2		
3		10	0,058841
	5		
7		11	0,081685
	10		
13		12	0,097221
	18		
23		13	0,092664
	30		
37		14	0,075594
	50		
63		15	0,039061
	83		
103		16	0,010939

Tab.3 Distortion d(j) caused by transmission errors, depending from error position xj in a nonlinear DPCM codec.

$p(x)=\ln 10 - 10^{-2\|x\|}$	4 Bit Dual Code		4 Bit Gray Code	
d(j)	$d(.)=\|y_l - y_k\|$	$d(.)=(y_i - y_k)^2$	$d(.)=\|y_i - y_k\|$	$d(.)=(y_i - y_k)^2$
d(0)	11,27	240	11,27	240
d(1)	26,59	1216	27,05	1526
d(2)	48,15	2953	36,19	2169
d(3)	51,86	3142	42,29	3557
d_{max}	104	10816	206	42436
$\bar{d}$	34,47	1888	29,2	1873

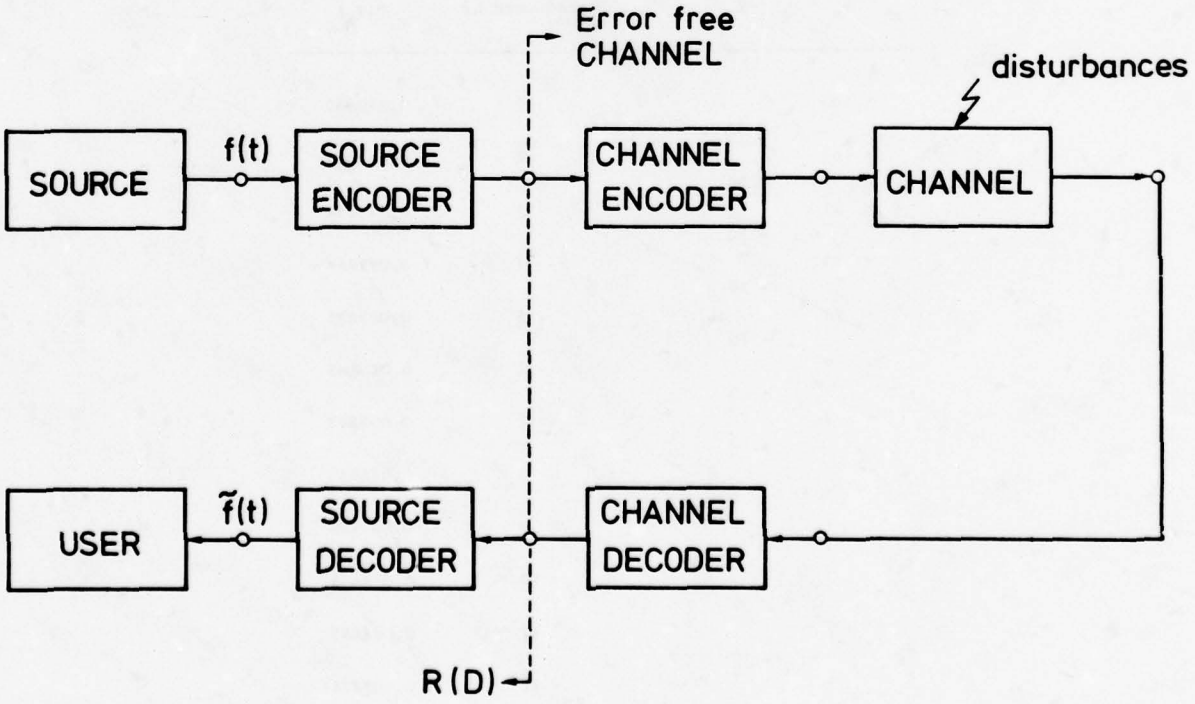


Fig.1 Blockdiagram for reliable information transmission using Source- and Channel-Coding

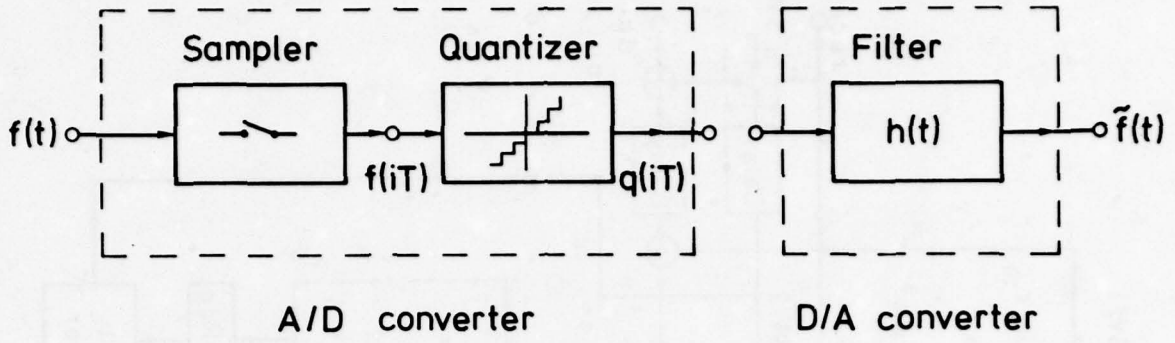


Fig.2 Blockdiagram of A/D - D/A conversion

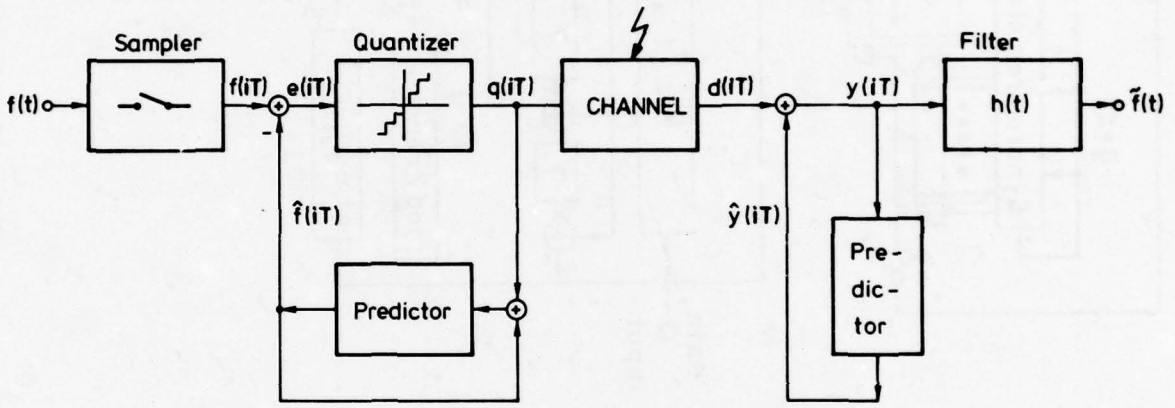


Fig.3 Blockdiagram of a DPCM transmission system

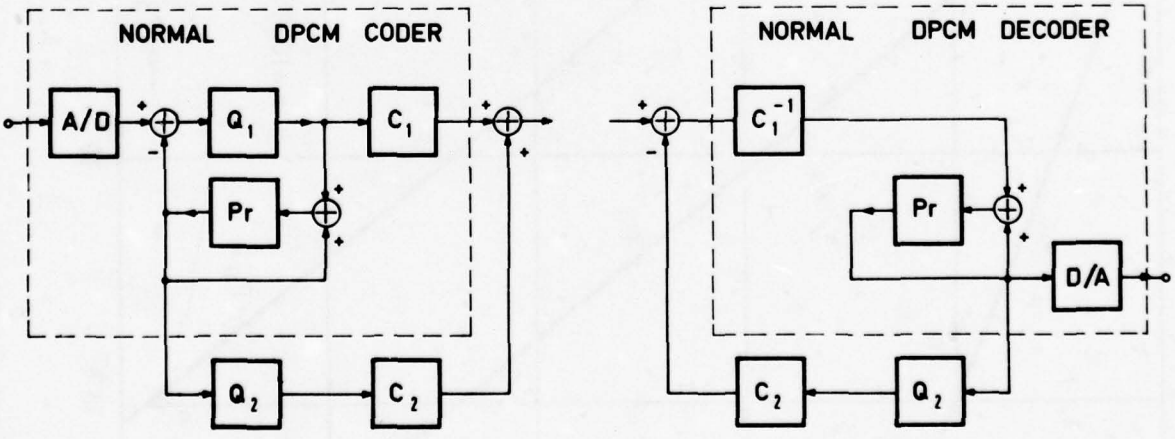


Fig.4 Blockdiagram of HYBRID-DPCM codec from van Buul

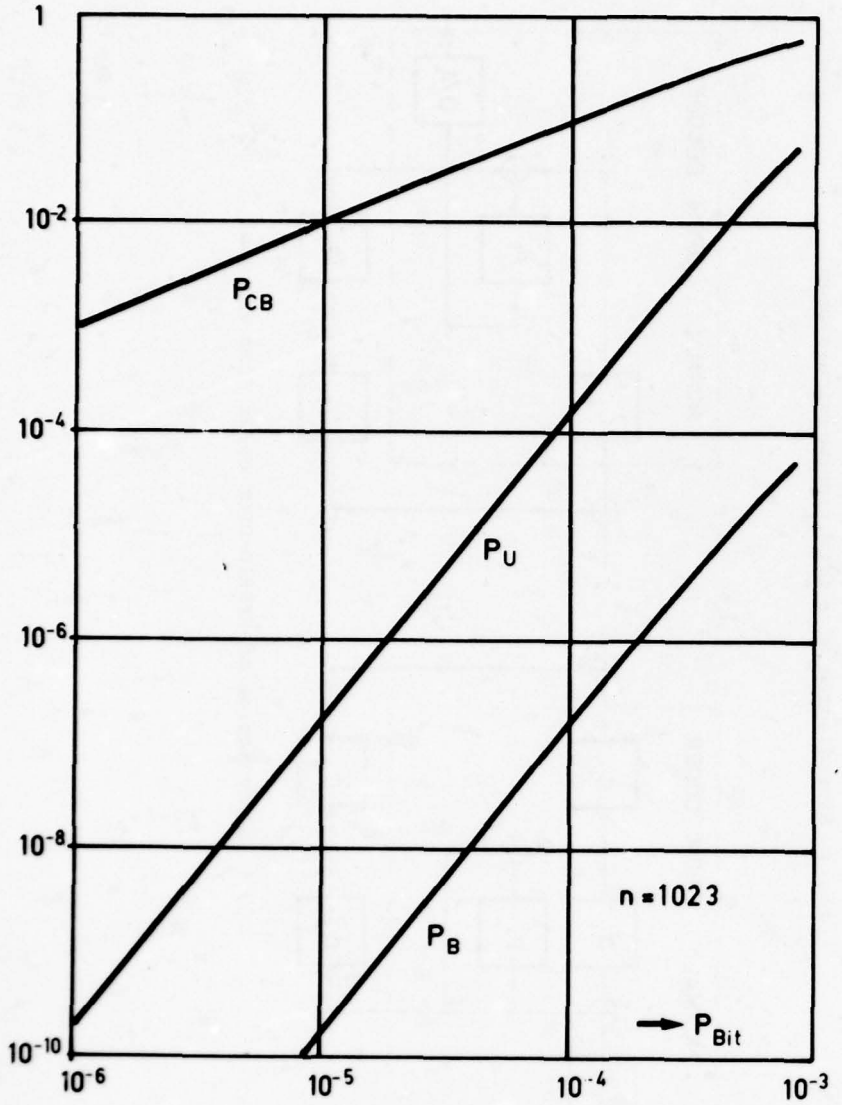


Fig.5 Residual block error probabilities

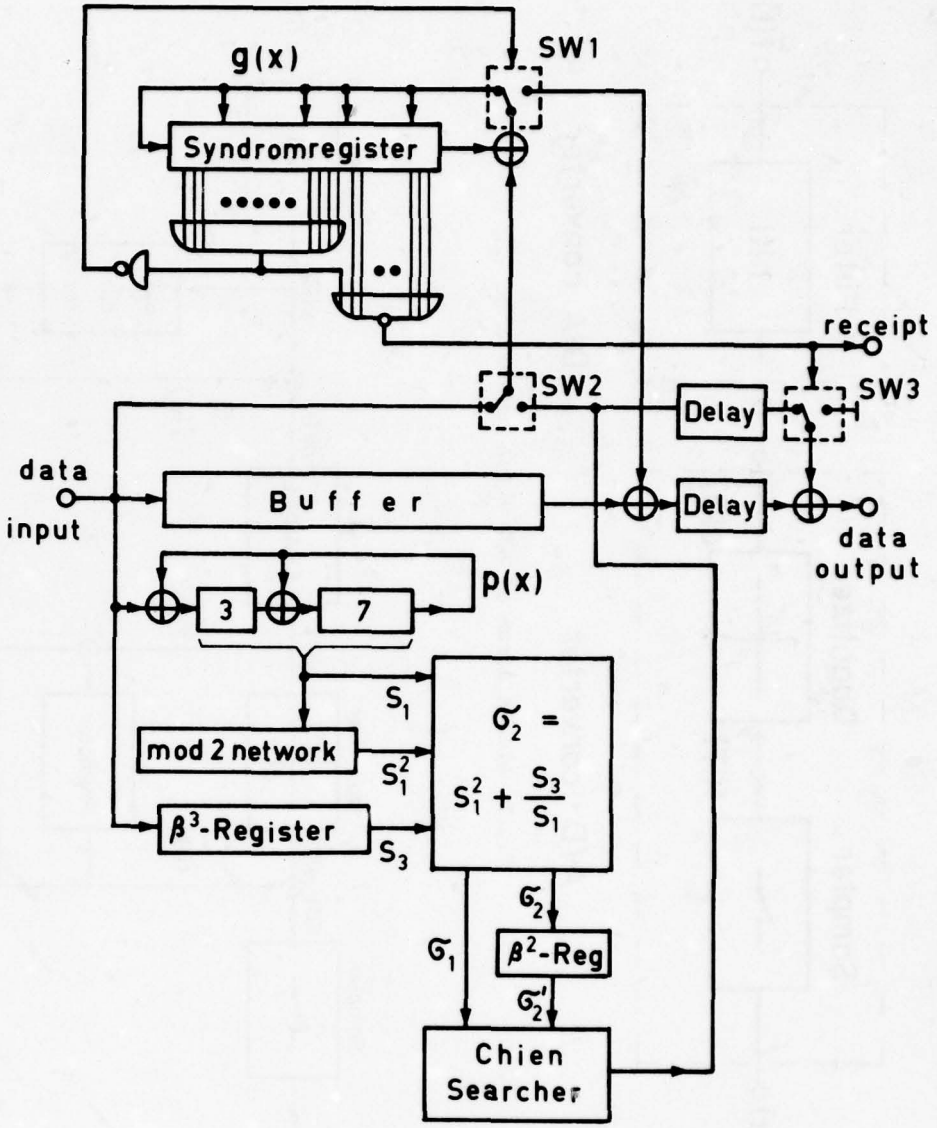


Fig.6 Blockdiagram of tri-state channel codec

SEGMENTATION OF PICTURES INTO CHANGING AND MOVING PARTS FOR FRAME REPLENISHMENT CODING TECHNIQUES

Jürgen Klie
Lehrstuhl für Theoretische Nachrichtentechnik
und Informationsverarbeitung
Technische Universität Hannover
Callinstraße 32
D-3000 Hannover

SUMMARY

A frame replenishment coder takes advantage of similarity between successive frames of a television system in two ways for reducing the transmission bit rate:

(a) Only the parts of the picture which change their information from frame to frame are transmitted, the unchanged parts can be reconstructed at the receiver by repeating from a frame memory.

(b) The changing parts can be coded with different spatial resolutions depending on subjective requirements by varying the sampling frequency.

To make use of these advantages it is necessary to segment the picture in changing and unchanging parts. This is done by comparing the average absolute amplitude of five adjacent frame-to-frame differences to a certain threshold. The threshold is found by subjective tests.

A changing part within a picture consists normally of a moving object and background which is uncovered by the moving object. Due to the temporal low-pass characteristic of the human eye, it is allowed to encode a moving object with a reduced horizontal and vertical resolution but not the uncovered background. Therefore it is useful to distinguish within a changing part between moving object and background. It will be shown that this segmentation can be done by temporal filtering of the video signal and comparing the output of the filter with the unfiltered signal. The difference of these two signals is unequal zero in moving parts of the frame.

1. INTRODUCTION

In a video telephone or surveillance system cameras are stationary. Scenes consist mainly of relatively small objects moving in front of a stationary background. A frame out of such a TV system can therefore be segmented for the case of a moving object in different parts. Fig. 1 shows a ball which is moving in front of a background. The ball was in the position marked by the dashed lines one frame before and has moved to its new position in the actual frame.

The actual frame can be segmented into the stationary background area which is unchanged compared to the previous frame, into the moving object itself and in an area of uncovered background which was covered by the ball one frame before. Both parts, the moving object and the uncovered background, have changed their information compared to the previous frame. Together they can be called to be the changing area. A frame replenishment coder takes advantage of this segmentation in two ways for the reduction of the transmission bit rate.

Only those parts of the picture which change their information from frame to frame, namely the areas of uncovered background and moving objects, are transmitted. The unchanged parts can be reconstructed at the receiver by repetition from a frame memory. This leads to a considerable saving in transmission bit rate. The required bit rate is determined mainly by the amount of changing picture elements. The segmentation in changing and unchanging areas is done with the help of a change detector. The input information for a change detector consists of the frame-to-frame differences of a digitized video signal (Mounts, F.W. 1969). The simplest method leading to an ideal segmentation would be to consider each picture element as a changing one, if its difference from the corresponding picture element in the previous frame is not zero.

This kind of ideal detector works (measured by objective standards) without any fault, but it can not distinguish between frame-to-frame differences due to a change of relevant information and those due to noise. Some complicated algorithms have been developed to find these parts of a frame (Conner, D.J. 1973; Limb, J.O. 1976).

Since the human observer is the receiver of the transmitted pictural information it is better to develop a subjectively optimized change detector, which detects only those changing parts of a picture, which change their information perceptible by the human observer. The task of a change detector which is optimal in the reduction of the transmission bit rate in conditional replenishment coders is therefore the segmentation of a frame into changing and unchanging parts with no percectible degradation in picture quality and with the lowest required bit rate. The development of the optimal change detector is based on an existing algorithm. This existing algorithm uses for the decision on change detection three different thresholds applied to frame-to-frame differences. The

algorithm is optimized by the subjective criterion of showing no visible difference between the processed and unprocessed signal and under the additional constraint of having a minimal amount of detected picture elements. The result is a change detector optimized by subjective criteria which reduces the transmission bit rate by about 30 % as compared to the ideal change detector mentioned above. The algorithm for change detection is based on the comparison of the absolute value of five adjacent frame-to-frame differences within a line to a threshold of 1.2 out of 256.

A changing part within a picture consists usually of a moving object and a background which is uncovered by the moving object (Fig. 1). Due to the temporal low pass characteristic of the human eye it is often allowed to encode a moving object with a reduced spatial resolution but this is not permitted for the uncovered background. If a reduced spatial resolution should be used for saving transmission bit rate it is necessary to distinguish within a changing part of a TV frame between moving object and uncovered background. For this purpose the signal is first filtered by a temporal low pass filter. Calculations show that a temporal filter effects not only the temporal bandwidth of the signal but also the spatial bandwidth. This happens only to moving objects. The spatial bandwidth of the unchanged parts and of the uncovered background is not influenced.

The spatial low pass effect of the temporal filter within moving parts of a frame can also be recognized in a changing of the amplitude of the picture elements which belong to the moving object. Therefore the moving object can be detected by comparing the output of the temporal filter within the unfiltered signal at the same time instant. The difference of these two signals is unequal zero only in moving parts of the frame.

2. CHANGE DETECTOR

2.1 Basic Idea

In a video telephone system or in surveillance systems the cameras are stationary. Relatively small objects are moving in front of a stationary background. Since the moving objects do usually not cover the whole field of vision, at least a part of the picture will not change during a longer period. A considerable saving in transmission bit rate can be achieved by segmenting the picture in changing and unchanging parts. Only those parts which have changed their information are transmitted. The stationary parts of the picture are repeated at the receiver from a frame memory. A change detector makes its decision on the basis of the frame-to-frame differences. The simplest way to detect the changing parts of a frame is to check the difference between each picture element of the present frame and the corresponding picture element of the previous frame. Whenever a difference is unequal zero, the picture element of the present frame which belongs to this difference is considered to be a changed one.

This kind of change detector works without fault measured by objective criteria and can be called an ideal detector but it is not the optimum with respect to a bit rate reduction judged by subjective criteria.

As in all visual communication systems the human observer is the receiver of the transmitted information. Therefore it is useful to transmit only those parts of each frame whose information change is visible to the observer. A change detector for conditional replenishment coding techniques can be regarded as being optimal judged by subjective criteria, if all parts of the frame, whose information change is visible, are detected while the amount of detected picture elements is minimal. Besides this subjective criterion one must pay attention also to the fact that in conditional replenishment systems positional information must be added to the visual information. The receiver is able to insert the transmitted information into the previous frame into the correct place only by using addressing techniques. Contiguous areas of detected picture elements within a line of a frame are usually called clusters (Candy, J.C. 1971). All the picture elements belonging to a cluster are transmitted with a common address. A change detector which produces on the average longer clusters by the same total number of detected picture elements is therefore preferable to others with respect to the aim of a high bit rate reduction.

2.2 Proposal for a Change Detector

The existing algorithm which has been mentioned above is shown in Fig. 2. It is characterized by the three thresholds T_1, T_2 and T_3 based on which a decision of change detection is made.

2.2.1 Coarse Threshold T_1

A coarse decision for each picture element is made in order to recognize significant changes, which lead to large frame-to-frame differences. The magnitude of each frame-to-frame difference is compared to the threshold T_1. If the threshold is exceeded, the picture element is considered as belonging to a changing area. The threshold should be approximately 1.5 percent of the maximum. In the case of an amplitude quantization with 256 levels the threshold is 4 out of 256.

2.2.2 Sign Decision

If the magnitude of the frame-to-frame differences is below the coarse threshold a fine decision is made in order to distinguish between a perceptible change in picture infor-

mation and an unperceptible change which may be caused by camera noise, small displacements of objects (less than one picture element per frame) or slight changes in lighting. Complete picture areas which are changed by the movement of an object have frame-to-frame differences with the same sign. This is demonstrated in Fig. 3. Noise for example, which is uncorrelated with the signal, leads to frame-to-frame differences where the signs tend to alternate. Therefore a second criterion for the visibility of a change in information has been used which checks the signs of several adjacent frame-to-frame differences within a line and determines the sum of identical signs. This method does of course not consider zero differences. If the sum of identical signs exceeds a certain threshold the picture element in the middle of the block is regarded to be a changing one. However, to keep the bit rate low, the average absolute amplitude of the frame-to-frame differences must exceed a threshold T_2.

The number of adjacent frame-to-frame differences used for the sign decision should be smaller than the average length of a contiguous changing area (cluster). Therefore five adjacent differences have been taken. Since it has been assumed that a frame-to-frame difference of two out of 256 is not noticable to a human observer the threshold T_2 has been chosen to be 2/256.

2.2.3 Threshold T_3

Quasiperiodic textural structures like a persons hair or a fence in a landscape lead also to frame-to-frame differences with alternating signs similar to noise. Therefore the sign decision is completed by a third thresholding. If the average absolute value of the five frame-to-frame differences exceeds threshold T_3, the corresponding picture element is regarded to be a changing one without looking at the sign decision. The magnitude of T_3 is determined by the fact that the change of being exceeded by noise is almost zero.

Assuming a camera signal to noise ratio of 45 dB, which is rather realistic, and a Gaussian distribution the probability that a noise amplitude will be higher than a certain threshold is shown in Table 1.

There is a considerable difference in the probability between a threshold of two and three. Therefore a threshold of $T_3 = 3/256$ provides a clear distinction between noise and a low amplitude and visible change in picture information.

2.3 Optimization of the Detector Thresholds

The proposed principle of a change detection algorithm using coarse and fine decisions will be optimized based on subjective and objective criteria by using computer simulation. Several different typical head-shoulder video telephone scenes of 30 seconds each have been used. The scenes were taken out of a video telephone system with 1 MHz analog bandwidth, 313 lines per frame and 50 fields per second. The sampling frequency was 2 MHz and the signal was quantized in 256 steps respectively 8 bit per sample. To check on the subjective criterion processed and unprocessed scenes have been displayed on a monitor. Test persons have to decide whether there was a difference to be seen between the two scenes or not.

The objective criterion has been applied by calculating the amount of changing picture elements and the average length of the clusters. The number of detected picture elements has been compared with the number, which has been detected by using the objective faultless detector with a coarse threshold of one out of 256 and which can be regarded as being objectively the ideal detector. First the usefulness of each of the three thresholds for changing detection shall be checked individually. The coarse decision alone leads to a subjective optimized change detector, if the threshold comes down to two out of 256. This decision leads to highly noncontiguous changing areas and requires a large number of bits for the positional information. The number of changing picture elements is reduced to 74 percent compared with the objectively ideal detector. The sign decision alone never leads to a subjectively faultless change detector, even if the size of the regarded picture area is enlarged from five differences up to seven or nine and of the threshold T_2 is put to zero. Besides this it was detected that the decision of the sign thresholding is overlapped by the two other detection criteria. This result shows that the probability of objects, which lead to frame-to-frame differences with alternating signs, is higher in normal video telephone scenes as it has been assumed before. Therefore the sign decision has been dropped completely. However, the sign decision may be still of some interest, when the camera noise is amplified by a quantizing process before change detection is done.

The decision based on threshold T_3, which includes a local averaging, is the best one in the subjective and objective sense. When the threshold T_3 is varied in such a way that the amount of changing picture elements is nearly the same as for $T_1 = 2/256$, the average length of contiguous changing areas is increased by about 15 percent and the required positional information is reduced. Besides this the two decisions of threshold T_1 and T_3 detect approximately the same changing parts if the total number of changing picture elements is the same in both cases. Therefore a change detector which uses only the absolute average difference criterion with threshold T_3 is the optimal one in the sense of bit rate reduction. The value of the threshold T_3 must be determined by the subjective criterion. The correct value is found, if test persons can no longer distinguish between processed and unprocessed scenes (error probability = 0.5). The probability of a right decision between the processed and unprocessed picture depends on the value of T_3. The result of the subjective tests are shown in Fig. 4. The error probability of 0.5 is reached with a threshold of $T_3 = 1.2/256$. Fig. 5 shows the subjectively and objectively

optimized change detector. The absolute average amplitude of five adjacent frame-to-frame differences is compared to the threshold $T_3 = 1.2/256$. The individual results of the optimization process are shown in Table 2.

The optimal change detector reduces the amount of changing picture elements to 69.5 percent compared with the ideal detector. In addition the optimal change detector generates in the average increased cluster lengths and shows no visible difference between the processed and unprocessed scenes.

3. MOVEMENT DETECTOR

Due to the temporal low pass characteristic of the human eye moving objects in a scene become blurred. This effect is amplified by the integrating properties of a camera tube and can be amplified by temporal filtering of the video signal. One can use these effects especially the low pass filtering to reduce the spatial resolution within the moving parts of a frame. A changing part detected by the change detector consists usually of a moving object and background which is uncovered by the moving object. Since the reduction of the spatial resolution is permitted only in the moving part and not in the uncovered background it is necessary to distinguish between these two parts within a changing area. This becomes quite simple when a temporal filter is used. The decomposition of a scene into single frames can be regarded as a sampling process with the frame frequency as the sampling frequency. Filtering in the temporal axis of a video signal is therefore done with a digital filter. If the filtered signal should not show any deterioration in picture quality as it may be caused by frequency depending group delay or by a change in the average amplitude a symmetrical FIR filter must be used. A block diagram of such a filter is shown in Fig. 6. The frequency response of the filter is given by

$$|H(j\omega_t)| = a_1 + 2a_o \cos\omega_t\tau_B \tag{1}$$

The sum of the coefficients must be one.

$$\sum_i a_i = 1 \tag{2}$$

ω_t is the temporal frequency and τ_B is the delay between two frames.
Important for the purpose of movement detection is the fact that a temporal filter has effect not only on the temporal bandwidth of the signal but also on the spatial bandwidth.

Fig. 7 shows an object consisting of five picture elements in the position A, B, C, D, E moving in x-direction with a speed of one picture element per frame. If the temporal filter with the frequency response of equation (1) is used, the output of the filter at the time instant t_i would be a function of the picture elements in the positions A, B, C. The same output can be achieved using a spatial filter, if the delay of this filter corresponds to the velocity of the object. For the example of Fig. 7 this delay must be the time of one picture element.

The spatial filter which is equivalent to the temporal filter described by equation (1) has the following frequency response:

$$|H_{eq}(j\omega_x, j\omega_y)| = a_1 + 2a_o \cos(\omega_x\tau_x + \omega_y\tau_y) \tag{3}$$

ω_x, ω_y are the two dimensional spatial frequencies. $\tau_{xy} = |\tau_x + \tau_y|$ is the delay of the filter which depends on the actual velocity of moving objects. If v_x and v_y are the velocity components measured in picture elements per frame time and lines per frame time then τ_{xy} is given by

$$\tau_{xy} = \left| v_x \tau_B \frac{1}{f_{ax}} + v_y \tau_B \frac{1}{f_{ay}} \right| \tag{4}$$

f_{ax}, f_{ay} are the sampling frequencies in the x- and y-direction. v_x and v_y are considered to be positive, when the object moves in the sampling direction, and they are considered to be negative in the other direction.
With equation (4) equation (3) can be written as

$$|H_{eq}(j\omega_{xy})| = a_1 + 2a_o \cos\omega_{xy}\tau_B \tag{5}$$

with

$$\omega_{xy} = \frac{\omega_x v_x}{f_{ax}} + \frac{\omega_y v_y}{f_{ay}} \tag{6}$$

The velocity dependence is now shifted from the delay term τ_{xy} to a frequency term ω_{xy}. ω_{xy} are frequency components which are created by a moving object in the signal. The two equivalent filters are shown in Fig. 8.

Under the condition of equation (2) it can be seen that $|H_{eq}(j\omega_{xy})|$ is equal one in stationary parts of the picture ($v_x = v_y = 0$) and differs from one whenever a velocity component is unequal zero.
The moving parts of a frame can be detected by comparing the output of the temporal filter with the unfiltered signal as it is shown in Fig. 9.
The difference between the two signals is equal zero in stationary parts of the frame,

independently whether the stationary parts belong to the changing or unchanging areas. The difference is unequal zero if it is calculated in the area of a moving object. Simple thresholding with a threshold of zero allows the segmentation of the frame in moving and stationary parts. Because it is sufficient to distinguish within the changing parts between moving and stationary picture elements the difference between the unfiltered signal is calculated only in the changing area. The decision is done for the temporal filtered signal.

4. CONCLUSION

A method is described which allows the segmentation of a video signal in changing and unchanging parts. The segmentation is done by comparing the average absolute value of five adjacent frame-to-frame differences to a threshold of 1.2/256.
The segmentation algorithm detects only those frame areas which have changed their information visibly. This is done for a minimum number of picture elements by using relatively long clusters, which in turn increases the bit rate reduction for conditional replenishment coders.

In addition to the segmentation of the frame in changing and unchanging areas a method is explained which allows an exact distinction between moving areas and uncovered background. This can be used for instance for the purpose of bit rate reduction by adapting the spatial sampling frequency to the reduced spatial resolution in the moving areas which depends on the velocity.

5. REFERENCES

CANDY, J.C., FRANKE, M.A., HASKELL, B.J., MOUNTS, F.W., 1971, "Transmitting Television as Clusters of Frame-to-Frame Differences", BSTJ Vol. 50, pp. 1889 - 1917

CONNER, D.J., HASKELL, B.J., MOUNTS, F.W., 1973, "A Frame-to-Frame Picturephone Coder for Signals Containing Differential Quantizing Noise", BSTJ Vol. 52, pp. 35 - 51

LIMB, J.O., PEASE, R.F.W., Walsh, U.A., 1976, "Combining Intraframe and Frame-to-Frame Coding for Television", BSTJ Vol. 53, pp. 1137 - 1173

MOUNTS, F.W., 1969, "A Video Encoding System Employing Conditional Picture-Element Replenishment", BSTJ Vol. 48, pp. 2545 - 2555

Table 1 Probability that the amplitude of Gaussian noise exceeds a certain threshold. The camera signal to noise ratio is 45 dB

threshold (n/256)	0.5	1	2	3	4
P (noise amplitude exceeds the threshold)	0.726	0.483	0.16	0.04	0.006

Table 2 Average length of clusters and percentage of changing picture elements which have been detected by the use of different methods for change detection

Threshold T_1 n out of 256	Threshold T_3 m out of 256	Average Length of Clusters in Picture Elements	Changing Picture Elements in Percent
1		25.4	100
2		15.7	74
	2	20.4	58.8
	1.2	25.7	69.5

STATIONARY BACKGROUND

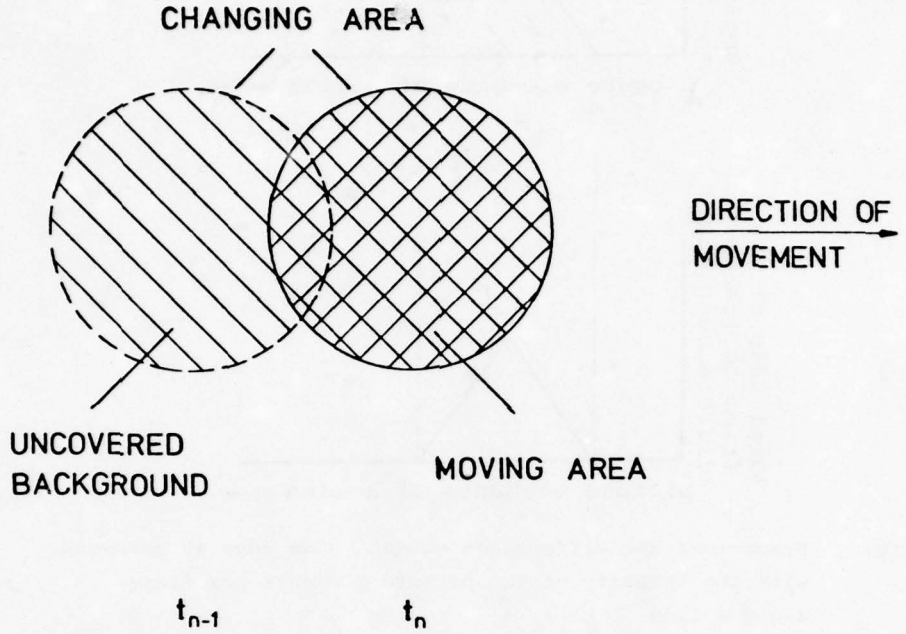


Fig. 1 Segmentation of a TV-frame in the case of a moving object

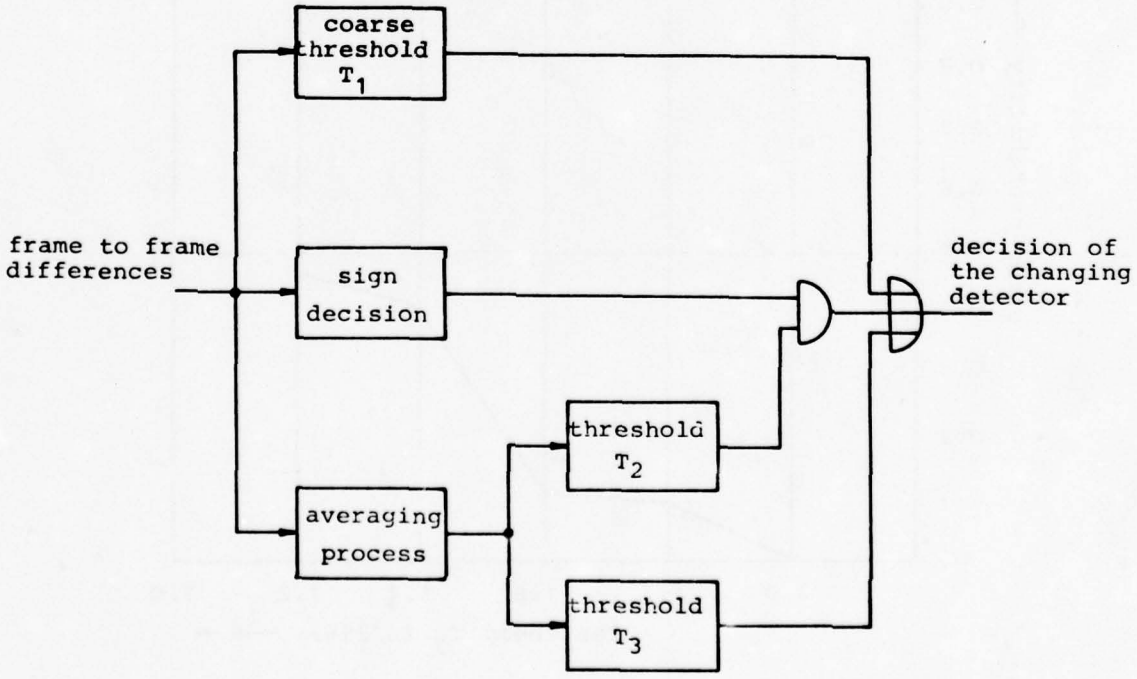


Fig. 2 Block diagram of a change detector which detects the temporal changing parts of a television frame by thresholding the frame-to-frame differences in three different ways

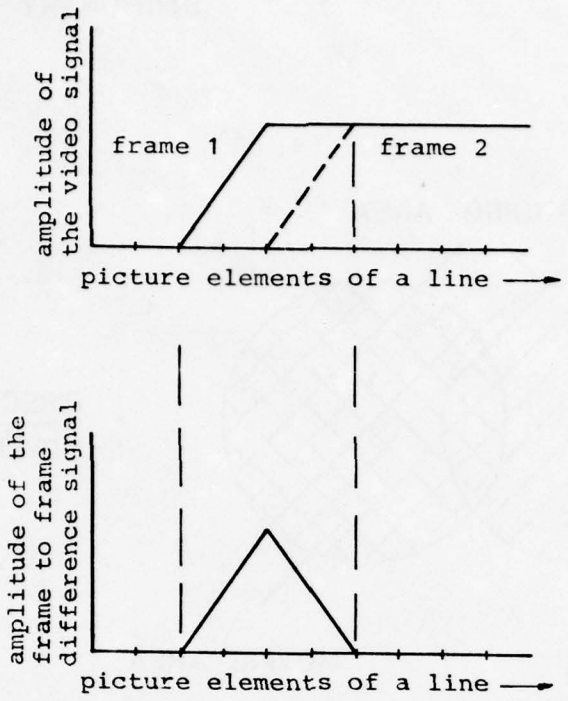


Fig. 3 Frame-to-frame differences caused by an edge which moves with the velocity of two picture elements per frame along a line

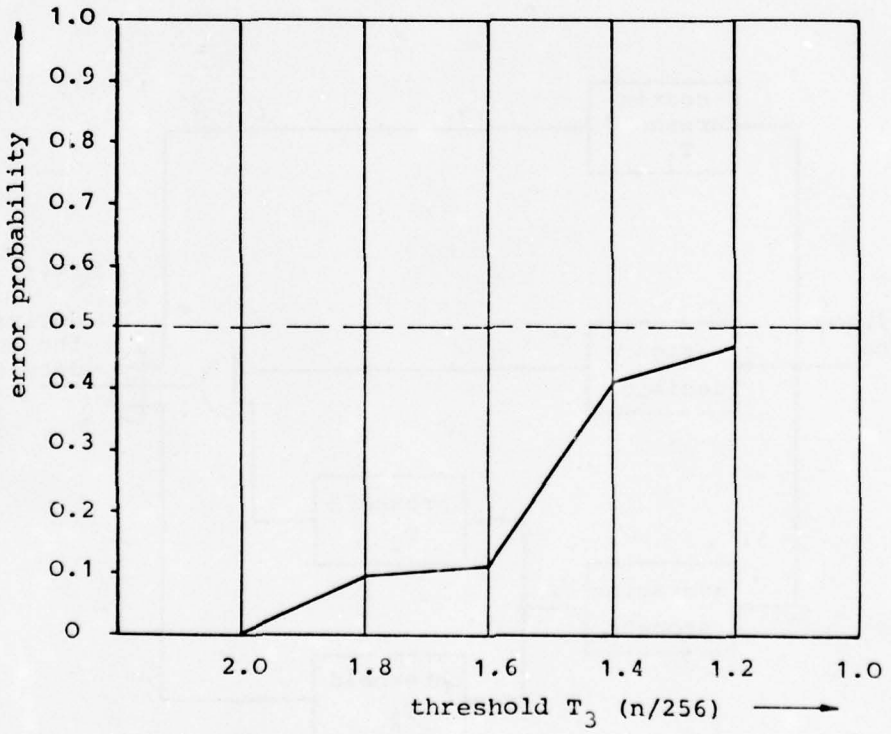


Fig. 4 Probability of a decision error of test persons who have to distinguish in subjective tests between an unprocessed or processed video sequence. The processing is done by a change detector which uses as criterion the absolute of the average difference

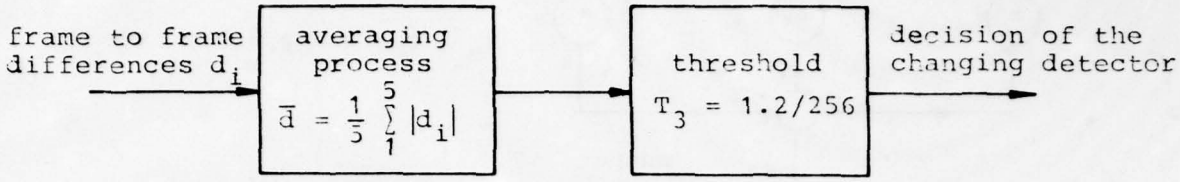


Fig. 5 Block diagram of a subjectively and objectively optimized change detector

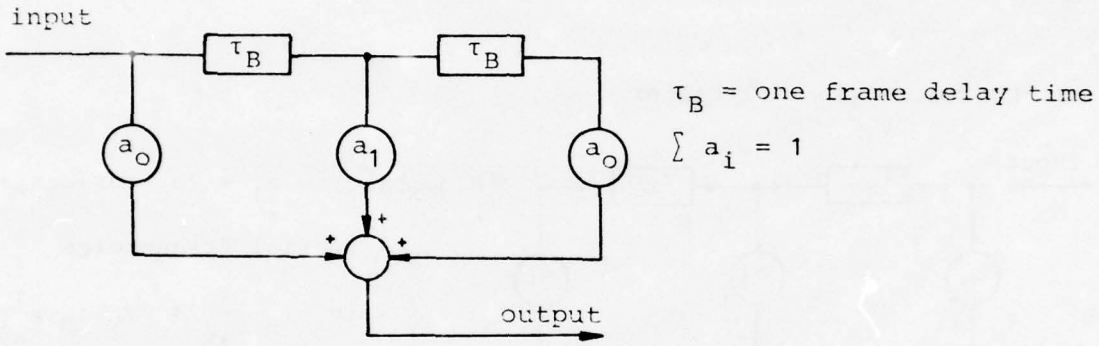


Fig. 6 Block diagram of a digital symmetrical FIR filter of second order with the sum of the coefficients being one. The filter is used for the temporal filtering of video signals

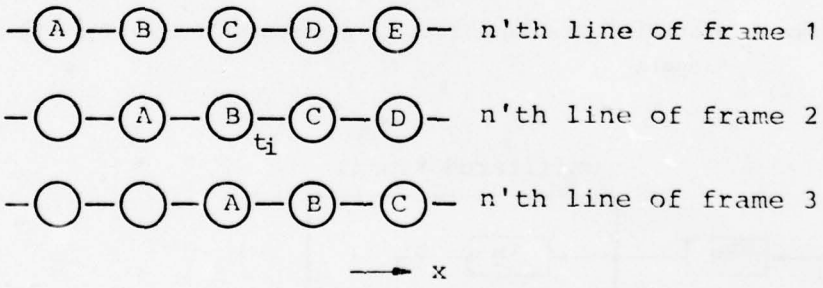
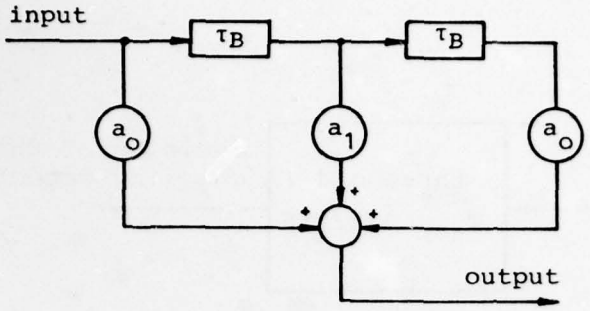


Fig. 7 Object consisting of five picture elements in the positions A, B, C, D, E which moves with the velocity of one picture element per frame in x-direction

Temporal Filter

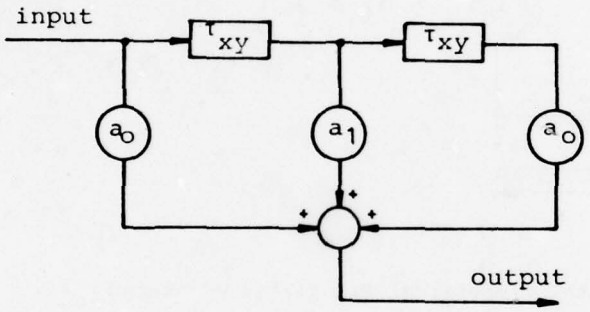


$$|H(j\omega_t)| = a_1 + 2a_o \cdot \cos\omega_t \tau_B$$

ω_t = temporal frequencies

τ_B = one frame delay time

Equivalent Spatial Filter



$$|H(j\omega_x, j\omega_y)| = a_1 + 2a_o \cdot \cos(\omega_x \tau_x + \omega_y \tau_y)$$

ω_x, ω_y = spatial frequencies

$$\tau_{xy} = \left| v_x \cdot \tau_B \cdot \frac{1}{f_{ax}} + v_y \cdot \tau_B \cdot \frac{1}{f_{ay}} \right|$$

$v_{x,y}$ = velocity components

f_{ax} = sampling frequency in in x-direction

f_{ay} = sampling frequency in y-direction

Fig. 8 Spatial filtering effect of a temporal filter for video signals

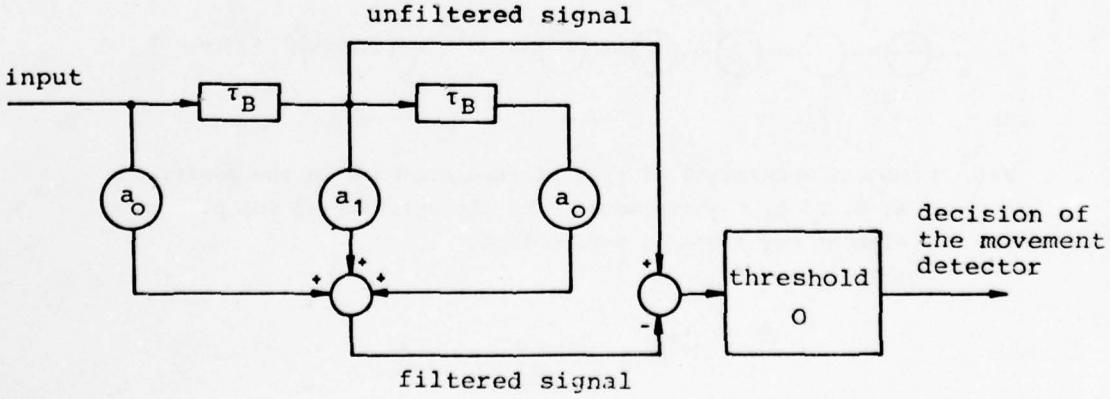


Fig. 9 Block diagram of a movement detector

STATE OF THE ART IN DIGITAL SIGNAL PROCESSING WITH APPLICATIONS TO MULTIPLE ACCESS SYSTEMS

Lester A. Gerhardt
Professor and Chairman
Electrical and Systems Engineering
Rensselaer Polytechnic Institute
Troy, N. Y. 12181
USA

ABSTRACT

This invited paper serves as the introductory address for the session on Multiple Access at the AGARD-NATO Symposium on Digital Communications in Avionics. The introduction sets the stage for the papers to follow by reviewing some of the more advanced developments and trends in the fields of digital signal processing and digital communications. The effect of the computer is then briefly covered as it has influenced this development, especially the trend to minicomputers and microprocessors, and distributed computing. The availability of new devices, device technology and directions are then summarized with respect to the emphasis on speed. This leads to a discussion of CDM, CDMA, SSMA, and TDMA. The Joint Tactical Information Distribution System (JTIDS) is described as a system concept. The papers for the session are then reviewed in context of the scope of this paper.

Introduction

Inherent in digital signal processing is the idea of discretizing the data in time (sampling), PAM, and often in amplitude (quantization), PCM. As applied to communications, any set of calculations performed on the so generated set of samples, usually performed in real time on line, is considered digital signal processing. This includes filtering, modulation, transformation of data, etc. Originally, design was performed in the analog world, and these optimized designs were converted to digital form. Now, fortunately, design is usually done directly in the digital domain.

The need for digital signal processing has been well established. In the way of review, digital processing offers reduced costs, ease of LSI, greater flexibility, error detection/correction capability, signals are easily regenerated, increased reliability, etc.

The trend to digital signal processing is a multifaceted motivation. First there are the system requirements which require an approach with the advantages cited above. On the other hand, there is the improvement of devices and the technology that offers the possibility to meet these demands in a realistic time frame using digital tools. Last, there is the ever increasing numbers of computers as a source of data, further demanding the transmission of information in digital form.

The beginnings of digital signal processing which saw transformations of analog designs to a digital domain, was particularly applied in filtering and resulted in large numbers of transformations and designated the indirect design approach.

It wasn't long before design was done directly in the digital domain. Commonplace now are the IIR and FIR filters. The applications to filtering soon expanded into many parallel efforts. High speed transforms were discovered (such as the Fast Fourier Transform) which revolutionized signal processing. Digital spectral analysis became commonplace. Simultaneously, came the revolution in computers resulting in the microprocessor. For the first time the computer could now be considered as an integral part of the system.

Applications followed almost immediately which used these approaches with digital communications being at the forefront. The use of digital phase locked loops (Costas), SSB modulation, spectral analysis are all examples. Overall the impact was great on voice transmission, data communication, image transmission and radar to name a few.

Computers and Trends

As mentioned, digital computers, themselves a source of discrete time and amplitude information, served as a major impetus to determine the course of digital signal processing and digital communications. From the bigger and faster machines such as the IBM 3033 and Cray at one extreme to the microprocessor at the other, the revolution was here to stay. However, it is my opinion that the major effect on communications was due to the microprocessor. Recall when the Intel 4004 was no sooner on the streets than the 8008 was announced. In turn came the 8080 with a ten times speed factor improvement. We now have available to us 16 bit microprocessors, with virtually any peripheral desired, accounting for the majority of the computer market. Computer stores are commonplace as are home computers and computerized test equipment. The advent of the microprocessor has caused the computer to permeate all walks of life all within the last half dozen years. The computer as an aid to digital communications is clear, not only as a source of data to be transmitted, but as a processor for inherently analog data, such as voice or video.

Another major factor affecting digital communications, is the trend towards distributed computing, in part again caused directly by the microprocessor and the direction to smaller less costly digital systems. Being a discrete source of data and being distributed, it further served to enhance and establish the need for more effective means of communicating digital data.

The motivation of the field toward distributed computing and networks may be better understood by clarifying the types of distribution involved.

Distributed computing capability is most often thought of as distributed location (on a national scale or on a local scale). Usually, the geographical distribution of computer facilities is already in existence and the motivation is to interconnect these in some sort of a network to increase overall computing power and/or

capability.

Distributed computing is many time stimulated by the appeal of distributed cost. Simply, one Section within a company, or one computer Center, cannot afford the cost in purchase or maintenance, of a facility it desires with the power it needs.

Distributed capability computing may arise from a tendency of distributed functionality. This exemplifies a trend to perform only restricted types of computing on a single machine, with different machines set aside for specific functions. This is typified by a self contained interactive graphics facility or a separate engineering computer at several major universities.

Again an additional overall impetus for the trend to distributed systems, is the greater availability for lower cost of minicomputers and microprocessors, thereby making the number of disjoint installations grow, yet each with its own limited capability.

Computer networks are the only practical means available to permit sharing of computer resources, information handling equipment etc. They offer and permit equality of access and quality independent of location. Today, about one-third of major public and private information dissemination centers provide services by computer networks, and we are experiencing an explosion of added users.

Many networks presently in existence attest to the need and significance of this trend. One of the most well known is the ARPANET in operation since 1971. Other industrially developed networks include INFONET (by Computer Sciences Corp.), DECNET (Digital Equipment Corp.), CYBERNET (Control Data Corp.), IBM-TSS and GE Information Services Network. University developed networks range from the ALOHA system (a satellite network based in Hawaii), the MERIT network (Michigan Educational Research Information Triad interconnecting the three largest university computer centers in the state), to TUCC (Triangle University Computer Center at North Carolina), the DCS (Distributed Computing System at the University of California at Irvine), and the DCN (Distributed Computing Network at the University of Maryland). Specialized commercial networks such as NASDAQ (National Association of Securities Dealers Automated Quotations - an over the counter stock quotation network), and the Interairlines Network (involving all US major carriers) must be added to the list. Of course this is not to leave out TYMNET (a packet switched system involving over 100 computers developed by Tymshare), OCTOPUS (at Lawrence Livermore Laboratory) or others. The network concept has expanded to an international scope as well, as represented by the EIN (European Informatics Network) involving France, Italy, Norway, Portugal, UK, Switzerland, and Yugoslavia; the EPSS (British Experimental Packet Switching Service), the CYLADES system (interconnecting French universities and research centers), and DDX (Denden Kosha Data Exchange) done by NIPPON Electric of Japan.

Yet the field is only a few years old and remains in its developing stages, hastened by the rapidly accelerating technology that surrounds it. The situation is such that it has become difficult to separate out the fields of computers, communications, and computer communications which is perhaps best for the sake of the integration needed for success.

Devices, Processing Technology and Speed

The key word is still speed. This is of particular importance in communications where the time constants require more rapid computation. Speed is brought about in two basic ways: one is by improvement of architecture and the other by improved device technology or new devices.

In the area of architecture, the 60's and 70's saw many new innovations including pipelining, and array processors. Systems such as the Floating Point Systems array processor provides for a jump discontinuity of improvement in processing speed and capability for communications applications. The entire question of parallel vs. serial processing became one raised often.

In a software sense, high level languages gave way to assembly languages (again) as we headed to the minis and then micros to attain higher speeds. At the other end of the spectrum was the development of new devices. These included surface acoustic wave devices, charge coupled devices (CCD) among others, the latter offering time bandwidth product of 250 and 100 Mhz operation. These permitted performing such operations as an inner product, so popular in communications, in much faster times than realized by conventional logic. They did however, realize this in what must be viewed as substantially an analog device! Operations such as correlation, convolution, matched filtering are now performed by such devices in many advanced systems.

Further, there is the trend to new processing technologies and the realization of submicron line widths. With finer geometries, higher speed is again possible, as is higher density. The use of silicon and GaAs integrated circuits offer the potential of still higher speed. Emitter coupled logic (ECL) is yet another option. The use of electron beam technology for device fabrication and/or optical methods directly tied into computer aided design approaches offer much improved processing speeds, yields, and size reduction all aimed at speed improvement.

In addition to the speed improvement sought for logic devices, there exists a need for improved memories. Read only memories, (ROM), and programmable ROM(PROM) helped with the speed problem. A formidible research thrust exists in this direction as well. Examples such as the bubble memories demonstrate the push for larger faster memories as one of the weaker links in the computer building blocks. Already over one million bit arrays have been developed.

In short, it should be clear that from the systems side as from the device and processing side, the motivation is to go smaller and faster; such forces can only result in faster processing speeds, a consequence of direct benefit to digital communications.

It is this environment in which we find ourselves, one which cannot help but foster the growth of digital technology, and digital communications.

Spread Spectrum Communications and Multiple Access

Thus far, the paper has dealt with the general area of digital signal processing and digital communications and the impact of computer technology on these fields. There have been major repercussions in several related subfields, due to these rapid if not revolutionary changes. One of these has been that of spread spectrum communications, a subset of the more general field of digital communications. Broadly speaking, a spread spectrum signal is generated by modulating a data signal onto a wideband carrier so that the resultant transmitted signal has bandwidth which is much larger than the data signal bandwidth and which is relatively insensitive to the data signal.

Spread spectrum developments actually began in the 1940's as the result of "clever engineering" on a selected basis, with most applications involving military problems. The technology has only recently become popularized in the general technical community. There were many reasons for its inception, one of the principle factors being the desire for an antijam capability, usually in a single user application - which helped maintain its low profile. As the advancements progressed, the spread spectrum signaling concepts have been found to be well suited to precision range and position location, and most recently have been successfully applied to multiple access situations involving many users simultaneously. In fact, the recent upsurge in open discussions of SS techniques has probably been caused in part by applications of the concept to multiple-user communication situations where large amounts of interference are encountered, and in part by a rapidly advancing technology which is making more intricate signal processing feasible.

Only recently has technology come to the point of making circuitry and systems reasonably small, reliable, and inexpensive so as to enable practical implementations of spread spectrum concepts. Viewed as a motivating force encouraging the growth of the field, this relatively recently developed capability for practical spread spectrum systems must be reinforced by the additional pressure of more and greater demands being made on communications systems than ever before. Increased message traffic from a higher number of users is creating a need for protection of information from interference and eavesdropping, not only in a military, but in a commercial environment as well. As a result of these two major forces, the availability of systems and components coupled with the need for improved communications, the field of spread spectrum communications has rapidly emerged in recent years as a major thrust in the technical community.

Nonetheless, over the last five years, the work in the field still has been primarily associated with military applications. JTIDS, PLRS, USC-28, and others are advanced systems that stretch today's design capability to the fullest. SEEK TALK, SINCGARS, advanced JTIDS, and their contemporaries are beginning to provide further impetus toward smaller, lighter, and more capable systems that are readily adaptable to volume production. (JTIDS will be briefly discussed in this paper and then again in detail in a paper by McMillan to follow.)

For the most part, during this period use was made of the same basic techniques, frequency hopping, time hopping, chirp, etc., noting perhaps some innovations such as offset QPSK. However, major strides were taken in the development of new devices and LSI techniques which offered the potential of reduction to practice for many of the heretofore theoretical spread spectrum concepts. Overall there was a decided improvement in and availability of means for implementation brought about by microminiaturization. At the same time, new device developments, such as surface acoustic wave (SAW) and charge coupled devices (CCD), offered still further opportunities for spread spectrum implementations. All of these led to simplification of receiver structures, synthesizers, approaches to synchronization, etc. Theoretical work also progressed, in the area of coding waveform development among others, but in perspective the major contributions were in the implementation domain.

In consideration of the future, with every new advance in integrated circuitry or more capable microprocessors, improvements in spread spectrum techniques are being and continue to be made. Great strides are being made in the area of matched filter correlators - primarily in the charge coupled and digital matched filter areas. Where synchacquisition of spread spectrum receivers once took minutes or seconds, we now have the promise of acquisition in milliseconds or even microseconds.

One should reasonably expect more emphasis on multiuser applications for both the commerical and military marketplace. Spread spectrum systems with more than gigahertz bandwidths, personal telephones that give fully portable wireless service to hundreds of users in a time division multiplexed frequency hopping format, and new frequency assignments allowing coexistence of conventional radio systems with newer spread spectrum systems are anticipated at relatively low costs. More use of adaptive techniques is expected in advanced systems.

In the way of problem areas, there are, as always, still some remaining. One always needs improved technology, and this is especially true of the spread spectrum field, with higher device speeds being a requirement. Longer codes continue to be sought with ever increasing time bandwidth products. Synchronization, although not as significant a problem as it once was, still remains in need of better solutions, and network timing confronts the future engineer as a problem area.

Overall, the task of the new systems is not only to accept the older systems without significant degradation, but to provide minimum interference or loss in performance to those systems.

Much of today's prime effort is being spent in the area of developing compatible, low density signaling structures that cannot only improve spread spectrum system performance but provide peaceful coexistence with other systems. What is today's capability? Where are we going? And what will it take to get there? Only imagination, a great deal of advanced design, and some very understanding frequency allocations.

As in most fields of rapidly advancing engineering and science, results are not always published or otherwise disseminated in a logical order. In the field of spread spectrum, this has been further compounded by the classified nature and origin of the work. This tends to continue in the area of multiple access as well.

As a result, to maintain this paper as unclassified, some care has beenexcercised to reference material already available in the open literature. The description that follows involves multiple access spread

spectrum, and is taken directly from the recently published IEEE Special Issue on Spread Spectrum (1). In particular, the summary comments are taken from two of the four papers which appear in this issue on the subject of multiple access (2,5). The credit for this following portion resides solely with these authors (M. Pursley and N.C. Mohanty).

Code Division Multiplexing (CDM) has in recent years become as well known as its more conventional counterparts of time division multiplexing (TDM), and frequency division multiplexing (FDM). The advantages of the CDM approach are its lack of complexity, random access, privacy, immunity to jamming and ranging using spread spectrum, as pointed out before and in the paper by Lindner to follow later in this session.

In recent years there has been increased interest in a class of multiple-access techniques known as code-division multiple access (CDMA). The CDMA techniques are those multiple-access methods in which the multiple-access capability is due primarily to coding and in which, unlike traditional time- and frequency-division multiple access, there is no requirement for precise time or frequency coordination between the transmitters in the system (2). CDMA techniques have been considered for a variety of satellite systems including the NASA tracking and data-relay system, systems to provide communication to aircraft and other mobile users (see Bernstein's paper, the third in this session for example), air traffic control systems, and military satellite communication systems (see Baerwald's paper, the 5th in this session). In certain satellite communication systems, CDMA techniques can be designed to provide multiple-access capability and simultaneously, to reduce the effects of multipath distortion (11). All of these approaches are integrated in an ICNI concept in the last paper in this session.

The most common form of CDMA is spread-spectrum multiple access (SSMA) in which each user is assigned a particular code sequence which is modulated on the carrier along with the digital data. The SSMA techniques are characterized by the use of a high-rate code (i.e., many code symbols per data symbol) which has the effect of spreading the bandwidth of the data signal. The two most common forms of SSMA are frequency-hopped SSMA and phase-coded SSMA. The first of these two was used in the TATS modulation system for the Lincoln Experimental Satellites and is described in detail in (12). Phase-coded SSMA (also known as direct-sequence spread spectrum (13,14) utilizes the most common form of spread-spectrum modulation: the carrier is phase modulated by the digital data sequence and the code sequence. Although phase-coded spread-spectrum modulation has been considered for a wide variety of purposes (12,13), fewer have been concerned with its use in achieving multiple-access capability (2).

In Pursely's paper he considers phase-coded SSMA system analysis, concentrating on communication performance rather than on acquisition and tracking performance, so that the performance measures of interest are error rate and signal-to-noise ratio. Although various aspects of phase-coded SSMA communication were discussed in a number of publications which appeared in the mid-1960's (e.g., 15-19), there were very few analytical results on asynchronous phase-coded systems and little had been done to identify the important code parameters for asynchronous phase-coded SSMA applications. Most of this work implicitly or explicitly assumed a synchronous model and therefore dealt only with the periodic cross-correlation properties of the code sequence. Further, nearly all of the results on cross-correlation properties of sequences dealt with only the periodic correlation.

One of the first detailed investigations of asynchronous phase-coded SSMA system performance which dealt with aperiodic cross-correlation effects was published in 1969 by Anderson and Wintz (20). They obtained a bound on the signal-to-noise ratio at the output of the correlation receiver for a SSMA system with a hard-limiter in the channel. The need for considering the aperiodic cross-correlation properties of the code sequences is clearly demonstrated in their paper. Since that time, many additional results have been obtained (e.g., 3,11) which help clarify the role of aperiodic correlation in asynchronous phase-coded SSMA communication. (For example, Yao (3) especially shows that the Gaussian interference model popularly used, is not valid for SSMA with small numbers of users, low length codes, and high SNR.)

The major advantages of SSMA are: (5)

1) it does not need any timing coordination;
2) it has simultaneous random access;
3) more importantly, repeater bandwidth of a satellite is utilized efficiently and no guard bands are inserted; and
4) it provides simultaneous ranging with telephone communication.

The system parameters used for ATS-1 satellite are given below: (5)

MODULATOR	
operating frequency	70 MHz
reference frequency stability	1×10^{-9}/Hz
input signal	voice: 300 Hz to 3.4 kHz BW FSK data: 100 bands
primary modulation	FM with 3.4 kHz to 12 kHz peak frequency deviation
code type	maximal length (PN)
code clock rate	16.376 MHz
code length	2047
PN carrier modulation	phase reversal
DEMODULATOR	
operating frequency	70 MHz
message demodulator	FM discriminator
correlator	phase reversal
cross correlation tracking	delay lock discriminator
acquisition	sweep by clock frequency offset up to 500 Hz

measuring range	
time accuracy	1×10^{-8} s
reading digits	8 digits
sampling rate	about 1 sample/s

The problem of synchronization with direct sequence spread spectrum is reduced in frequency hopping (FH) spread spectrum where the frequency of the transmitted signal stays on a specific frequency for an interval and hops to another frequency selected pseudorandomly. Some features of the PN sequence modulations and frequency hopping are as follows: (5)

CHARACTERISTIC	PN	FREQUENCY HOPPING
Multiple Access (with time division)		
Near margin	20 dB	60 dB
Sync time	3 x uncertainty x processing gain	10 data bits to prevent false clock
Long code serial search	several seconds	milliseconds

The major requirement in FH systems is the need for a fast settling time, which permits a rapid hopping to accommodate high data rates with little off time. Error correction coding is extremely important with FH M-ary FSK. Reed-Solomon codes are usually used for this purpose. Whereas in the direct sequence the phase was coded, in frequency hopping, the frequencies are coded (random). When the data are stored in bursts the transmission of bursts is done in coded time. The slot in which the transmitted pulse occurs is selected by a code generator. The data bits within a frame are stored for a transmission at high speed at selected time intervals and this information is known a priori at the receiver. The time hopping system (TH) not only scatters the timing information but utilizes the slots of the idle users.

In the TDMA system, assuming a satellite application,due to assignment of time frames, a single signal is present at a time (5). This eliminates intermodulation products within the satellite transponder. With phase modulated or digitally encoded signals the TWT amplifier can be operated at saturation without any power loss. The uplink power need not be controlled. The frequency of signals received is different from that of the transmitted signal to avoid interference. The frequency stability is not critical. The TDMA is ideal for baseband transmissions including voise and data transmission. As the number of users increases, TDMA systems perform much better than FDMA systems (5). In TDMA systems analog messages have to be digitalized and the messages are to be transmitted in bursts which require buffer storage, unique word detection, and burst synchronization.

TDMA modems are identical wideband burst modems for all stations. Unique word detection and burst synchronization have been dealt with for the TDMA system by Schrempp and Sekimoto (21) and Gabbard (22). Each earth station consists of three major subsystems: (5)

1) PCM Coder
2) Control Subsystem and
3) PSK Modem.

The control system is the central part of the TDMA system. The burst length is the length of a single uninterrupted transmission from an earth station, a frame contains a burst from each accessing station, and guard time is the time between the end of one burst and the beginning of the next burst. Each station burst has a format made up of preamble bits and information bits. The preamble has the following functions:

1) It contains a sufficient number of bits for recovery of a coherent carrier and bit timing for the demodulation (for example PSK). This has to be done independently for each burst since it may not be possible to have coherency of carrier phase and bit timing between transmitting stations.

2) It contains a sufficient number of bits of unique words that are used for station addressing and word and burst synchronization. The guard time is provided between bursts to prevent interference due to overlapping or defective equalization.

A typical PCM coder has the following characteristics: (5)

modulation	pulse code
number of channels	24
sampling frequency	8000 samples/s/channel
clock frequency	1.544 MHz
sampling interval	7 bits for voice plus 1 bit for signaling

The PSK modulator-demodulator may have the following features:

modulation	two phase PSK
demodulation	coherent
IF frequency	70 MHz
frequency stability	10^{-5} per day
carrier recovery	1000 μs
phase ambiguity removal	differential bit coding
operating bit rate	6.176 Mbits/s

The structure of a typical preamble has the following patterns: regular pattern for 30 symbols (60 bits) for carrier phase and bit clock recovery; a unique word of 10 symbols (20 bits) which is used with a correlation detector, to generate a time signal at the tenth symbol from which word-length timing for the rest of the burst is reckoned; four symbols (8 bits), consisting of a six bit station identification signal, plus two bits used for supervisory signals, to control the activity of standby reference stations; and a sequence of symbols to provide service circuits and various housekeeping signaling functions.

In the Defense Satellite Communications System, frame rate for TDMA is fixed at 1200 Hz (5). The time base at each terminal is synchronized to that of the master terminal. The guard band varies from 0 to 100 ns.

The signals are translated in frequency and then retransmitted to the earth terminal. Burst rates range between 1.2288 and 78.6432 Mbit/s. Convolutional coding is used for forward error correction, the modulation is QPSK type, and signaling rates range between 1.2288 and 78.6432 MHz. The range between earth terminal and satellite is over 23,000 miles. The terrestrial subsystem consists of (5)

1) earth terminal complex (ET)
2) an interconnect facility (ICF)
3) a technical control facility (TCF).

The TCF contains Pulse Code Modulation (PCM) multiplex and asynchronous time division multiplex units. The objective of ET is to provide synchronization timing, control, buffer, modulate and demodulate code and decode when necessary and to provide Doppler correction. The buffer is interfacing storage devices which are unique to TDMA systems. The most efficient buffer would have a capacity equal to the terminal aggregate receive rate divided by the frame rate. The memory read/write clocks are bounded by the highest burst rate 78.6432 MHz, and frame rate 1200 Hz. Data are temporarily stored in memory elements and data for any channel can be assigned to a continuous memory block on a first come first served basis or on a priority basis. The memory block length is determined by the I/O capabilities, random access, and shift register. The Random Access Memory supports memory blocks of arbitrary size. The Monitor, Alarm, and Control System determines the operation flexibility and availability. This monitors

1) error correction
2) QPSK type modulation
3) IF and carrier frequency
4) baseband interference
5) burst rate and length including preamble.

In spite of all these complexities, most of the future satellite systems would be TDMA systems.

In his paper (5), which serves as the source for the above, Mohanty nicely describes a hybrid system, a combination of SSMA and TDMA, which is stated to be within the state of the art for both military and commercial applications. Implicit in this discussion, there is an underlying emphasis on satellite applications. This is continued in the papers of the session as well (Bernstein, Baerwald). Therefore, this area as a viable application is briefly covered using Mohanty (5) again as an open literature reference, to put the use of the various methods in proper perspective.

Satellite repeaters provide communication links to users separated by large distances, or in inaccessible locations on sea, land, or in space. The capacity of the satellite link is as good as the terrestrial link and three geostationary satellites can cover most of the earth. Any earth station can listen to any signal including its own and it can detect errors using error correcting codes. But the number of channels in the satellite transponder is limited by the power and bandwidth of the repeater. These channels might be Frequency Division Multiple Access (FDMA), Time Division Multiple Access (TDMA), Spread Spectrum Multiple Access (SSMA), or Code Division Multiple Access (CDMA). The multiple access schemes employ modulations which have disjoint frequency, time or distinct codes. The various multiple access modulations already have been discussed. The FDMA system is the simplest of all existing multiaccess systems as it uses only traditional frequency division multiplexing hardware. The commercial satellites Intelsat I through Intelsat IV and all traffic control units use FDMA systems and many military satellites are still FDMA. The disadvantages of the FDMA systems is that the system capacity is limited by the intermodulation in the satellite repeater. The satellite repeater has a nonlinear TWT amplifier along with other filters. An uplink power control is required to make full use of the repeater capacity. The total capacity undergoes a rapid drop between one and four accesses. In view of these disadvantages, the Intelsat V, Canadian domestic satellite system, and USA military systems among others are switching to TDMA systems. The SSMA system has been used exclusively in military systems under various names such as frequency hopping, pseudonoise systems, and jamming systems.

Finally, to round out this discussion, the JTIDS system will be briefly discussed. It is discussed in detail in the paper by McMillan to follow. JTIDS is a time division multiple access communications system.

The basic building block of the JTIDS message for any of the candidate architectures is a 6.4 microsecond pulse which conveys five bits of information. Each successive pulse is transmitted on a different frequency with a different phase code to provide protection against jamming and interference. In the TDMA signal structure there is an option to transmit the same five bits of information in two successive pulses. The redundancy of this "double pulse" waveform provides additional jamming protection.

The exact same symbol pulse is used in all JTIDS architectures. Furthermore, the same pulse structure is used for both synchronization and data signals. Thus, the basic building block for TDMA, ATDMA, DTDMA, and HTDMA communications is the same.

In the case of multiple nets, each terminal that transmits simultaneously will use a different pulse-to-pulse frequency pattern. Furthermore, each symbol pulse will employ a different phase code. Thus, the individual pulses on one net will be separated in both frequency and code from those on another net. This separation minimizes the chance of interference between pulses and permits a receiver to select the pulses from the net of interest.

The essential characteristic of Time Division Multiple Access, or TDMA, is that time is considered to be divided up into a series of intervals called "time slots". All time slots are of equal length, and all terminals synchronize to a common system time so that each terminal knows the boundaries of the time slots. A TDMA user broadcasts a complete message of fixed length within a time slot on any one of 128 nets. (A net being characterized by a pseudorandom frequency and phase code sequence distinct from all other sequences that could be used in that time slot for any other net.) In the basic TDMA structure, normally only one user at a time broadcasts in a net. Multiple user transmissions, other than net entry and relayed messages, can only occur in a time slot if each transmission is in a different net. A user can transmit or receive in a given time slot, but cannot do both simultaneously. Transmission or reception can be in any one net, and is dynamically programmable between nets, time slot by time slot. A "guard" period or silent period occurs after the end of each message to permit the signals to propagate over a wide area before the transmission from the next user begins. Thus, messages in each net appear in a serial fashion.

The TDMA structure was designed primarily for transmission of fixed format position and track messages. As the requirement developed for transmission of different types of messages of varying duration, such as digitized voice, other slot sizes and message lengths were developed to optimize system throughout for these applications. These variations are grouped under the title Advanced TDMA, or ATDMA.

Unlike the TDMA and ATDMA architecture where an entire message is transmitted in a single burst, each DTDMA transmission interval contains only a single 6.4 microsecond pulse. Users are assigned transmission intervals at an average rate commensurate with their reporting requirements. These intervals occur pseudorandomly in time and are interleaved with transmission opportunity intervals of other users. The system is time ordered in the sense that within each DTMA net only one user broadcasts in a transmission interval. However, since the propagation time of a pulse is much longer than the transmission interval (milliseconds versus microseconds), and since users are at random locations with respect to each other, time ordering is not maintained at the receiver. The receiver relies on code division processing techniques to recover the data.

HTDMA provides for simultaneous transmission or reception of message bursts within time slots in addition to distributed transmission or reception of pulses as described in the DTDMA section. Unlike DTDMA, tracking of sources is not required. Unlike TDMA, a terminal can process multiple messages simultaneously. The HTDMA capability is achieved by allocating a small number of time slots or half slots in a TDMA/ATDMA net for transmission of HTDMA synchronization bursts. All other slots in the net are available for normal TDMA/ATDMA transmissions.

With this as background, the reader should be in a position to more fully appreciate the papers to follow and place them in the proper context.

The Session's Papers

This invited paper serves as the introduction to the Session on Multiple Access for the AGARD-NATO Symposium on Digital Communications in Avionics. Consequently, it concludes with a brief description of the papers to follow, presented in the context of this talk.

All papers deal with multiple access systems. The first, Modem Telegraphique 75 Bauds a Etalement de Spectre by D. Brisset describes a spread spectrum multiple access system (SSMA). Various methods of multiple access are first reviewed followed by a detailed treatment of the advantages and disadvantages of the SSMA approach including the effects of amplitude and phase distortion, doppler, etc. It makes a particularly fitting paper with which to start off the session.

The Performance of Code Division Multiplexing with Pulse Position Modulation by Lindner is self-explanatory by its title. The PPM system described enables the transmission of analog samples directly without the need for PCM or delta modulation. The PPM system is analyzed in detail for various cases. The similarity to methods propsed by Viterbi is shown for the case where the binary sequences act as carrier signals. The PPM-CDM case is also analyzed and compared with a PCM-CDM system using both computer simulation and actual equipment for speech transmission, with the results that PPM-CDM is found to be more advantageous.

The third paper, A Terminal Access Control System for Fleetsat by Bernstein, describes a demand assignment multiple access systems, DAMA, for use with data and voice primarily intended for mobile users. The system is inherently a TDMA based design, uses convolutional coding and Viterbi decoding, and a dual microprocessor control section in the subscriber unit. Implemented on a minicomputer (combining the ideas of this paper with respect to the trend to mini and microcomputers) there is only need for one central access controller. Testing is planned for Spring 1978 with the thrust of the program towards the Navy's UHF DAMA TRI-TAC program.

Implementing JTIDS in Tactical Aircraft by McMillan, describes the Joint Tactical Information Distribution System, previously addressed in this paper. It deals with the requirements, system architecture, critical technologies, and system integration. The program goals are the implementation of digital communications for tactical aircraft and as such provides a focus of the methodology offered in previous papers to a complete tactical system.

The fifth paper is also oriented to satellite communications with military applications. TDMA for Relayed Communications by Baerwald is an excellent review of the state of the art in TDMA as applied to satellite communications and the impact on the various aspects of communications systems operation and performance. It treats electronically steered phased arrays, adaptive processing, timing, AJ capability, and a demand assignment methodology. TDMA is reviewed both for the present and future potential, where it is offered as one of the most viable alternatives for the applications suggested.

The last paper, Integrated Communication, Navigation and Identification (CNI) in the 1980's by Kennedy provides an appropriate ending to the session. It considers the emergence of distributed TFCDMA, Time-Frequency-Code Division Multiple Access technology, thereby combining the ideas presented in the previous papers in one unifying concept. The integrated CNI provides a nice vehicle for presenting the case effectively. The possibility of receive while transmit is described as is the use of different CNI networks by a user, with the paper completed by a discussion of the operational impact of the proposed approach.

CONCLUDING REMARKS

Overall, the major approaches to multiple access have been presented in this session, both independently, and in combined use for maximum effectiveness. Some have been realized, some are in work, and some are in the conceptual stages. The stage for the development of such sophisticated digital communications systems and concepts has been set in this paper. Given the literal explosion in computer technology from both the systems and device technologies, and the recent developments, coupled with the impact of spread spectrum systems and the need for secure multiple access, the concepts described will be a virtual reality in the near future. It is not to be viewed as a passing fad, but a major thrust in technological development;

a trend to a new type of spectral allocation fully utilizing a digital approach, a trend begun in circles where control can be excercised over the complete communications network (such as in military applications), but one which may well spread to broader international communications for both military and commercial purposes in the not too distant future.

REFERENCES

1. Special Issue on Spread Spectrum Communications - IEEE Transactions on Communications - Gerhardt, L. A. and Dixon, R. C., Guest Editors, August 1977.

2. Performance Evaluation for Phase-Coded Spread Spectrum Multiple Access Communication - Part I System Analysis, Pursley, M. B., IBID.

3. Performance Evaluation for Phase-Coded Spread Spectrum Multiple Access Communication - Part II Code Sequence Analysis, Pursley, M. B., Sarwate, D. V., IBID.

4. Error Probability of Asynchronous Spread Spectrum Multiple Access Communication Systems, Yao, K., IBID.

5. Spread Spectrum and Time Decision Multiple Access Satellite Communications, Mohanty, N. C., IBID.

6. Industry and Computers, Shuey, R. L., IEEE Transactions on Manufacturing Technology, December 1975.

7. Spread Spectrum Receiver Using Surface Acoustic Wave Technology, Milstein, L. B. and Das. P., IEEE Transactions on Communications, August 1977.

8. Communications Technology: 25 Years in Retrospect, Andrew, F. J., IEEE Communication Society Magazine, January 1978.

9. Tutorials on Signal Processing for Communications, Mina, K. V., Lawrence, V. B. and Werner, J. J., Tewksbury, S. K., Kreburtz, R. B., Thompson, J. S., and Verna, S. P., IBID.

10. Digital Signal Processing, Oppenheim, A. V. and Schafer, R. W., Prentice Hall, 1975.

11. "Sub-baud coding", Proceedings of the 13th Annual Allerton Conference on Circuit and Systems Theory, Massey, J. L. and Uhran, J. J., October 1975.

12. "Satellite communications to mobile platforms", Proceedings of the IEEE, Lebow, I. L., Jordan, K. L. and Drouilhet, P. R., February 1971.

13. Spread Spectrum Systems, Dixon, R. C., New York: Wiley, 1976.

14. Spread Spectrum Techniques, Dixon, R. C. (editor), IEEE Press, 1976.

15. "Multiple access to a hard-limiting communication-satellite repeater", Aein, J. M., IEEE Transactions on Space Electronics and Telemetry, December 1964.

16. "Multiple access to a communication satellite with a hard-limiting repeater - Volume II: Proceedings of the IDA multiple access summer study", Aein, J. M. and Schwartz, J. W. (editors), 1965.

17. "A comparison of pseudo-noise and conventional modulation for multiple-access satellite communications", Blasbalg, H., July 1965.

18. "Multiple access to a communication satellite with a hard-limiting repeater -- Volume I: Modulation techniques and their applications", Kaiser, J., Schwartz, J. W. and Aein, J. M., January 1965.

19. "Modulation techniques for multiple access to a hard-limiting satellite repeater", Schwartz, J. W., Aein, J. M. and Kaiser, J., May 1966.

20. Unique Word Detection in Digital Burst Communication, IEEE Trans. on Commun., Schrempp, W., Sekimoto, T., August 1968.

21. Design of a Satellite Time-Division Multiple Access Burst Synchronizer, IEEE Trans. on Commun., Gabbard, O. G., August 1968.

Acknowledgment

Special mention must be made of the fine work done by M. Pursley and N. Mohanty in their papers (2,5), major excerpts of which are used herein to describe SSMA and TDMA systems.

DISCUSSION

J.T.Martin, UK

When comparing Microprocessors with Mini Computers you stated that a disadvantage of Microprocessors was an increased Development Cost. Would you please expand on this statement.

In the UK Microprocessors exist which have available high level languages (e.g. CORAL) which are fully transportable. This means that Micros can be developed in the same way as Minis. For small jobs where high level language is not used it is perhaps unfair to compare Micros with Minis as, in this case, Micros should be compared with hard wired logic.

Author's Reply

The development costs referred to include both hardware and software. The software units are becoming more a part of the total costs and result in increased development costs due to (1) the rapidity of technological change (software requires updating) and (2) the general lack of printability when written in machine or assembler language needed to attain speed.

The US manufacturers of course all have compilers for high level languages (FORTRAN, BASIC etc.) but this is not the context in which I described the use of the microprocessor. I agree with your last statement and showed that based on a recent survey 62% of applications of microprocessors are used to replace hard wired logic.

J.Majus, Ge

In the paper you put emphasis on Spread Spectrum Techniques. Especially in military systems, Spread Spectrum has been used for channel security. On the other hand there is a trend from Spread Spectrum Techniques to Cryptic Encoding Methods. What are the further advantages of Spread Spectrum Techniques?

Author's Reply

Originally, spread spectrum was used to effectively communicate in the presence of a jammer (hostile environment). The discussions in my paper of spread spectrum deal with multiple access, a multiple user configuration. The trend you speak of, is a long standing one, in my opinion, and similarities exist between encrypting methods and spread spectrum technology; but more cannot be said due to classification. Certainly, PN sequences have an inherent property of encryption as a consequence of their characteristics.

C.E.Tate, UK

Future systems complexity is increasing very fast and we have the choice of using standard VLSI chips, Micros, or custom VLSI. In each case CAD, logic and test simulations are needed, but current software falls far short of the necessary capability for complex circuits. Suites such as D.LASAR, TEGAS even are inadequate. The problem is that technology is moving faster than the CAD progress and renders software suites obsolete. Do you see this as a serious bottleneck?

Author's Reply

I do not regard it as a serious bottleneck, but as the price we pay for a rapidly advancing technology.

Modem Télégraphique à étalement de spectre

D. BRISSET - G. AUGER
TH-CSF Division Télécommunications
16, rue du Fossé Blanc - 92231 GENNEVILLIERS

1 - INTRODUCTION

Le modem télégraphique à étalement de spectre objet de cette présentation a été réalisé dans le cadre du développement d'une station expérimentale de télécommunications par satellite, entrepris par le Centre Electronique de l'Armement (CELAR). Cette station est en cours d'évaluation dans une expérimentation mettant en oeuvre le satellite SYMPHONIE.

L'information transmise est un rythme télégraphique à 75 bits/s, utilisant un codage différentiel suivi d'une modulation PSK bi-phase. L'étalement de spectre est réalisé dans une bande de plusieurs dizaines de MHz au moyen d'une séquence PN à rythme élevé.

Par rapport aux autres techniques d'accès multiple (FDMA, TDMA, sauts de fréquence) les avantages de cette transmission, de type SSMA sont : l'accès multiple sans obligation de gestion centralisée du réseau, la discrétion, la résistance au brouillage, une complexité modérée des équipements impliqués dans la station sol et à bord du satellite. En contrepartie la capacité de transmission du réseau (nombre d'utilisateurs simultanés se partageant la bande d'étalement W) est inférieure à celle des autres modes d'accès multiple. Le tableau de la figure 1 résume ces points de comparaison.

2 - RAPPEL DES PRINCIPES DE BASE

Le schéma de principe de la liaison est représenté sur la figure 2.

- A l'émission, l'information à transmettre, codée en mode différentiel, est additionnée modulo 2 à une suite PN de débit L fois supérieur, de périodicité L' chips. Le spectre émis est indiqué sur la figure 3, il présente trois propriétés remarquables :

(1)
- étalement dans la bande W = $\frac{2L}{T}$
- pseudo période $\frac{W}{2L'}$,
- densité de puissance réduite dans le rapport L = $\frac{WT}{2}$

- A la réception l'étalement de spectre est supprimé dès le premier mélange utilisant un hétérodyne modulé par une réplique synchrone de la suite PN utilisée à l'émission.

Une boucle de verrouillage de code (BVC) assure l'acquisition et le maintien du synchronisme des deux séquences PN. Le signal comprimé est filtré à bande étroite (bande B_1) et démodulé de manière cohérente, au moyen d'une porteuse reconstituée par la boucle de verrouillage de porteuse (BVP).

Les propriétés de la liaison résultent directement de cette description :

- Discrétion : Tout récepteur ne possédant pas la réplique du code PN doit se fixer de démoduler les chips et pour cela disposer d'un rapport énergétique $\frac{WT}{2}$ fois supérieur à celui de la liaison utile.

- Résistance au brouillage : L'opération de compression du signal reçu effectuée par le récepteur, provoque l'étalement d'un brouilleur étroit, avec les caractéristiques citées en (1).

  Si ce brouilleur est distant de $\Delta\omega$ de la porteuse SSMA, la puissance perçue par le démodulateur dans la bande adaptée (1/T Hz) est réduite dans le rapport :

$$\frac{WT}{2}\left(\frac{x}{\sin x}\right)^2 \qquad x = \frac{\Delta\omega}{W}$$

  Plus généralement, pour un brouilleur de caractéristiques quelconques, le spectre perçu par le démodulateur s'obtient par la convolution du signal reçu et de l'hétérodyne de compression puis filtrage dans la bande 1/T.

../..

3 - CAPACITE DU RESEAU

3.1 - En l'absence de brouillage

Soient les paramètres suivants :

- M : Nombre de communications simultanées dans le réseau
- N_o, N_J : Densités de bruit aux entrées du récepteur et du satellite
- P : Puissance d'une communication à l'entrée du satellite
- P_R : Puissance totale reçue au sol
- K_1/WT : Coefficient de corrélation des codes.

Après compression, le rapport $\frac{C}{N}$ dans la bande $\frac{1}{T}$ s'écrit :

$$(2) \quad \frac{C}{N} = \frac{\frac{P_R}{N_o} T}{M + \frac{N_o W}{P}\left(1 + \frac{P_R}{N_o W}\right) + \frac{K1(M-1)P_R}{N_o W}}$$

M s'extrait aisément de cette formule. Sur la figure 4 on a représenté ses variations en fonction de W avec les valeurs numériques de la liaison expérimentale, K_1 étant pris comme paramètre.

3.2 - En présence de brouillage

En présence d'un brouilleur de puissance J_B, la formule (2) se récrit en remplaçant le terme N_{JW} par $J_B + N_J W$. On peut alors exprimer le paramètre $\frac{J_B}{P}$ représentatif de la résistance au brouillage de la liaison. La figure 5 établie pour $K_1 = 1$ et $M = 10$ et 100 liaisons simultanées montre que $\frac{J_B}{P}$ passe par un optimum en fonction de W.

4 - FONCTIONNEMENT DU MODEM EN PRESENCE DES IMPERFECTIONS DE LA CHAINE DE TRANSMISSION

Les imperfections inévitables de la chaîne de transmission sont :

- Des écarts de fréquence (dérives, effet Doppler)
- Des bruits de phase (transpositions)
- Des distorsions de phase et d'amplitude.

Les contraintes résultantes sur les circuits du démodulateur sont examinées ci-après :

4.1 - Boucle de verrouillage de code (BVC)

La BVC utilise en phase poursuite un corrélateur actif de type 1 Δ (Réf. 1) choisi pour sa gigue de phase $\frac{\sigma_T}{\Delta}$ minimale :

$$\frac{\sigma_T}{\Delta} = \sqrt{\left(\frac{N_o}{2C} + \frac{B_1 N_o^2}{C^2}\right) B_n}$$

Δ = Durée du chip

B_n = Bande de bruit de la BVC

Erreurs de fréquence : Les erreurs de fréquence à considérer sont :

Les écarts Doppler : $\Delta f_1 < \pm 0{,}5$ kHz (porteuse)

Δf_2 (fréquence d'étalement)

L'imprécision des fréquences d'étalement : Δf_3

Dans la mise en oeuvre opérationnelle envisagée de la liaison, les autres erreurs de fréquence peuvent être considérées comme négligeables.

L'écart Doppler porteuse conduit à choisir $B_1 = 1$ kHz ; pour C/No = 30 dBHz et $\frac{\sigma_T}{\Delta} = 0{,}15$, on obtient $B_n = 13$ Hz.

En phase d'acquisition, la recherche du synchronisme entre les deux séquences PN est effectuée en décalant le rythme de la séquence locale de R_{Hz}. La vitesse de recherche effective est en réalité comprise entre $R - |\Delta R|$ et $R + |\Delta R|$, avec $\Delta R = \Delta f_2 + 2\,\Delta f_3$.

La plage de capture en fréquence de la boucle 1 Δ est égale à 0,95 Bn (Réf. 1) ; on doit donc respecter l'inégalité :

$$R + |\Delta R| < 0{,}95\, B_n = 12 \text{ Hz}$$

Pour un $|\Delta R|$ de 5 Hz ceci implique $R < 7$ Hz. La vitesse minimum de balayage est alors $R - |\Delta R| = 2$ chips/s.

Pour améliorer ce résultat et minimiser ainsi la durée de la phase d'acquisition on utilise pour cette phase une boucle de type 2 Δ (Réf. 2). Ce type de boucle possède une plage de capture en fréquence plus élevée et l'on est de plus autorisé à élargir légèrement sa bande de bruit au détriment de la gigue de phase pendant la recherche. Pour le même $|\Delta R|$ de 5 Hz, la vitesse minimum de balayage est égale à 20 chips/s, soit dix fois plus élevée qu'avec la boucle 1Δ.

../..

La durée de la phase d'acquisition, pour une séquence PN de longueur L' est alors :

$$Ta < \frac{L'}{R - |\Delta R|}$$ Soit par exemple Ta < 2,5 s pour L' = 500 chips

Distorsions : on a étudié par simulation les pertes de corrélation par rapport au cas idéal résultant :

- du filtrage sans distorsion des lobes secondaires, le spectre émis étant alors réduit à la bande $\pm \frac{W}{2}$.

  La perte de puissance est égale à 0,8 dB au niveau du corrélateur.

- du filtrage sans distorsion du lobe principal, le spectre émis étant réduit à la bande $\omega_0 \pm \frac{W}{4}$.

  La perte de puissance dans le corrélateur est égale à 2 dB.

- d'une distorsion du temps de propagation de groupe appliquée au lobe principal. Les distorsions rencontrées se manifestent principalement en bord de bande. On a en particulier simulé une variation de la forme :

$$\Delta\tau(\omega) = \frac{a}{W} \sin \frac{b(\omega-\omega_0)}{W}$$

  qui évoque la distorsion rencontrée dans les TOP et dans les filtres simples.

  Avec a = b = 2 on a calculé une perte de 1,2 dB.

4.2 - Boucle de verrouillage de porteuse (BVP)

La boucle de verrouillage de porteuse opère sur le signal comprimé, doublé en valeur absolue et limité.

Deux phénomènes interviennent au niveau du démodulateur, qui sont la gigue de phase de l'oscillateur asservi et, résultants de celle-ci, les glissements de phase occasionnels d'amplitude égale à π.

4.2.1 - La gigue de phase ε^2 (rd/S)^2 de la porteuse reconstituée a deux origines :

- le bruit de phase de l'oscillateur asservi : $2\Delta\omega$ étant la largeur à 3 dB du spectre de l'oscillateur, le carré de l'écart type de l'erreur de phase correspondante est :

$$\varepsilon_1^2 = \frac{3}{4} \quad \frac{\Delta\omega}{2BL}$$ 2 BL : bande de bruit de la BVP

- le bruit additif gaussien du récepteur : $\varepsilon_2^2 = \frac{N_0}{2C} 2B_L$

La gigue totale $\varepsilon^2 = \varepsilon_1^2 + \varepsilon_2^2$ peut être minimisée en choisissant 2 BL telle que :

$$2BL = \sqrt{\frac{3C\Delta\omega}{2N_0}} \qquad \varepsilon^2 = \sqrt{\frac{3}{2} \frac{\Delta\omega}{C/N_0}}$$

Pour C/No = 30 dBHz et $\Delta\omega$ = 0,3 rd/S on obtient ε^2 = 0,02 (rd/S) et 2 BL = 20 Hz.

4.2.2 - Les glissements occasionnels de phase d'amplitude π entraînent autant d'erreurs au niveau du décodeur différentiel. Le temps moyen T_{AV} entre deux glissements est :

$$T_{AV} = \frac{1}{B_L} e^{\frac{\pi}{2\varepsilon^2}}$$

On en déduit le taux d'erreur résultant :

$$\frac{T}{T_{AV}} = T B_L e^{-\frac{\pi}{2\varepsilon^2}}$$

Avec les valeurs calculées précédemment ce taux d'erreur est parfaitement négligeable. Toutefois il convient de noter qu'il varie très vite avec ε^2, qui doit donc être minimisé. Par exemple avec ε^2 = 0,15 il serait égal à 10^{-5} et ne pourrait plus être négligé.

4.3 - Démodulation cohérente

Les abaques habituelles fournissant le taux d'erreurs en fonction du rapport énergétique E/No à l'entrée du démodulateur doivent être utilisées en tenant compte de l'erreur de phase de la porteuse. Il faut considérer que ε s'ajoute quadratiquement à l'écart type de l'erreur de phase du signal incident dont la valeur est $\sqrt{\varepsilon^2} = \sqrt{\frac{N_0}{2E}}$.

Le rapport énergétique auquel s'appliquent les abaques est donc :

$$\left(\frac{E}{N_0}\right)_{Eq} = \frac{E/N_0}{1 + 2\varepsilon^2 E/N_0}$$

Avec ε^2 = 0,02 et E/No = 12 dB on calcule une perte inférieure à 2 dB permettant d'obtenir un taux d'erreurs inférieur à 10^{-4}.

5 - CONCLUSIONS -

Les caractéristiques du modem télégraphique à étalement de spectre décrit ci-dessus permettent de tenir compte des imprécisions de fréquence et des bruits de phase présents dans la chaîne de transmission.

Les résultats expérimentaux font apparaître un taux d'erreur inférieur à 10^{-4} pour un $\frac{C}{No}$ de 30 dBHz à l'entrée du récepteur.

L'optimisation de la vitesse de recherche en phase d'acquisition a conduit à utiliser deux types de corrélateurs commutés. Par exemple l'acquisition d'un code de 500 chips est obtenue en moins de 25 s avec une probabilité de non acquisition inférieure à 10^{-4}.

Ce matériel devrait permettre d'assurer plusieurs dizaines de communications simultanées en présence de brouilleurs de niveaux très supérieurs à ceux des signaux utiles.

BIBLIOGRAPHIE

1) - Walter J. GILL
A comparison of binary delay-lock tracking loop implementations
IEEE Transactions on Aerospace and electronic systems
Vol. AES-2, N° 04, July 1966

2) - P. TOLSTRUP Nielsen
On the acquisition behaviour of binary delay-lock loops
IEEE Transactions on Aerospace and electronic systems
May 1975

3) - Floyd M Gardner
Phaselock techniques.

Qualités	FDMA	TDMA	F H	SSMA
Emploi de la bande du répéteur	+	+ +	+	- -
Emploi de la puissance du répéteur	-	+ -	-	+ -
Capacité	+	+ +	-	- -
Gestion du réseau	-	- -	- -	+ +
Complexité de la station sol	+ +	- -	- -	+
Vulnérabilité	-	- -	+	+ +
Discrétion	- -	- -	-	+ +

\+ avantage
\- inconvénients
} selon le critère de comparaison

Fig. 1 Comparaison des procédés d'accès multiple

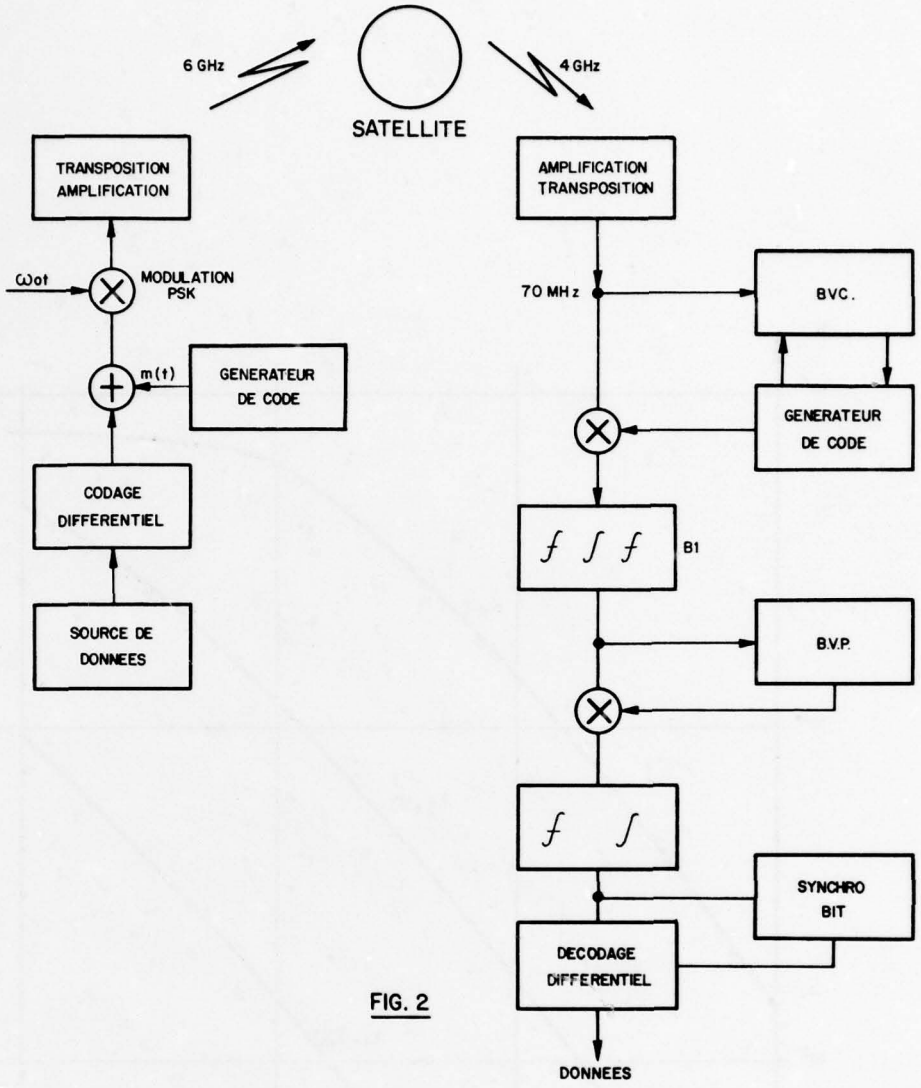

6 GHz
4 GHz
SATELLITE
TRANSPOSITION AMPLIFICATION
ωot
MODULATION PSK
m(t)
GENERATEUR DE CODE
CODAGE DIFFERENTIEL
SOURCE DE DONNEES
AMPLIFICATION TRANSPOSITION
70 MHz
B.V.C.
GENERATEUR DE CODE
B1
B.V.P.
SYNCHRO BIT
DECODAGE DIFFERENTIEL
DONNEES

FIG. 2

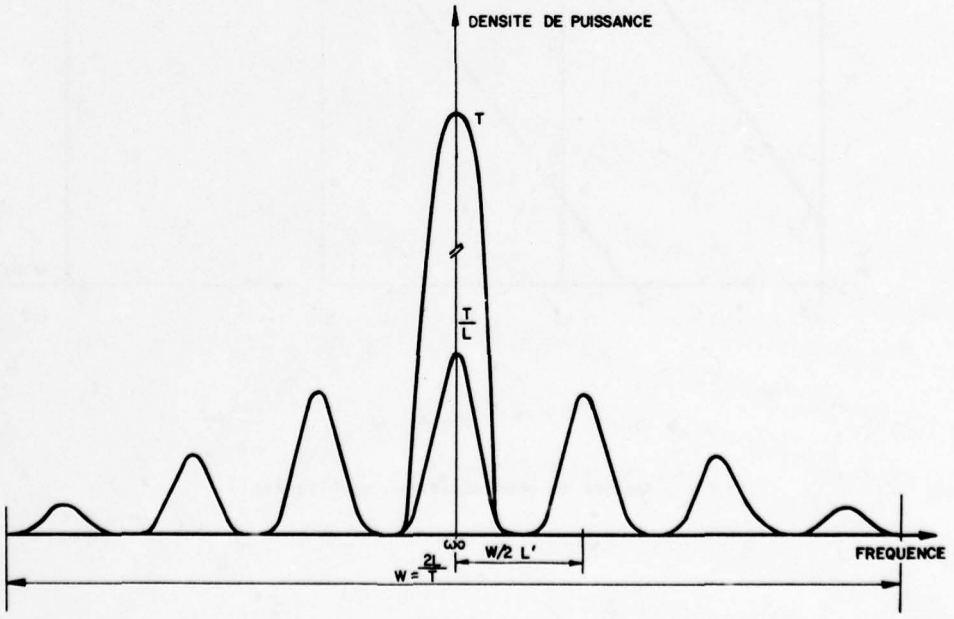

DENSITE DE PUISSANCE
T
T/L
ωo
W/2 L'
W = 2/T
FREQUENCE

FIG. 3

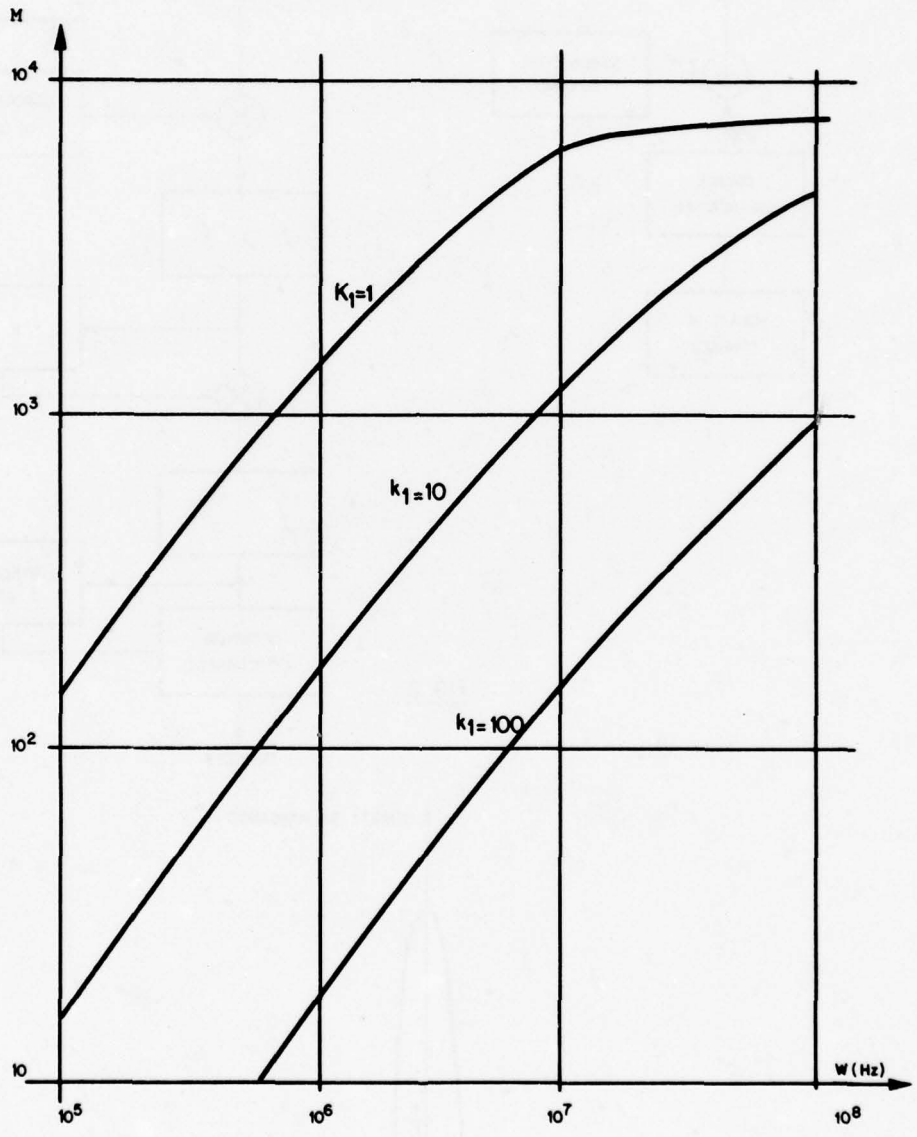


Fig. 4

Nombre de communications simultanées

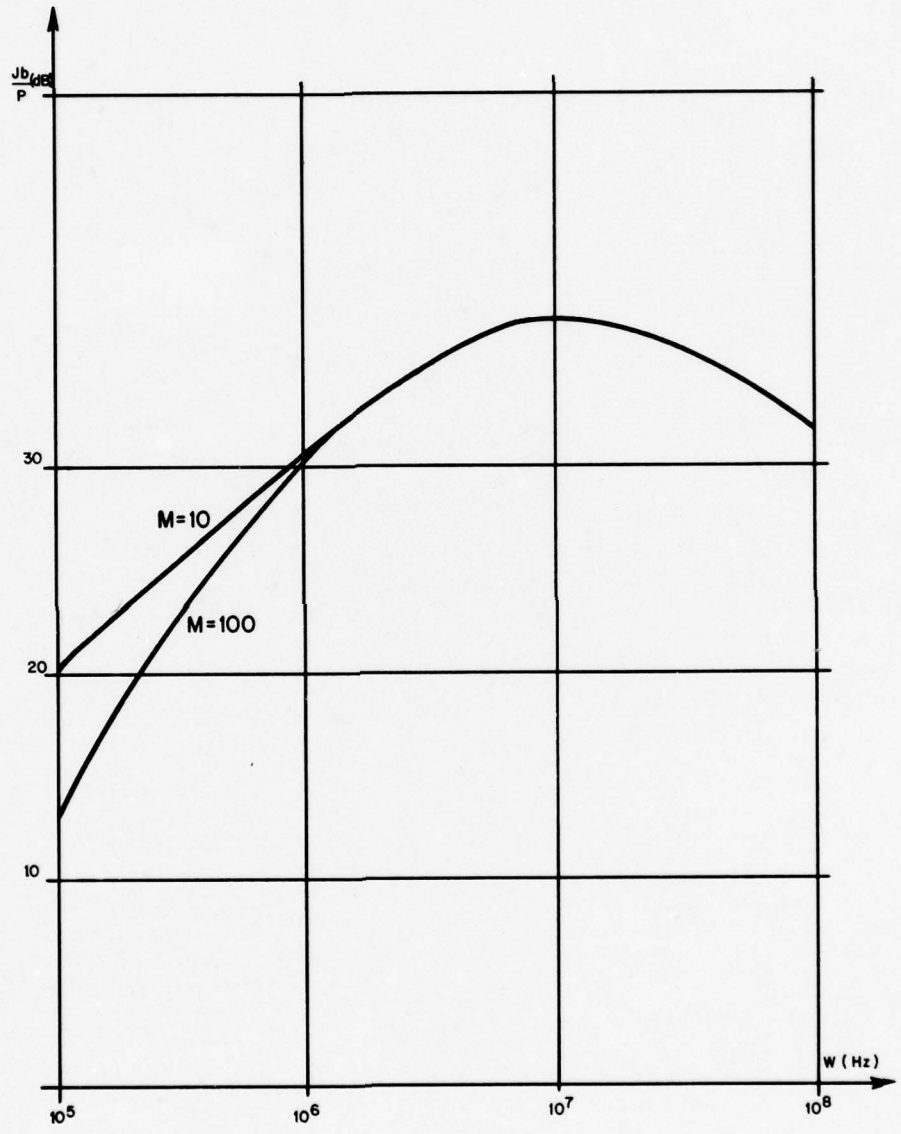


FIG. 5

Protection contre le brouillage

THE PERFORMANCE OF CODE DIVISION MULTIPLEXING WITH PULSE POSITION MODULATION

Jürgen Lindner

Institut für Elektrische Nachrichtentechnik,

TH Aachen, 5100 Aachen, W.-Germany

SUMMARY

Code division multiplexing (CDM) became known as a means to share a given communication channel by more than one user, alternatively to time and frequency division multiplexing. In this paper pulse position modulation (PPM) is considered for CDM instead of the more common PCM. The goal is, to find attractive alternatives to PCM for the transmission of analog data over a CDM system

For this purpose the relations between the output signal to interference ratio, the number of users and the bandwidth on the channel are calculated and an attempt is made to optimize the binary sequences used as "carrier functions".

Binary sequences are presented having autocorrelation functions (ACF) which are relatively good approximations to the optimal ones. By means of some known results it is shown that the crosstalk properties, given by the crosscorrelation functions (CCF) of the binary sequences, are fixed.

After a theoretical comparison of CDM/PPM with CDM/PCM on the basis of these results, a PPM receiver is presented, leading to low cost and effort necessary for the implementation of a CDM/PPM system, e.g. for speech transmission. One very attractive feature is the absence of any synchronization within the system, similar to frequency division multiplexing with AM and incoherent reception, but with the properties of CDM and the resulting advantages in some applications.

1. INTRODUCTION

Spread spectrum techniques have been known for many years [1], but only for a much shorter time the components for building systems with reasonable size and cost have been available.

Message privacy, the immunity to jamming, the possibility of ranging by pulse compression and the possibility of uncomplicated multiplexing with random access to a nonlinear channel are very attractive features for military applications. Hence the first systems built fall into this category [2]. On the other hand cost and complexity are still high enough today to prevent many applications in the non-military field. For this reason some investigations have been made with the intention to simplify the correlation receivers (see e.g. [3]) and to find optimal signals (see e.g.[4,5]). In the following only the multiplexing properties of spread spectrum are considered and the starting point is a special case of the well known "direct sequence" spread spectrum technique. For this special case, called code division multiplexing (CDM),in [6] an attempt was made to find sets of optimal binary sequences for a binary data transmission over such a CDM system. The results are used in this paper and therefore they are briefly described in the next section.

To transmit speech signals in this way over a common channel, the signals must be sampled, quantized and converted to binary numbers, i.e. PCM must be applied. Moreover bit and word synchronization are necessary for each transmitter/receiver pair and the frequency band occupied on the channel seems to be relatively large. Because of these disadvantageous properties another modulation scheme will be examined here for use in CDM, i.e. pulse position modulation (PPM).

A similar CDM/PPM system was analyzed already in 1965 [7] but only with fixed parameters. In the following the optimization results of [6] shall be applied to CDM/PPM to find relations between the output signal to interference ratio and the number of active multiplexing subscribers, the effectiveness in use of bandwidth and the influence of the assumed set of binary sequences on these quantities. A comparison of CDM/PPM with CDM/PCM together with some realization considerations and a conclusion in the last part of this paper may give some hints for possible applications of CDM with PPM.

2. KNOWN RESULTS RELATED TO CDM

Fig. 1 shows a simple model of a CDM system, using binary "carrier functions" for each information bit. These carrier functions consist of N subpulses of equal duration and can therefore be represented by binary sequences of length N with elements +1 or -1.

Multiplexing is performed by assigning a different carrier function to each transmitter and random access means that the transmitters operate asynchronous, i.e. there is no common time base for the transmitters.

In [6] this system is considered with respect to the set of carrier functions or binary sequences which are necessary to achieve a bit error probability as small as possible for each transmission link. In the following

$$\underline{x} = (x_1, x_2 \ldots, x_N) \quad ; \quad x_i \in \{-1, 1\}$$

is a binary sequence of length N belonging to a carrier function and

$$\varphi_{xy}(k) = \begin{cases} \sum_{i=1}^{N-k} x_i \cdot y_{i+k} & ; \quad 0 \leq k \leq N-1 \\ \sum_{i=1}^{N-k} y_i \cdot x_{i+k} & ; \quad -(N-1) \leq k < 0 \\ 0 & ; \quad \text{otherwise} \end{cases} \tag{1}$$

is the aperiodic crosscorrelation function (CCF) of two sequences $\underline{x}$ and $\underline{y}$. For $\underline{y} = \underline{x}$ this CCF becomes the autocorrelation function (ACF) $\varphi_{xx}(k)$. For short the CCF of two sequences $\underline{x}_k$ and $\underline{x}_j$, taken from the set of carrier functions, is abbreviated by $\varphi_{kj}(l)$ in the following.

The interference for receiver k in Fig. 1 is given by a sum of shifted, overlapping CCFs $\varphi_{kj}(l)$, $j \neq k$, weighted by +1 or -1 according to the data bit of the j-th transmitter. If the clock rates for the generation of the carrier functions for all transmitters differ a little from each other and if the number of active subscribers is greater than 4 or 5, the interference is nearly gaussian distributed with zero mean. Therefore only the second moment or the power R determines the bit error probability P_e for the k-th receiver:

$$P_e = \frac{1}{2} \operatorname{erfc} \sqrt{\frac{S}{2R}} \tag{2}$$

S is the signal power at sampling instants and erfc is the error function complement. One result of [6] is that the crosstalk power R for many sets of binary sequences is given by

$$R \approx \frac{4}{3} L \, \overline{\hat{\varphi}^2_{kj}(l)} \tag{3}$$

$\overline{\varphi}$ means averaging over all l ($-N+1 \leq l \leq N-1$) and $\hat{}$ means averaging over the set with respect to j ($j = 1,2,\ldots,L+1$; $j \neq k$). $L+1$ is therefore the number of active multiplexing subcribers. Another formula derived in [6] gives a connection to the ACFs:

$$\overline{\varphi^2_{kj}(l)} \approx \frac{N}{2} + \frac{1}{N} \sum_{i=1}^{N-1} \varphi_{kk}(i) \cdot \varphi_{jj}(i) \tag{4}$$

It can be shown that the second term is approximately equal to zero if the average $\hat{}$ is taken over large sets of binary sequences. Therefore sets of random sequences of length N give a mean crosstalk power of

$$R \approx \frac{2}{3} LN \tag{5}$$

The signal to interference power ratio, abbreviated by SIR in the following, is with $S = N^2$ for these sets:

$$SIR \approx \frac{3N}{2L} \tag{6}$$

Fig. 1 shows the necessity of synchronization for every transmitter/receiver pair. To achieve this, the ACF of the carrier function must have small "sidelobes". This means that the values $\varphi_{kk}(i)$ are small for every $i \neq o$ compared to $\varphi_{kk}(o) = N$. Eq. (4) shows that for sets of sequences with so-called "good" ACFs, i.e. with small sidelobes, eq. (6) is also true. From now on a set of binary sequences is assumed, having an SIR which is exactly equal to 3N/2L. This may be justified by the fact that eq. (6) is derived by averaging over 2N-1 CCF values and L CCFs.

For eq. (6) only the crosstalk interference was considered. If additional white gaussian noise of (two sided) spectral power density R_o is assumed on the channel eq. (6) must be modified to

$$SIR = \frac{3}{2} \, \frac{N}{L + \frac{3}{2} \left(\frac{E}{R_o}\right)^{-1}} \tag{7}$$

E is the energy for one subpulse of the carrier function: $E = A^2 t_o$ if A is the amplitude and t_o the duration of one subpulse of the carrier function. To get the same simple form as eq. (6), the quantity $L + \ldots$ in the denominator of eq. (7) is abbreviated by L_{eff} in the following.

3. PPM TRANSMISSION

3.1. The model

Fig. 2 shows the model of the PPM system assumed throughout this paper (with the exception of section 6). The incoming signal is sampled in the transmitter and every sample shifts a carrier function according to the sample value. Positive values mean a shift to the right with respect to a reference point and negative sample values mean a shift to the left with

respect to this point. The receiver consists of a matched filter (MF), giving at its output the ACF of the carrier function with the same shift as that produced by the transmitter, but superimposed by interfering signals. After comparing with a threshold a spike occurs and the distance between this spike and the reference point gives the sample being transmitted.

In the following it will be assumed that the samples can take 2b + 1 discret levels between -b and b and that the interference at the MF output is gaussian with zero mean and variance R. The 2b + 1 discret samples give also 2b + 1 discret shifts in time and the receiver must decide at these 2b + 1 instants, wether the treshold ist exceeded or not. Fig. 3 illustrates this. The continuous time ACF, as shown in this figure, can be replaced by the discrete time ACF $\varphi_{xx}(k)$ of the binary sequence $\underline{x}$ which represents the carrier function (in this section only one carrier function or sequence $\underline{x}$ is considered; therefore the index k of $\underline{x}_k$ is omitted).

Now the output signal to interference ratio of the receiver shall be calculated and its dependance on the ACF and the threshold c. For this purpose it is reasonable to consider first the error probability P_e.

3.2. Error probabilities

An error occurs, either if the threshold is exceeded at false instants or if it is not at the right instant. The probability that the threshold is exceeded at instant i is

$$Q_i(k) = \frac{1}{2}\,\mathrm{erfc}\,[(c - \rho_{xx}(i-k))\sqrt{\tfrac{1}{2}\,\mathrm{SIR}}\,] \quad ; \quad i \neq k \tag{8}$$

This probability depends on the actual sample k ($-b \leq k \leq b$) and on the sidelobes $\rho_{xx}(j)$, $j \neq 0$ of the normalized ACF ($\rho_{xx} = \varphi_{xx}/N$). SIR is the signal to interference ratio valid for the main peak of the ACF. The threshold c can be between 0 and 1. If

$$P_o = \frac{1}{2}\,\mathrm{erfc}\,[(1-c)\sqrt{\tfrac{1}{2}\,\mathrm{SIR}}\,] \tag{9}$$

is the probability that the threshold is not exceeded at the right instant (sometimes called probability of false detection), then the error probability P_e is given by

$$P_e = \frac{1}{2b+1} \sum_{k=-b}^{b} P_{ek}$$

$$P_{ek} = 1 - (1-P_o) \prod_{\substack{i=-b \\ i \neq k}}^{b} (1-Q_i(k)) \tag{10}$$

If the maximum value of the normalized ACF sidelobes is m, then

$$Q_i(k) \leq \frac{1}{2}\,\mathrm{erfc}\,[(c-m)\sqrt{\tfrac{1}{2}\,\mathrm{SIR}}\,] = Q \tag{11}$$

Therefore

$$P_e \leq 1 - (1-P_o)\cdot(1-Q)^{2b} = P_{emax} \tag{12}$$

P_{emax} is not a function of k but only of c, m and SIR.

Supposing that <u>all</u> ACF sidelobes are equal to zero, then the PPM model considered here is identical with an 2b + 1 - ary transmission system using orthogonal functions. It can be shown (see e.g. [8]) that this kind of data transmission approximates the Shannon bound for a channel disturbed by white gaussian noise if the channel bandwidth and 2b+1 go to infinity and if "greatest of" decision is made instead of threshold decision.

3.3. Signal to interference ratio at the output

The interference power R_a at the output of the receiver cannot be calculated easily, but a simplifying assumption leads to an upper bound for R_a. The assumption is that all events leading to an error are mutually exclusive. For a single reception operation either one positive threshold crossing at a false instant can occur or a negative crossing at the right instant. This leads to

$$R_a \leq \sum_{i=1}^{b} i^2 Q_i + P_o\, S \tag{13}$$

Q_i are the probabilities of exceeding the threshold at instant i for the ACF with zero shift, i.e. $Q_i = Q_i(0)$, and S is the signal power at the receiver output. From eq. (13) follows a bound for the signal to interference power ratio SIR_a at the output of the receiver

$$SIR_a \geq \frac{1}{P_o + \frac{1}{S}\sum_{i=1}^{b} i^2 Q_i} \tag{14}$$

If $c \geq 0.7$; $m \leq 0.1$ and $SIR \geq 17$ dB then it can be shown by using eq. (12) that the errors are only produced by false detection, i.e. $P_e \approx P_o$ and eq. (14) can be simplified further:

$$SIR_a \approx \frac{1}{P_o} \quad , \text{ if } \quad c \geq 0.7 \quad , \quad SIR \geq 17 \text{ dB} \tag{15}$$

It is remarkable that in this equation the SIR_a depends not on the signal power S. Usually SIR_a is a function of S so that a decreased power S gives also a decreased SIR_a.

3.4. Optimum autocorrelation functions

Now the upper bound for R_a given by eq. (13) is minimized to get an optimum ACF for the discussion of the binary sequences which are of interest here. Of course, this optimum does only exist if a constraint is introduced. If the constraint is for simplicity

$$J = \frac{1}{b} \sum_{i=1}^{b} \rho_{xx}(i) = \bar{\rho} = \text{const.} \tag{16}$$

and if the functional to minimize is given by eq. (13), i.e.

$$I = \sum_{i=1}^{b} i^2 \, Q_i + P_o \cdot S \tag{17}$$

then the optimization can be carried out simply. Eq. (16) says that the sidelobe mean of the normalized ACF $\rho_{xx}(i)$ has the given value $\bar{\rho}$ within the decision window. The resulting optimal ACF sidelobes are given by

$$\rho_{xx}(i) = c - \frac{2}{\sqrt{SIR}} \sqrt{\ln \frac{i}{i_o}} \tag{18}$$

Fig. 4 shows these values versus i for b = 64 and optimal c . The value i_o must be choosen to satisfy the constraint (16) . Obviously the SIR is an important quantity. For SIR greater than 20 dB or 21 dB the optimum ACF tends to an ACF with a constant sidelobe level of about $\bar{\rho}$, if $\bar{\rho}$ is the sidelobe mean given by eq. (16). If SIR is about 16 dB or less, then the sidelobes differ considerably from the constant level ACF. How far this can influence the interference power R_a or the SIR_a , respectively, depends on the threshold c. This can be seen if eq. (18) is put into eq. (14) with regard to eq. (8):

$$SIR_{aopt} \geq \frac{1}{P_o + \frac{1}{S} \sum_{i=1}^{b} i^2 \, \frac{1}{2} \operatorname{erfc} \sqrt{2 \ln \frac{i}{i_o}}} \tag{19}$$

In this lower bound only P_o depends on c and a decrease of c means a decrease of P_o which increases the lower bound and therefore SIR_a. Fig. 4 and eq. (18) show the resulting difficulties: the given $\bar{\rho}$ is a function of c and if c is decreased, the ACF sidelobes are shifted to negative values and the ACF may become unrealistic. Another problem arises if one considers the ACF of binary sequences. For useful sequences $\underline{x}$ the height of the ACF sidelobes does not fall monotonously as shown in Fig. 4 and in addition to this the sidelobe mean $\bar{\rho}$ is about equal to zero. On the other hand the optimum ACF for $\bar{\rho} = 0$ cannot be approximated by ACFs of binary sequences. The problem can be solved if one considers absolute values of $\rho_{xx}(i)$ which must be done in any case if envelope detection is applied.

Fig. 5 demonstrates the differences between the lower bounds for SIR_a if different classes of binary sequences are used. For curve a) an optimum ACF was taken for every value of SIR (see Fig. 4) and the normalized sidelobe mean was set to $\bar{\rho}=0.04$ which is a reasonable value for b=64. The same value $\bar{\rho}$ was taken to form a constant-sidelobe level ACF, i.e. $\rho_{xx}(i)=\bar{\rho}$, $i \neq 0$ (curve b)). It can be seen that this ACF comes relatively close to the optimum one, especially if SIR $\geq$ 20dB. This seems to be the same tendency which has been recognized just before in Fig. 4.

Both ACFs, a) and b) cannot be realized by +1,-1,-binary sequences, but that is no fatal consequence as can be seen by curve d). This curve is derived from an average over the ACFs of the ensemble of all 2^N binary sequences which are possible for the length N=b. It can be shown that the average over all $|\rho_{xx}(i)|$ is approximately given by

$$\widehat{|\rho_{xx}(i)|} \approx \frac{2}{N} \sqrt{\frac{2}{\pi} (N-i)} \quad ; \quad i=1,2,\cdots,N-1$$

The differences to curves a) or b) of about 5dB or 8dB in SIR_a respectively give rise to the question whether binary sequences exist which come closer to the optimal ones. One way to find some, may be to search for the better ones within the ensemble of all 2^N binary sequences, but this can be very expensive, especially if N is, e.g. 64 or 100 or greater. Another way is to use pseudo noise (PN) sequences as done in [7] (without any optimization considerations). But the mean, given by curve d) holds also for PN sequences and there are only a few PN sequences of a given length $N=2^n$; $n=1,2,3,\cdots$ which are really better than d).

In [9] a class of binary sequences is described, which seems to be especially attractive in this connection, because they have their greatest sidelobes within the region of the ACF main peak and they can be constructed very simply by using so-called generalized E sequences. One special example of such E sequence is

$$(\underline{a}, \underline{a}, -\underline{a}^-, \underline{a}^-)$$

$\underline{a}$ may be any binary sequence of length N_1, but the CCF of $\underline{a}$ and the inverse time sequence $\underline{a}^-$ should be as small as possible to get a binary sequence $\underline{x}$ with good properties. $\underline{x}$ is simply formed by concatenation of the subsequences, e.g.:

$$\underline{a} = (1,1,1,-1) \; ; \quad \underline{a}^- = (-1,1,1,1)$$

$$\underline{x} = (1,1,1,-1,\ 1,1,1,-1,\ 1,-1,-1,-1,\ -1,1,1,1)$$

$$\underline{a} \qquad \underline{a} \qquad -\underline{a}^- \qquad \underline{a}^-$$

In this example the resulting sequence $\underline{x}$ has the length $N=4N_1$.

For the "approximation" (curve c)) in Fig. 5, a sequence $\underline{x}$ like the one described above was used, but with a different sequence $\underline{a}$ of length $N_1=16$:

$$\underline{a} = (1,1,1,1,1,-1,-1,1,1,-1,-1,1,-1,1,-1,-1)$$

It can be seen that this approximation is almost as good as the unrealizable constant-sidelobe ACF.

No quantizing noise was taken into account for Fig. 5; this becomes important, if $SIR_a \geq 39dB$ for b = 64 (=> 7 bit).

4. CDM WITH PPM

Now the results of the last section shall be applied to CDM. The interference assumed in the last section is now equal to the crosstalk from the unwanted transmitters. Because it is reasonable here, to assume a set of carrier functions with "good" ACFs, eq. (6) gives the SIR at the instant of the ACF main peak. If additional white gaussian noise of power density R_o is assumed, then

$$SIR = \frac{3N}{2\,L_{eff}}$$

$$L_{eff} = L + \frac{3R_o}{2\,E} \tag{20}$$

The effective number of interfering subscribers L_{eff} can be expressed in terms of the bandwidth expansion

$$\beta = \frac{f_{gm}}{f_{gs}} \tag{21}$$

where f_{gm} is the bandwidth of the multiplexing signal on the channel and f_{gs} the bandwidth of the input signal. If t_o is the duration of one subpulse of the PPM carrier function, it is reasonable to set

$$f_{gm} = \frac{1}{2t_o} \tag{22}$$

With the duration of the whole carrier function $T = N \cdot t_o$ and the sampling period of the input signal $T_s = 1/2\, f_{gs}$, f_{gm} can be expressed by

$$f_{gm} = N \cdot f_{gs} \frac{T_s}{T} \tag{23}$$

The quantity T_s/T is a measure for the pause between two succeeding carrier functions and in the following it is abbreviated by r. The bandwidth expansion is therefore given by

$$\beta = rN \tag{24}$$

This gives with eqs. (20), (21), (22) and $E = A^2 \cdot t_o$:

$$L_{eff} = L + \frac{3R_o \beta f_{gs}}{A^2} \tag{25}$$

The quantity

$$\eta = \frac{L}{\beta} \tag{26}$$

is a measure for the effectiveness in use of bandwidth; η should be as large as possible. If the numerator of eq. (26) was L+1, then η would be at most equal to 1.

η_{eff} is defined similarly

$$\eta_{eff} = \frac{L_{eff}}{\beta} \tag{27}$$

Eqs. (20), (24) and (27) give

$$\eta_{eff} = \frac{3}{2r\,SIR} = \eta + 3\,\frac{R_o \cdot f_{gs}}{A^2} \tag{28}$$

These equations connect the signal to interference ratio to the effectiveness in use of bandwidth if the transmitter amplitude A, the input signal bandwidth f_{gs} and the noise power density R_o are given. If no thermal noise ist present, η_{eff} is equal to η, otherwise η is less than η_{eff}.

Up to now it was assumed that the signal to interference ratio SIR depends on N/L_{eff} as described by eq. (20). This is for sets of carrier functions or binary sequences with "good" ACF (small sidelobes) or for sets of random sequences of length N. It is most likely that eq. (20) is also true for a set of binary sequences with approximatly optimum ACF as described in the last section. Useful sequences $\underline{a}$, $\underline{b}$, ··· are those with small ACF sidelobes as described in [10,11]. A few examples for $N_1=16$:

```
+ +++ --- +-- -+- -+-
+ ++- +++ --- -+- ++-
+ ++- --+ --- +-- +-+
+ --+ --- +-- -++ ++-
+ +-+ -+- -+- --- -++
+ --+ --- +-- -++ ++-
+ -+- -+- --- --+ ++-
+ +-+ --+ +++ --+ +--
```

$+ \triangleq 1$

$- \triangleq -1$

Other examples of generalized E sequences are

$$(-\underline{a},-\underline{a},\ \underline{a},-\underline{a},-\underline{a},-\underline{a},-\underline{a},\underline{a},-\underline{a}^-,-\underline{a}^-,\underline{a}^-,-\underline{a}^-,\underline{a}^-,\underline{a}^-,\underline{a}^-,-\underline{a}^-)$$

$$(\underline{a},\underline{a},\underline{a},\underline{a},\underline{a},-\underline{a},-\underline{a},\underline{a},-\underline{a}^-,-\underline{a}^-,\underline{a}^-,\underline{a}^-,-\underline{a}^-,\underline{a}^-,-\underline{a}^-,\underline{a}^-)$$

5. COMPARISON OF CDM/PPM WITH CDM/PCM

On the basis of the last results PPM is now compared with PCM for the purpose of code division multiplexing with respect to the signal to interference ratio SIR_a at the output of the receiver and the effectiveness in use of bandwidth η. For the PCM case a derivation similar to that of section 4 leads to

$$\eta_{eff\ PCM} = \frac{L_{eff}}{\beta_{PCM}} = \frac{L_{eff}}{K \cdot N} = \frac{3}{2K \cdot SIR_{PCM}} \tag{29}$$

K is the number of bits necessary for each sample including the number of word synchronization bits K_s:

$$K = ld2b + K_s \tag{30}$$

ld is the logarithm to the base 2 and 2b+1 the number of quantizing levels used for the PPM case before. As above a given η_{PCM} leads with eq. (29) to SIR_{PCM} and the SIR_{aPCM} follows by the generally known relations[+]:

$$P_{ePCM} = \frac{1}{2}\ \mathrm{erfc}\sqrt{\frac{1}{2}\ SIR_{PCM}}$$

$$SIR_{aPCM} = \frac{1}{8 \cdot P_{ePCM}} \tag{31}$$

Fig. 6 shows SIR_a for PCM and PPM versus $\eta_{effPCM} = \eta_{effPPM} = \eta_{eff}$. It must be noted that the PPM curves are the lower bounds for SIR_a taken from Fig. 5. The influence of the parameter r is evident. r=1 means no pause between two succeeding unmodulated carrier functions and r=2 means a pause which is equal to the duration T of the carrier function. For PCM a mean value of K_s=1 bit for every word was assumed for synchronization. It can be seen that the differences between PPM and PCM are not very important if the more realistic case r=2 is considered and if one bears in mind that the PPM curves are only lower bounds. The relatively bad utilization of bandwidth,adherent to CDM, can be seen by the absolute values of η. η=1% means that the frequency band occupied by the CDM system is by a factor of $1/\eta$=100 greater than it is e.g. for frequency division multiplexing with SSB-AM. For PCM a factor of K=8 comes from the usual PCM bandwidth expansion and the remaining factor of 12.5 is due to the binary CDM transmission system.

$\eta_{PPM} = \eta_{PCM} = \eta$ can be calculated from eq. (28) and the number of active multiplexing subscribers is

$$(L+1)_{PPM} = \eta \cdot r \cdot N + 1$$

$$(L+1)_{PCM} = \eta \cdot K \cdot N + 1 \tag{32}$$

This is for equal effectiveness in use of bandwidth.

[+] The input samples of the transmitter are assumed to be triangularly distributed between -b and b. The same assumption was made for the PPM case in section 4.

6. REALIZATION CONSIDERATIONS

There are some possibilities for realizing a CDM/PPM system. Because every transmitter/receiver pair is independant of the others, only one transmitter/receiver pair must be considered.

Starting with the transmitter, one sees that there is no problem to build a sampler together with a PPM modulator. This is also true for the generation of the carrier function which must be shifted in time according to the actual sample value.

The receiver needs more careful considerations if the resulting complexity shall be low. The heart of the signal processing is the matched filter (MF). Advances in technology offer relatively simple solutions, e.g. bucket brigade tapped delay lines or complete CCD or SAW correlators. For SAW correlators the input signal must be modulated with a center frequency of about 100 MHz and it is reasonable to apply envelope detection. Then only absolute values of the ACF are important, but all theoretical results of section 4 are just for this case.

Another problem is the reference point in time. This problem can be solved by transmitting only differences of two succeeding samples followed by integration at the output of the receiver. Because the output signal of the MF after the threshold can then considered to be pulse frequency modulated, a phase locked loop may be used to convert the incoming differences of time shifts into a continuous time signal. This signal is fed through a low pass filter and a deemphasis circuit (for integration). Because all parts of this receiver will not have ideal behaviour and because the receiver model assumed in section 4 is not identical to the PLL model proposed here, the achievable SIR_a may somewhat differ from the SIR_a calculated theoretically in section 4.

7. CONCLUSION

PPM is an alternative approach to PCM for sharing a common communication channel by means of code division multiplexing. The theoretical optimization considerations in section 4 showed a relatively weak influence of the ACF on the signal to interference ratio at the output of the PPM receiver. Binary sequences with small ACF sidelobes are useful but up to now it is an unsolved problem to find sequences with smallest sidelobes if the length is 100, 200 or more. The proposed classes of sequences which have their greatest ACF sidelobes in the region of the ACF main peak are just as useful and they are easy to construct from shorter ones.

The effectivety in use of bandwidth is in the region of about one percent or less for PPM as well as for PCM if a reasonable signal to interference ratio at the output of the receiver is desired. One part of the relatively large bandwidth expansion gives a certain insensitivity to noise and jamming and the remaining part is needed for multiplexing.

With PPM one can benefit from the properties of CDM as well as this is for PCM. With PPM the message privacy should be valued somewhat higher, because for CDM/PCM simple receivers can be built which need no information about the code or carrier function used in the transmitter [3]. One important feature of CDM/PPM is the little effort necessary for implementation and the absence of any synchronization if a phase locked loop is used for demodulation.

8. REFERENCES

[1] Dixon, R.C., 1976, "Spread Spectrum Systems", Wiley, N.Y..

[2] Anderson, D.R., Wintz, P.A., 1969, "Analysis of Spread Spectrum Multiple Access Systems with a Hard Limiter", IEEE Trans. on COM-17, pp. 285-290.

[3] Annecke, K.H., 1977, "A Self-Synchronizing Receiver for a Code-Division-Multiple-Access System", 2nd Symposium on EMC, Montreux, pp. 161-166.

[4] Gold, R., 1967, "Optimal Binary Sequences for Spread Spectrum Multiplexing, IEEE Trans. on IT-13, pp. 619-621.

[5] Varakin, L.Y., 1971, "Selection of Signal Systems for Asynchronous Address Communication Systems with Coherent Reception", Telecom. and Radio Engrg. 25, Pt 1, No. 12, pp. 37-43.

[6] Lindner, J., 1977, "Kollektive von Binärfolgen für asynchrones Codemultiplex", AEÜ 31, pp. 231-238.

[7] Van Blerkom, R., Sears, R.E., Freeman, D.G., 1965, "Analysis and Simulation of a Digital Matched Filter Receiver of Pseudo Noise Signals", IBM Jour. 9, pp. 264-273.

[8] Viterby, A.J., 1966, "Principles of Coherent Communication", McGraw-Hill, N.Y. .

[9] Lindner, J., 1977, "Synthese zeitdiskreter Binärsignale mit speziellen Auto- und Kreuzkorrelationsfunktionen", Dissertation TH Aachen.

[10] Lindner, J., 1975, "Binary Sequences up to Length 40 with Best Possible Autocorrelation Function", Electronics Letters 11, pp. 507.

[11] see [10], Internal Report, Inst. für Elektr. Nachrichtentechnik, TH Aachen.

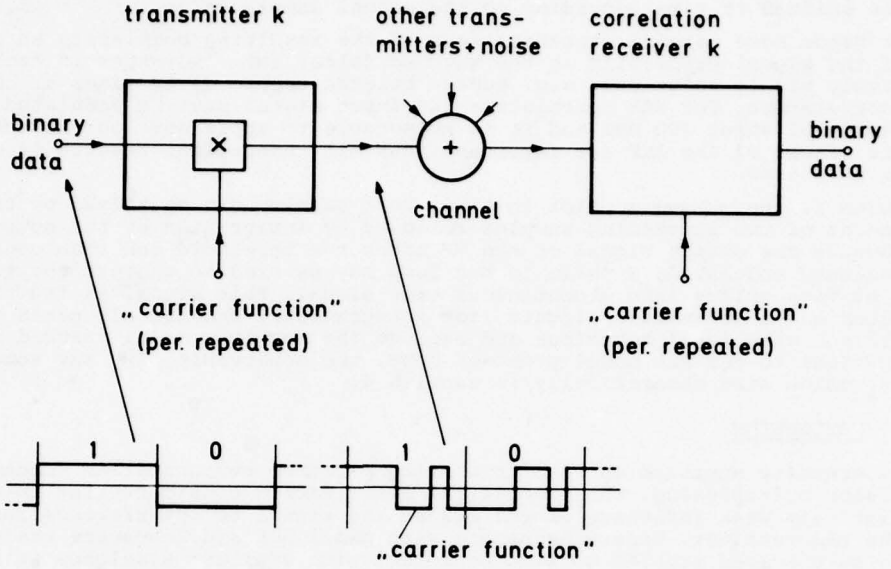


Fig.1 Principle of code division multiplexing for binary data transmission

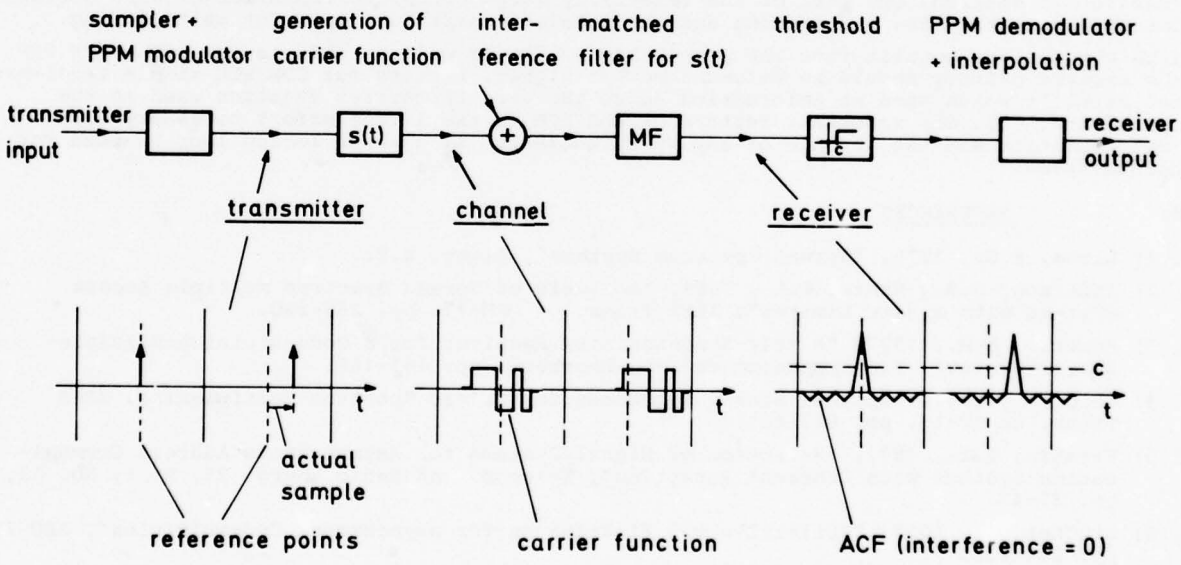


Fig.2 Principle of a transmission system using pulse position modulation

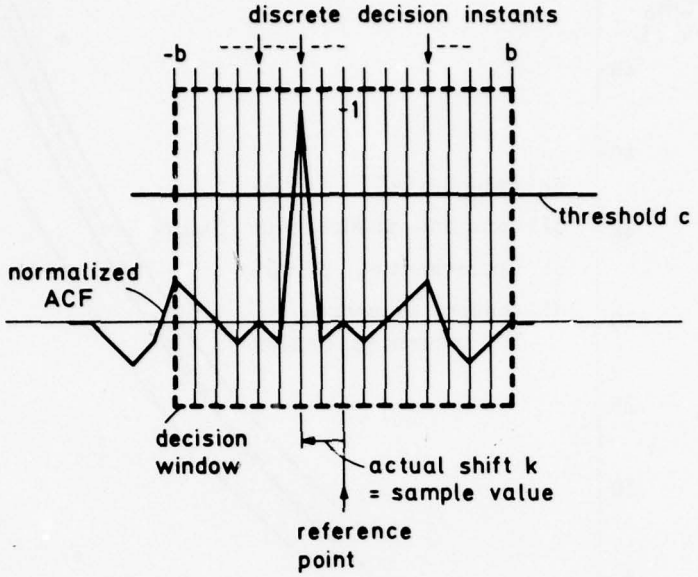


Fig.3 Autocorrelation function and decision window

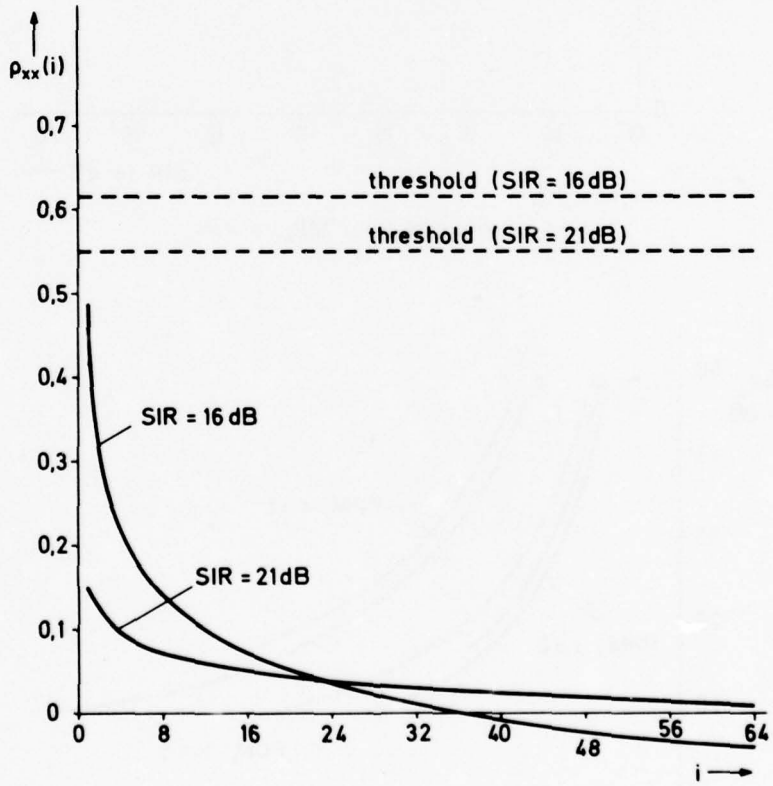


Fig.4 Optimum sidelobe functions; b = 64 , $\bar{\rho}$ = 0,04

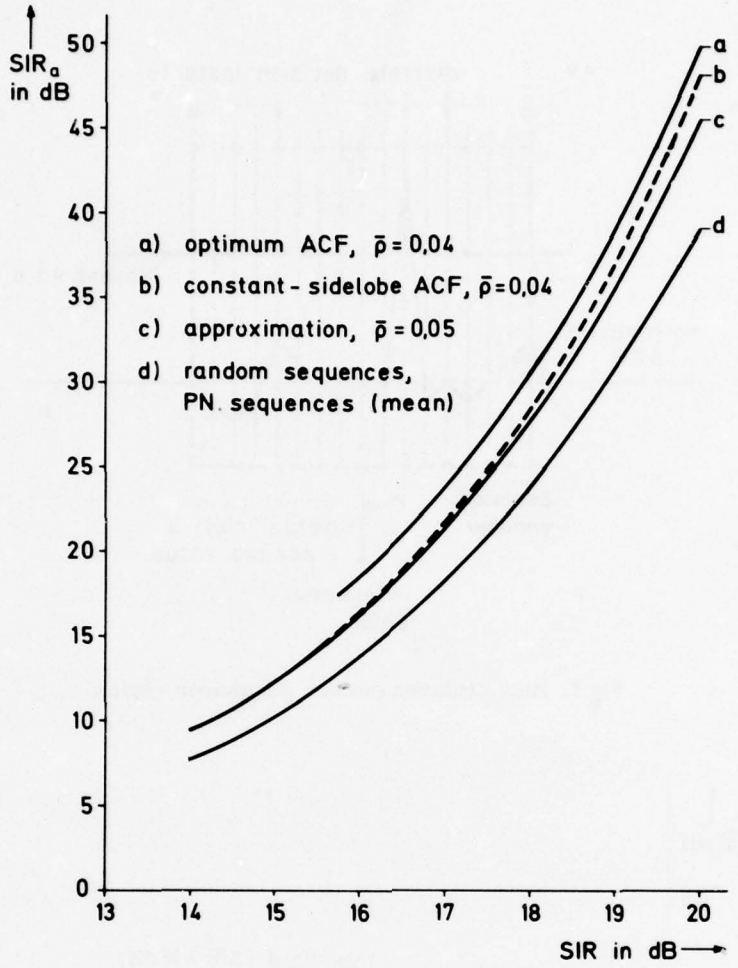


Fig.5 Lower bounds for SIR_a, b = 64

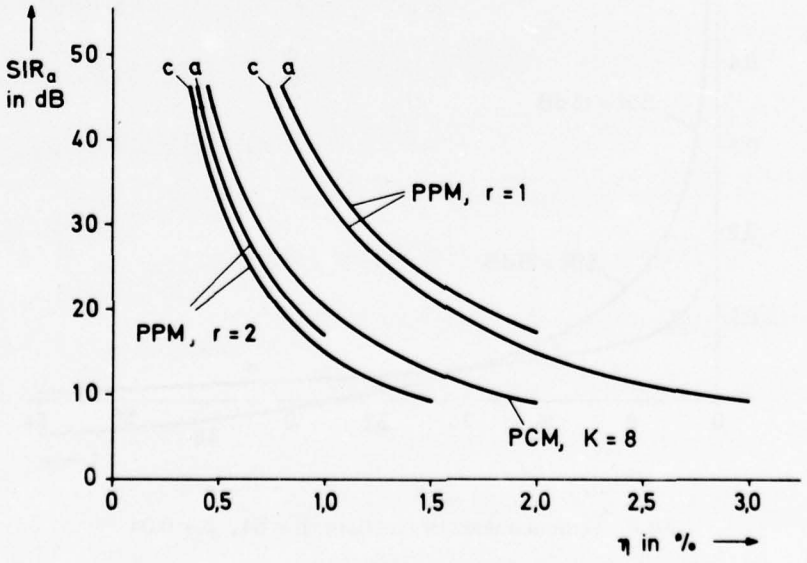


Fig.6 Comparison of CDM/PPM with CDM/PCM; a, c see Figure 5

DISCUSSION

P.Das, US

Would you please clarify why you don't need synchronisation.

Author's Reply

(1) All considerations were made in the baseband, i.e. you need no carrier synchronization.

(2) The PPM transmission can be regarded as the pulse frequency modulation of differences of sample values. Therefore a PLL can be used for demodulation; the adjustment of the natural PLL frequency is not critical.

A TERMINAL ACCESS CONTROL SYSTEM FOR FLEETSAT*

Steven L. Bernstein
Massachusetts Institute of Technology
Lincoln Laboratory
Lexington, Massachusetts 02173 U.S.A.

SUMMARY

The space segment of the UHF FLEETSATCOM system provides a substantial resource to support important communication needs of Navy and other users that are part of the U.S. Department of Defense. Because of the significant investment being made in space and ground segments to provide for these needs, there is strong motivation to examine techniques that make efficient usage of these assets. This paper reports on one such effort, the Terminal Access Control System (TACS) being developed at M.I.T. Lincoln Laboratory. TACS utilizes time division multiple access (TDMA) to derive several circuits from each of the several 25 kHz wide frequency channels available. Access to these circuits is controlled in real-time by a master control station according to user demand; thus TACS is an example of a demand-assigned multiple access (DAMA) system. A description of the access control techniques and efficiency of performance is provided. In order to verify system performance predictions, a master control station and mobile platform subscriber unit have been constructed and used in an extensive test program.

1. INTRODUCTION

The space segment of the UHF FLEETSATCOM system provides a substantial resource to support important communication needs of Navy and other users that are part of the U.S. Department of Defense. Because of the significant investment being made in space and ground segments to provide for these needs, there is strong motivation to examine techniques that make efficient usage of these assets. This paper reports on one such effort, the Terminal Access Control System (TACS) being developed at M.I.T. Lincoln Laboratory. TACS utilizes time division multiple access (TDMA) to derive several circuits from each of the several 25 kHz wide frequency channels available. Access to these circuits is controlled in real-time by a master control station according to user demand; thus TACS is an example of a demand-assigned multiple access (DAMA) system. A description of the access control techniques and efficiency of performance is provided. In order to verify system performance predictions, a master control station and mobile platform subscriber unit have been constructed and used in an extensive test program.

Each FLEETSAT satellite has nine 25-kHz-wide hard-limiting UHF transponders. Without TDMA, using existing AN/WSC-3 transceivers on ships and shore stations, the system would be capable of supporting a maximum of nine circuits per satellite. However, a straightforward system link calculations shows that considerably more than one circuit (with a nominal data rate of 2400 bits-per-second) can be supported via each channel. For example, at least ten such circuits can be carried ship-ship or ship-shore, with the limitation being primarily bandwidth, not signal-to-noise ratio (SNR). Time division multiple access (TDMA) is an applicable technique to increase the number of circuits per channel.

Using TDMA each FLEETSAT channel can support up to nine simultaneous 2400 bps circuits. Each circuit would operate at a burst rate of 32,000 binary symbols per second using QPSK modulation and rate 3/4 error protection coding.

2. DEMAND ASSIGNMENT

TACS implements TDMA and achieves a substantial further increase in effective capacity by pooling all the frequency channels and dynamically reassigning time slots to new users as old users finish their transmissions and relinquish their circuits. This is referred to as demand assignment. The use of demand assignment enables the system to provide satisfactory service to a large community of users as illustrated in Fig. 1.

Figure 1 show the number of equal duty factor communication nets that can be supported by a demand assignment system such as TACS as a function of the number of circuits available and the duty factor per net. (The Engset formula for system performance with a finite population was used.) The blockage probability is specified at 0.01. If the nets were pre-assigned to particular circuits then, of course, only one net could be supported per circuit. It can be seen that demand assignment permits a substantial increase in the number of communication nets that can be supported by a given number of circuits. In addition, the number of nets per circuit that can be supported with a given blockage probability increases as the number of available circuits in the pool increases. TACS permits a mixture of pre-assignment and demand assignment of circuits to accommodate the wide variety of user response time and availability requirements.

3. TACS SYSTEM DESCRIPTION

3.1 System Diagram

The constituents of TACS are shown in Fig. 2. Basically, the TACS Master Control Station equipment consists of the system Multiple Access Controller (MAC), which is a software demand assignment manager resident in a mini-computer; a microprocessor-based Terminal Access Controller (TAC), which implements the circuit-switching and TDMA commands of the MAC; a full-duplex transceiver system consisting of two AN/WSC-3 units;

*This work was sponsored by the Department of the Navy.
The views and conclusions contained in this document are those of the contractor and should not be interpreted as necessarily representing the offical policies, either expressed or implied, of the United States Government.

and a group of user I/O devices, such as vocoders, teletypes and data terminals. Mobile-platform TACS subscriber units are similar, except that they have no MAC (only one is needed in the system) and they may be half-duplex (requiring only one WSC-3). This figure emphasizes the fact that TACS can be implemented as an add-on to existing satellite communications equipment. Several hundred ships have already been outfitted with one or more WSC-3 transceivers and associated antenna subsystems.

3.2 Frame Structure

Figure 3 illustrates a typical TACS frame structure for use with a system burst rate of 32,000 binary symbols per second in data slots. Control and other system housekeeping signals are transmitted at a burst rate of 9600 symbols per second for greater reliability. A basic requirement is that the user data rate times the frame duration must not exceed the burst rate times the slot duration (after allowing for guard intervals, modem acquisition and burst synchronization). Beyond this, the details of the frame structure are influenced by half-duplex operational efficiency, byte organization requirements, control burst structure and other considerations. The number of data circuits that can be provided is a function of the user data rate, the system burst rate and the frame structure, as described in more detail below.

An advantage of the TDMA approach is that a half-duplex terminal, by operating in two or more time slots, can support several simultaneous, independent two-way circuits. Slots marked with an asterisk(*) in Fig. 3 can be used simultaneously by a half-duplex user. The time required for the Master Control Station to respond to a circuit request can be as short as a single frame duration, if the user is able to successfully use one of the shared request slots that are provided in every frame. If his attempt fails because of contention with another user request, he will repeat the attempt in a later frame. A "safety valve" that guarantees that his request will reach the Master Control Station even under conditions of very heavy request traffic is the dedicated request slot, which is scheduled once per minute for each user in the system.

Changes in link conditions may dictate changes in the burst and code rates. The Master Control Station can control these parameters to adapt them to the new conditions.

The control and ranging slot in Fig. 3 contains the control burst, completely specifying the usage of the remaining slots, which is assembled on a frame-by-frame basis by the software demand assignment manager (the MAC) at the Master Control Station. Every user in the system continuously copies the control burst, whether he is using any communication circuits or not. A user who has just turned on his equipment must first identify and interpret the control burst, to establish his receive timing. He then transmits an initial ranging burst in the second part of the control and ranging slot, using a timing pre-correction of 260 milliseconds (relative to his receive clock), which is the median round-trip propagation delay to the satellite. His actual propagation delay is measured at the Master Control Station by observing the time of arrival of his ranging burst, and the corrected value for his delay is transmitted as part of the next control burst.

The user is now ready to request a circuit, using a shared request slot. If he is unable to get through in repeated attempts because of contention, his request will be transmitted when his dedicated request slot comes up. Each subscriber unit always reports in his dedicated request slot, even if he has no circuit requests to make, so the MAC can compute range delay updates.

Circuit requests from a particular subscriber unit are actually initiated by a user on one of its four I/O ports. The MAC fulfills the request if possible and communicates the assignment information to the subscriber unit as part of the next control burst. The subscriber unit responds by coupling the I/O port in question to the assigned channel and slot.

As long as it is not required to do two different things at once - e.g., receive and transmit at the same time - a half-duplex terminal can accommodate more than one I/O port simultaneously. For example, if the terminal is operating in the fourth slot in Fig. 3, as indicated, it is locked out of slots 5 and 6 because the transmit lead time for those two slots overlaps the receive time for slot 4. Slot 7, however, can be used without interferring with slot 4.

4. SUBSCRIBER UNIT (TAC)

Figure 4 gives a more detailed description of the TAC subscriber unit. It is divided in two sections, each controlled by a microprocessor. The DAC performs administrative functions; it buffers user data, implements the commands of the MAC, and maintains frame and slot timing. The Processor/Decoder primarily does bit manipulation and receiver/transmitter control. In slots where data is to be transmitted, the Processor/Decoder convolutionally encodes high-rate burst data from the DAC, interleaves it, and sends it to the R/T. When the data is de-interleaved at the receiving end, multi-symbol bursts of erasures occurring at the receiver input are randomized.

In slots where data is to be received, the Processor searches for the burst synchronization sequence, de-interleaves the data, decodes it by means of a variable rate soft decision Viterbi decoder, and transfers it to the DAC for buffering and re-formatting to a continuous bit stream at the user data rate.

Figure 5 shows the TIC (Terminal Input Controller), a keyboard entry and alphanumeric display device. Its primary function is to allow a user to communicate circuit requests and other information to the MAC via his own TAC. Although TICs could serve any I/O device that TACS can accommodate, they will actually be used only for voice circuits. A TIC and its handset can be located at a considerable distance from the TAC; for example, they could be installed on the bridge or in the Combat Information Center of a ship. The user enters his circuit request via the TIC keyboard by going through a prompt-and-reply sequence, or by keying a short pre-set sequence for frequently-called addresses.

Figure 6 shows a fully assembled TACS subscriber unit. The actual TAC hardware takes up only 15" of vertical rack space even in this experimental versions. The rest of the rack space is taken up with peripheral hardware and a WSC-3 transceiver at the bottom.

5. CENTRAL CONTROLLER (MAC)

Figure 7 shows the key features of the MAC, which is resident in a mini-computer at the Master Control Station (Taylor, L.E..,to be published). The Data Base contains all the relevant information on the parameters and status of all terminals in the system, as well as the satellite channels. Using the information in the Data Base, together with circuit requests, ranging signals and reports, the MAC performs its three main functions: circuit assignment, control burst preparation and adaptive control of system parameters. Basically, the assignment problem is addressed by search for a match among the available satellite circuits and the slot and channel availabilities of the calling party and the called party.

The adaptive control algorithm of the MAC processes several kinds of input information to determine changes that must be made in the data base to reflect changing channel and traffic conditions. Among the changes that can be made are: the number of request slots; the code and symbol rates of data slots; and call precedence thresholds. The first is based on subscriber unit reports of the number of times they must repeat their circuit requests in shared request slots before they get through to the MAC. The second kind of adaption is based on subscriber units reporting their link conditions, both signal-to-noise ratio and pulsed radio frequency interference. The MAC then adjusts code rates, symbol rates and slot durations to tailor to the needs of each user. The range of burst rates is from 9600 to 32,000 symbols/sec and the code rate is selectable from 1/2, 2/3, 3/4, 4/5. The third kind of adaptive change is a response to increasing congestion on the satellite circuits. Traffic is reduced by setting a numerical precedence threshold, and informing users via the control burst that requests of a precedence lower than the threshold will not be honored.

6. PRELIMINARY TEST RESULTS

An extensive series of system tests are being conducted on TACS. In addition, analytical modeling of key performance aspects has been done where possible. Results will be given here in two areas: decoder performance and preliminary MAC circuit assignment performance.

6.1 Decoder Performance

All TACS traffic is protected by the user of convolutional coding and soft-decision Viterbi decoding. This is done to protect the links from the effects of pulsed interference and to provide increased link margin allowing disadvantaged users to remain in the system.

Figures 8 and 9 show the decoder performance when the channel is perturbed by Gaussian noise and pulsed RFI. The probability of error per bit is shown as a function of energy-per-bit to noise density ratio (E_b/N_0) with the RFI duty factor, f, as a parameter. Both simulated and measured performance is shown with good agreement between them. The rate 1/2 code is seen to provide efficient (E_b/N_0 = 7 dB) protection with 20% of the signal "erased" by RFI even to error rates as low as 10^{-5}. The rate 3/4 does this well with 5% erasures. Further details of these results are given in (Bernstein, S.L., et al.., 1977).

Recall that the MAC can assign burst and code rates to match the channel conditions. With the coding/decoding scheme being utilized, this can be done efficiently over a wide range in channel conditions.

6.2 Demand Assignment Performance

The performance of the MAC Central Controller is being measured via a software Call Request Simulator. This can generate traffic loads equivalent of hundreds of nets of tactical users.

One meaningful measure of the MAC's demand assignment performance is call blockage probability, which is equivalent to 1 minus the grade of service. The Call Request Simulator can be programmed to synthesize any desired communications scenario, and a statistics-gathering package in the MAC can measure the average call blockage rate over many thousands of frames. Experimental data obtained in this way can be used for verification of theoretical predictions, for the few simple, uniform communications scenarios that are analytically tractable. Having established the credibility of the system by this means, one can then use it to obtain performance characterizations for more complicated but more realistic scenarios which are impossible to analyze. An example of system demand assignment performance data obtained with the Call Request Simulator is shown below.

Duty Factor	Number of Nets Served	Blockage Probability with 27 circuits	
		Theoretical	Experimental
0.2	66	0.010	0.010
0.3	48	0.012	0.010
0.4	39	0.012	0.010

In this scenario it was assumed that the MAC had access to 27 shared circuits (e.g., a total of 27 time slots on several frequency channels), which were arranged in 3 independent groups of 9 circuits each. Using a blocked-calls-cleared queuing theory model with a finite user population, probability of call blockage was calculated for various populations of identical users, with various average per-user duty factors. The table entries indicate three combinations of duty factor and population size that had a predicted blockage probability close to 0.01. The experimental results were obtained by programming the Call Request Simulator for the same population sizes and duty factors, letting the system run for thousands of frames, and measuring the resulting blockage rates. Other examples have been worked out, with comparable results.

7. ACKNOWLEDGMENT

The conception and successful development of TACS was achieved by the members of the Terminal Technology Group at Lincoln Laboratory, the members of which are too numerous to mention here.

However, the author would like to acknowledge that he has drawn freely from (Heggestad, H.M..,1978) in describing the TACS subsystems and test results.

References

Bernstein, S.L., Heggestad, H.M., Mui, S., Richer, I., 1977, "Variable-Rate Viterbi Decoding in the Presence of RFI", Natl. Telecommunications Conf., Los Angeles.

Heggestad, H.M., 1978, "System Test Results on the Lincoln Laboratory TACS," AIAA 7th Communication Satellite Systems Conf., San Diego.

Taylor, L.E., to be published, "Terminal Access Control System (TACS) Circuit Allocation," Technical Report NSP-4, Lincoln Laboratory, M.I.T.

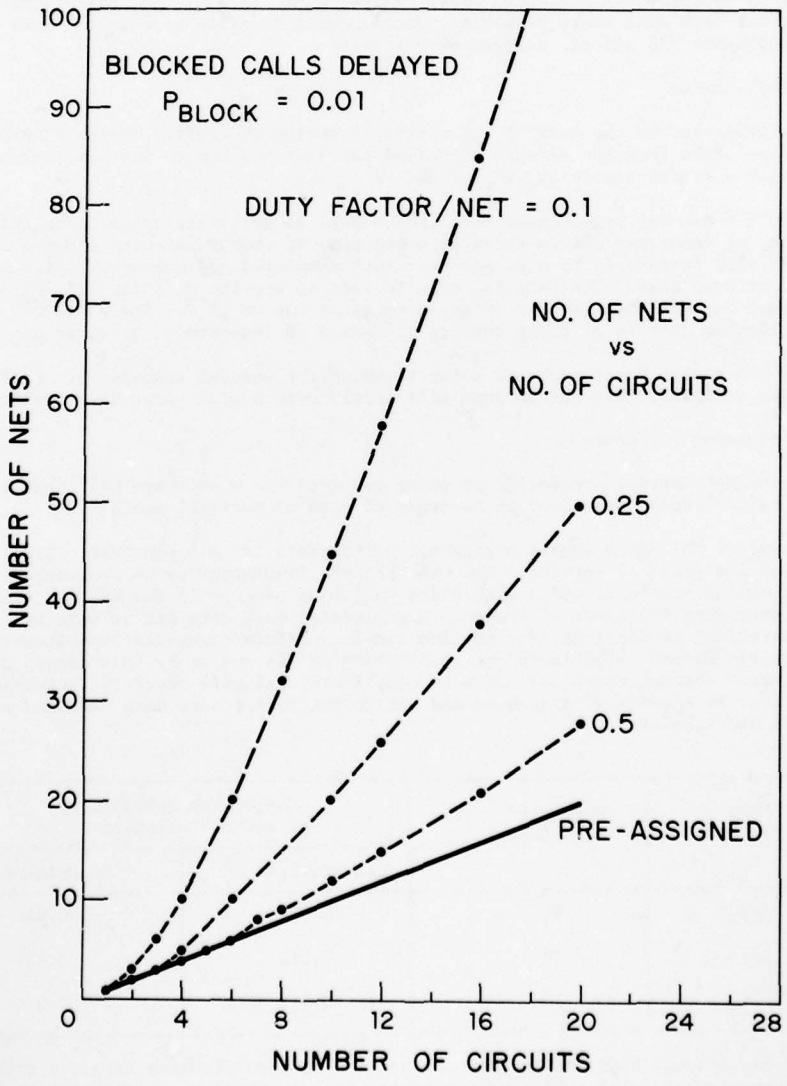


Figure 1. Number of Nets Versus Number of Circuits Available.

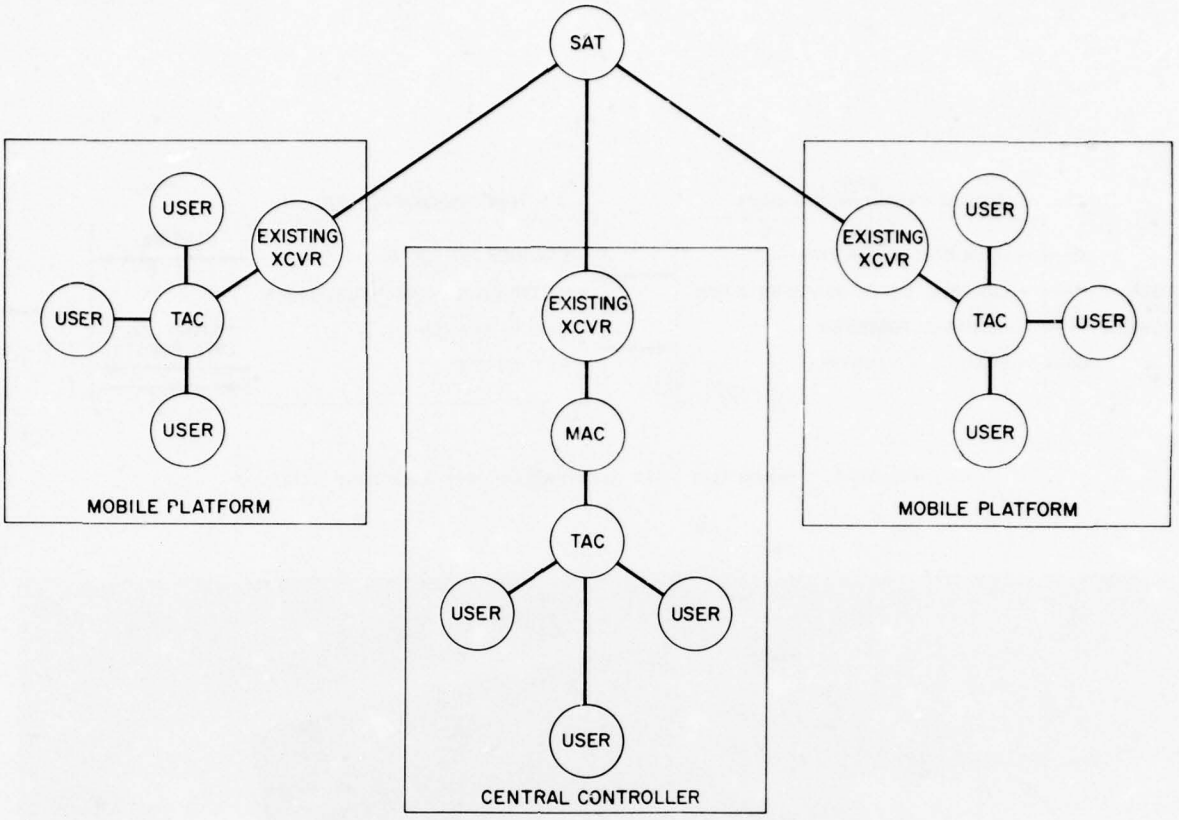
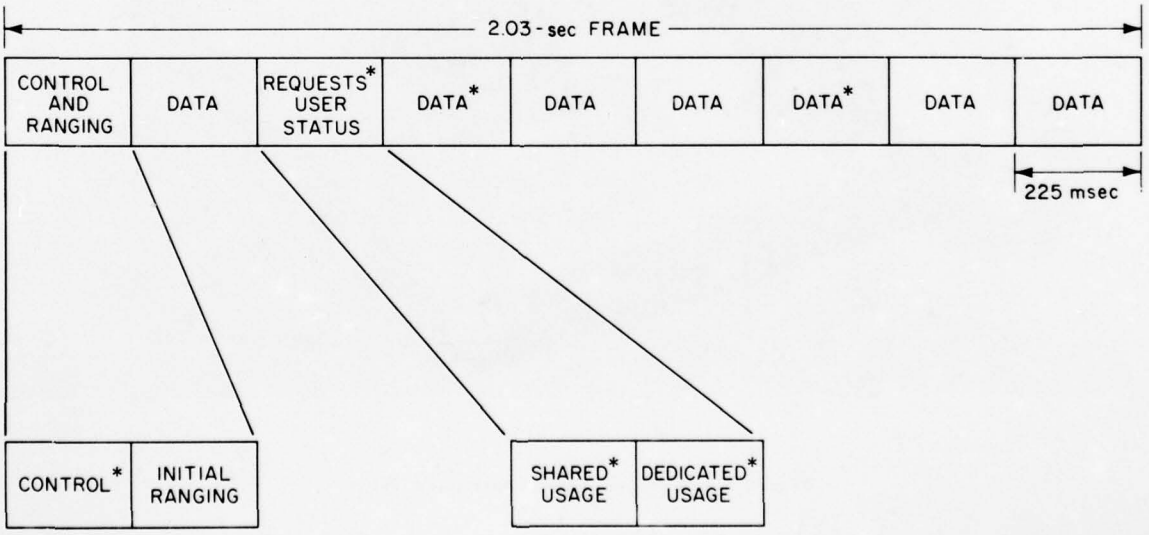


Figure 2. TACS System Configuration.



*EXAMPLE OF SLOTS THAT CAN BE USED SIMULTANEOUSLY BY H-DUX USER
NOTE: EACH DATA SLOT PROVIDES A 2400-bps CIRCUIT

Figure 3. Typical Frame Structure.

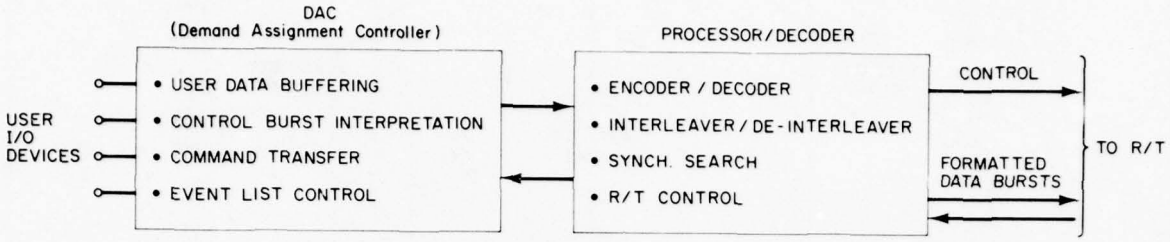


Figure 4. Subscriber Unit Terminal Access Controller (TAC).

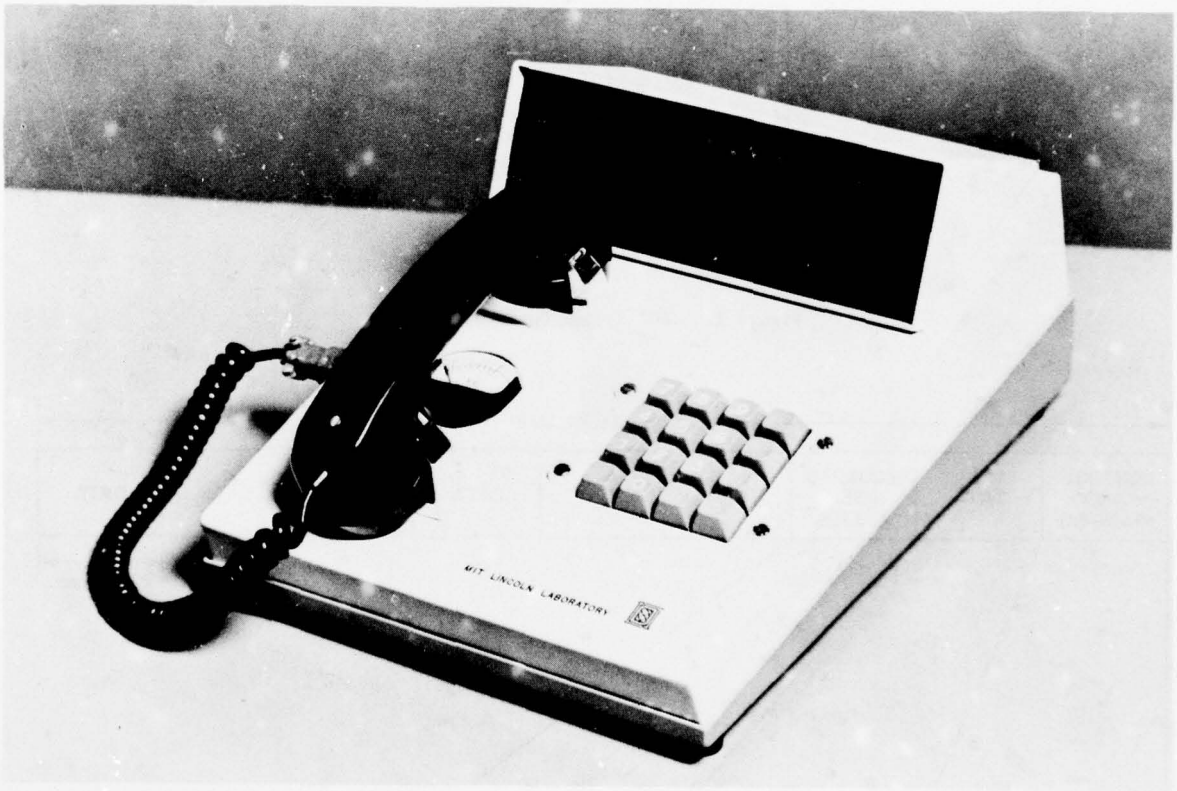


Figure 5. Terminal Input Controller (TIC).

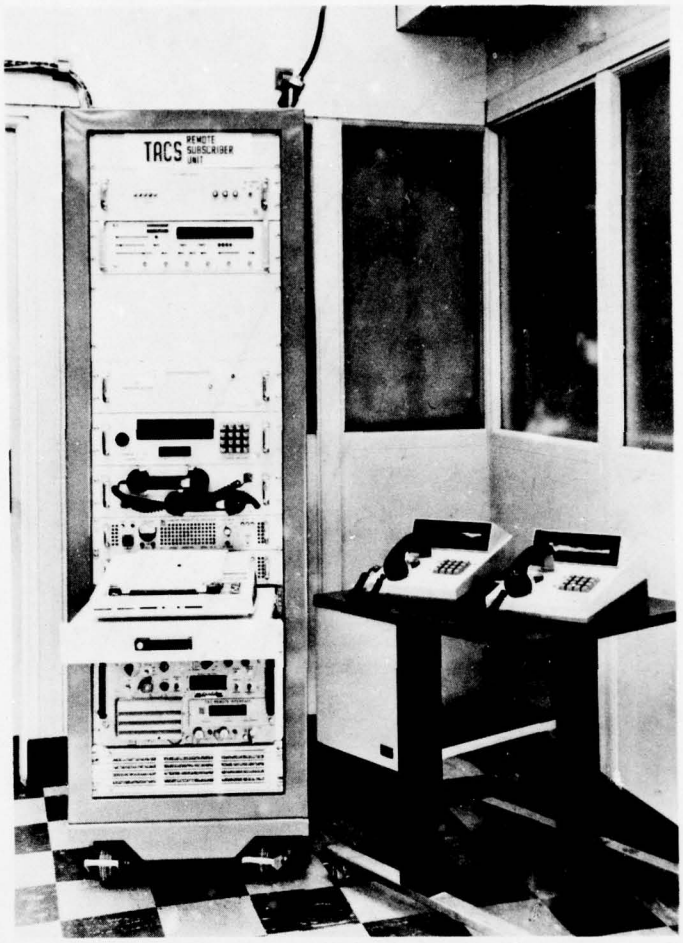


Figure 6. TACS Subscriber Unit.

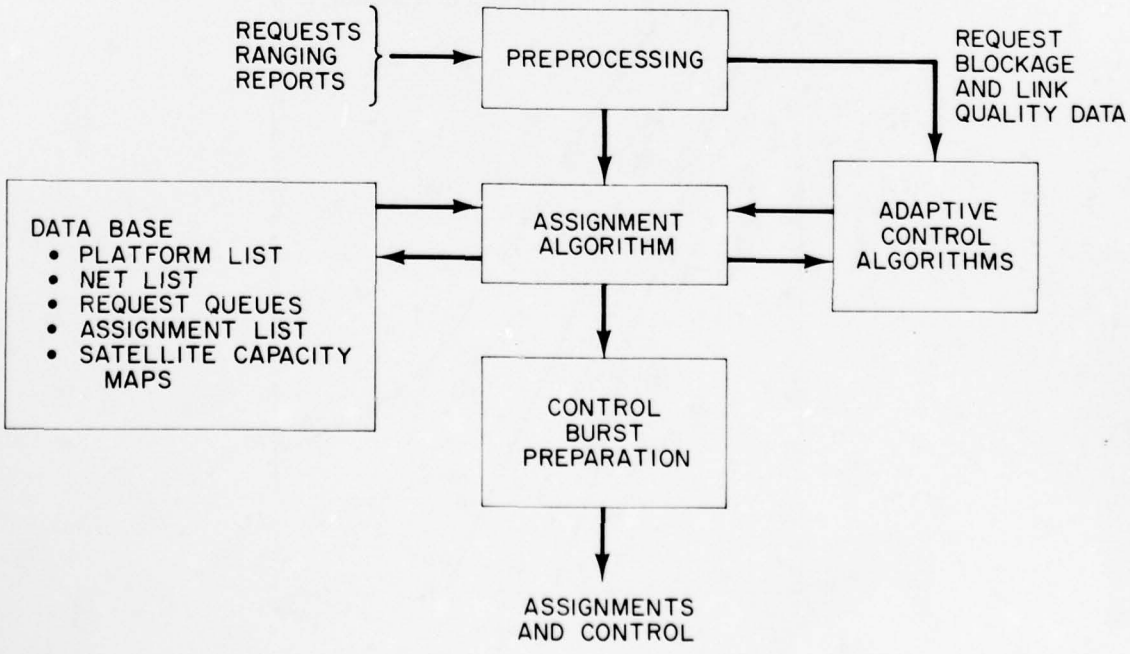


Figure 7. Multiple Access Controller (MAC).

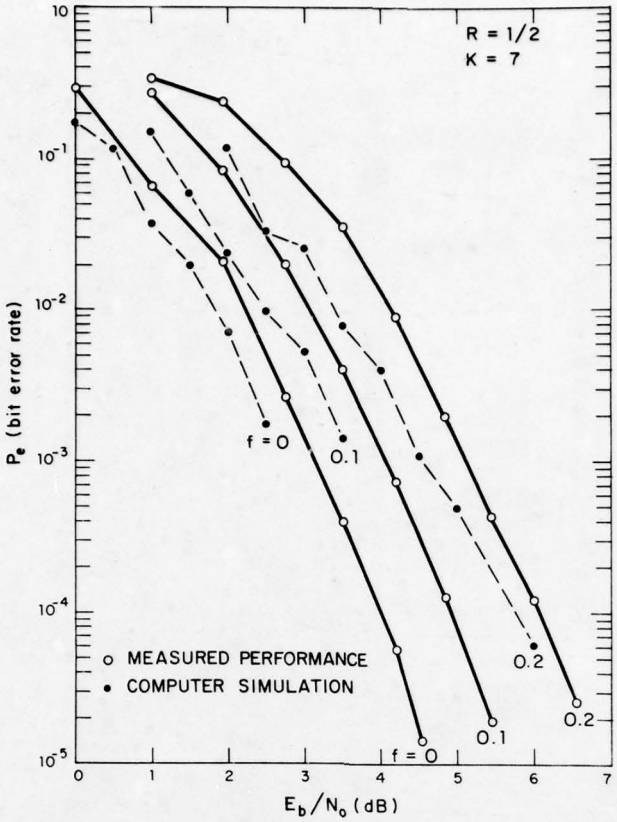


Figure 8. Rate 1/2 Decoder Performance.

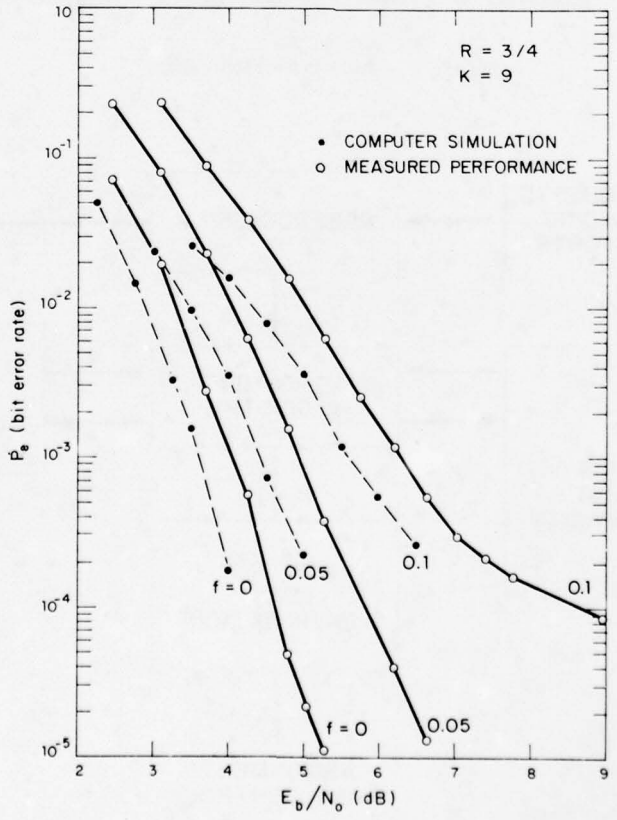


Figure 9. Rate 3/4 Decoder Performance.

DISCUSSION

A.Sewards, Ca

(1) Why was a Viterbi coder chosen for the system – was it to improve performance in a gaussian noise situation or to cope with bursts?

(2) Have any problems been experienced with loss of synchronization, given the large number of bits required to implement the Viterbi decoding process?

Author's Reply

(1) Convolutional coding, interleaving, and Viterbi decoding provides a robust combination for combatting both Gaussian noise and burst noise. It is insensitive to the burst statistics and provides for a low E_b/N_0 also.

(2) Synchronization is accomplished once each data burst with a short (32 symbol) sequence. Once synchronization is achieved in a slot it can not be lost.

IMPLEMENTING JTIDS IN TACTICAL AIRCRAFT

David R. McMillan
The MITRE Corporation
Bedford, Massachusetts, USA

SUMMARY

This paper addresses the general problem of implementing the Joint Tactical Information Distribution System (JTIDS) in tactical aircraft. Major characteristics of the JTIDS system are summarized and relevant elements of the JTIDS program plan are highlighted. The context for specific requirements to be satisfied with JTIDS aircraft installations are established in terms of operational functions, sources and sinks of information which support those functions, and current information distribution systems and man-machine interfaces which mechanize the transfer to and from tactical aircraft and their crews. The relative benefit of JTIDS over present solutions to operational information needs is established in this context. The paper concludes with an overview of key elements of the tactical aircraft implementation challenge.

1. OVERVIEW OF JTIDS

The Joint Tactical Information Distribution System (JTIDS) is a high-capacity, secure, jam-resistant time-division multiple access information distribution system providing integrated communications, navigation and identification capabilities. It is being developed by the U. S. Department of Defense for use by its tactical forces. DoD has also offered JTIDS to NATO as a standard ECM-Resistant Communications System (ERCS) for use by Allied forces. The system architecture and design features of JTIDS support the operational needs of U. S. and Allied tactical force elements in the severe threat environment of the 1980's and beyond. This section reviews the operational and technical characteristics of JTIDS, along with the program for implementing JTIDS in U. S. tactical aircraft.

1.1 The Operational Need

United States tactical air forces gained significant operational experience during the Southeast Asia conflict in an environment where real-time command, control, and communications were important to air operations. Subsequent analysis of that experience, augmented by similar analysis of Israeli operational experience in the hostile EW environment of the October 1973 Mideast conflict, highlighted the urgent need for a jam-resistant and secure high-capacity information distribution system. Key requirements include:

a. The need to transport information from its sources to its users, even when the users do not know who the sources might be or what information is available.

b. The need to distribute large volumes of information among many subscribers with the minimum possible delay.

c. The need to minimize interference caused by simultaneous transmission of information by two or more subscribers, or to reduce the loss of essential information if simultaneous transmissions must be made.

d. The need to support the information requirements of all subscribers, and enable each subscriber to receive and process only that information which is relevant or useful to his tactical mission.

e. The need to provide positive and unambiguous identification of all friendly tactical elements.

f. The need to enable the determination of position in a common geographic grid and the reporting of that position to other tactical elements on a continuous basis.

g. The need to provide the foregoing capabilities throughout the tactical theater with minimal imposition of artificial geographic boundaries or communication management functions.

h. The need to provide the foregoing capabilities with minimal degredation due to enemy jamming, and with maximum probability that the enemy cannot obtain or exploit anything useful from the system.

In a very direct and significant manner, the satisfaction of these needs will improve the ability of tactical force elements to achieve their objectives while it enhances their ability to survive. The information distribution system which meets these requirements will therefore pay for itself in terms of friendly resources saved and hostile resources destroyed.

1.2 JTIDS SYSTEM ARCHITECTURE

The fundamental architectural feature which gives JTIDS its performance capabilities is a communication net or bus which is shared by its users for both transmission and reception of messages. A frequency hopping and pulse coding pattern is used to provide multiple nets within the JTIDS band. As a system, JTIDS is designed to support the combined information transfer and management needs of various users. System-wide standards which must be satisfied by all users include signal structure, transmission timing, net management protocols, message contents, and data formats. These standards are necessary to provide interoperability among the users.

This notion is illustrated in Figure 1, where each JTIDS subscriber transmits his data in assigned time slots. When a subscriber is not transmitting, he can receive whatever data is being transmitted by others. When data messages are broadcast, the originator does not know the identity of all potential users of that data. Each terminal is programmed to monitor all transmissions and select for further processing only those messages which contain information useful to the subscriber's mission. In general, the breakdown of JTIDS system capacity will correspond to the following types of information distribution.

a. Tactical aircraft equipped with JTIDS will broadcast position and status reports on a periodic basis.

b. Surveillance, command, and control elements such as the E-3A and ground-based centers will broadcast track reports on hostile, unknown, and non-JTIDS equipped friendly aircraft.

c. Tactical aircraft and their control elements will use JTIDS to provide a number of secure, jam-resistant voice channels.

d. Many subscribers will transmit messages for command, control, mission coordination, and threat warning.

Examination of the capacity demands of various subscribers reveals that tactical aircraft primarily access the JTIDS net in a receive mode, while command and control elements impose a heavy transmit load on the net. This aspect will be elaborated in Section 3.

1.3 System Implementation Plan

Advanced development programs of the U. S. Air Force and U. S. Navy were merged in 1975 to form the joint program for JTIDS. Initial development efforts have been directed toward surveillance and command/control system applications. Operational capabilities will be achieved for the E-3A and ground-based centers in the early 1980's, thereby supporting high-capacity distribution of real-time track data in a hostile electronic warfare environment.

The major thrust of JTIDS is to implement terminals and associated interfaces on fighters and attack aircraft. Once JTIDS is implemented aboard tactical aircraft, most of the system's benefits will be realized. Highlights of the tactical aircraft implementation program include:

a. Delivery to the U. S. Navy and Air Force of twelve "fighter-sized" advanced development model (ADM) terminals during the summer of 1978.

b. Laboratory, cargo aircraft, and system lab testing of seven of these terminals by the U. S. Navy throughout 1979.

c. Installation of five of the ADM terminals in pods for flight testing aboard fighter and attack aircraft by the
U. S. Air Force beginning in mid-1979.

d. Initiation of full scale development for the production version of the Class 2 "fighter-sized" terminal, and development of the F-15, and F-16, and pod-mounted installations during the 1979 - 1982 period.

e. Initial installation into U. S. Air Force tactical aircraft by 1984.

Potentially, JTIDS terminals may be installed in over 6,000 U. S. tactical aircraft.

2. TACTICAL AIRCRAFT REQUIREMENTS

Since JTIDS is an information distribution system, the specification of individual JTIDS capabilities for tactical aircraft involves the information being distributed. Tactical information needs are characterized in terms of utilization, transfer, and man-machine interface. This characterization is accomplished by answering four questions about the information which could be distributed with JTIDS.

a. What information needs are associated with the key operational functions of tactical aircraft?

b. What are the sources and sinks of the information needed by tactical aircraft and crews to perform these operational functions?

c. What information distribution system(s) and man-machine interfaces are capable of transferring the information associated with each operational function between the tactical aircraft and the associated source or sink?

d. Which of these information distribution capabilities are best satisfied by JTIDS, and what priorities are attached to these JTIDS capabilities?

When question d. has been answered, the specific capabilities to be provided through the installation of JTIDS in tactical aircraft will be identified. This section establishes the framework for that evaluation by defining the range of possible answers to the first three questions.

2.1 Operational Functions

The starting point for establishing tactical aircraft information needs is a breakdown of operational functions. For this evaluation, a three-level breakdown has been developed. The top level includes three overall objectives:

a. Accomplish own mission,
b. Support command and control needs, and
c. Support other mission elements.

The second level under "accomplish own mission" is broken into eight general activities which are performed during any tactical aircraft mission. Each of these eight activities is then subdivided into specific operational functions with definite objectives and information needs. Similar second- and third-level breakdowns are made under the two support objectives, but those functions have even more explicit information implications.

Figure 2 presents the resulting breakdown of operational functions. The meaning of most items is clear from the titles, and the characteristics of the information needs associated with each operational function are elaborated in paragraph 2.4 below.

2.2 Information Sources and Sinks

Figure 3 presents a hierarchical breakdown of information sources and sinks, beginning at a generic level and continuing to the level of detail at which information handling characteristics are directly related to tactical aircraft information needs and operational functions. The first level identifies four generic categories: actual physical entities of interest, bodies of information, aircraft on-board systems, and other entities. Although the first two categories are not particularly relevant for JTIDS considerations, they are included to emphasize that some information needs may not be satisified by JTIDS.

The breakdown under aircraft on-board systems includes four major types according to their own function: navigation, weapons, electronic warfare, and observation. Further breakdowns within each type indicate specific systems which may already satisfy some information needs aboard tactical aircraft. Finally, the category entitled "other entities" is sub-divided into broad groups of information handling elements which constitute the outside world from the viewpoint of the tactical aircraft.

2.3 Information Distribution Systems

In the broadest possible sense, a variety of mechanisms can distribute information among the tactical aircraft, its crew, and the sources and sinks just described. Figure 4 shows the breakdown used for this analysis. As with the previous breakdowns, this is a hierarchy which begins with broad categories and subdivides into specific examples.

The top-level breakdown distinguishes among direct exchanges; the three major functional categories of communication, navigation and identification; and a category for combinations of these three. Characteristics of the lower subdivisions are appropriate to this analysis, although they may be incomplete outlines. Under communication, the key distinction is whether voice or data is being communicated. Navigation systems are usually considered in two categories: wide-area systems, which usually have lower update rates and accuracy standards commensurate with the scale factors of the problem; and terminal-area systems, whose information characteristics are geared to the need for highly precise guidance in a limited area where unacceptable errors could build very rapidly.

2.4 Cockpit Man-Machine Interfaces

This dimension is particularly important when considering potential JTIDS capabilities to be implemented in tactical aircraft. Except for totally automatic functions which provide some capability without crew intervention, the crew's total perception of a JTIDS capability will depend on what he sees and does in the cockpit. In aircraft such as the F-15, F-16, or A-10, the single crew member must perform major functions such as weapons management and navigation while also flying the airplane under very demanding circumstances. This creates a competition for his attention which he will resolve by focusing on those cockpit interfaces which are the most useful and useable for his immediate needs.

The breakdown in Figure 5 is keyed to this perspective. The only top-level distinction is between interactions which are simple, involving a single uni-directional action or information flow, and complex interactions which are compounded from two or more simple interactions. Most complex interactions will require some dedicated attention to execute the interaction, with a consequent impact on attention span for other primary functions and information. Because of the extremely high demands on the pilot's attention in a fighter cockpit, simple interactions are for more realistic in terms of effectively implementing JTIDS.

Simple interactions are categorized according to the crew member's faculty which is involved in the interaction. Visual transfer is strictly from a display system to the pilot. Aural transfers could occur in either direction, and manual transfers will only occur as inputs from the pilot. Complex interactions are subdivided into those which cause information transfer to the pilot, and those which cause information to be transferred outside the aircraft. The man-machine interactions associated with data systems such as JTIDS normally fall into these categories.

3. CAPABILITIES TO BE IMPLEMENTED

Given the inherent potential of JTIDS to support the implementation of various capabilities, along with the inherent constraints upon such factors as size and crew workload in tactical aircraft, it is extremely important to select the proper capabilities to be implemented.

This section considers qualitatively potential JTIDS capabilities for tactical aircraft in terms of the operational functions, information sources and sinks, information distribution systems, and cockpit man-machine interfaces which were indicated in Section 2. Quantitative assessments of potential JTIDS capabilities are underway. The discussion which follows is subdivided according to the operational functions outlined in 2.1 and listed in Figure 2.

3.1 Get Requirements to Crew

All three functions under this heading involve one-time transfers of information in batches from some command or control agency to the flight crew. Any tactical mission is defined by specifying the flight plan (or flight path), the objective location and characteristics, anti-air defenses to be encountered, alternate missions, and a variety of similar items. For a preplanned mission, a full set of these data items must be provided before takeoff. Both the modified and air-diverted missions involve changes after the flight has begun. In these cases, a subset of the data items required for preplanned missions must be furnished to the crew in flight. The present means for presenting this type of information to flight crews is verbal, either during the preflight briefing or by voice radio. Utilization of data communication techniques such as JTIDS raises possibilities such as a schematic pictorial display or alphanumeric printout in the cockpit.

The only way that JTIDS could supplant a preflight briefing would be to send mission data to each pilot in his aircraft, using an interface such as an alphanumeric display or printer. This might be useful as an alternative to physically copying a large volume of data, but it should not replace the preflight briefing which has other benefits. For inflight data transfers, JTIDS could provide significant benefits if a simple cockpit interface such as an alphanumeric display or strip printer were used. The primary benefit would be the removal of traffic from voice radio channels, which could be very significant due to the length of time required to read and acknowledge new mission assignment data.

3.2 Coordinate Flight with Other Elements

This activity encompasses all functions whose goal is to achieve and maintain physical proximity with other mission elements during coordinated maneuvers. Information to support these functions primarily involves location and kinematic data, along with projections or intentions for future values of this data. Factors such as the precision of this data, the need for static or kinematic data items, and the need for amplifying data such as fuel or weapon status are implied by each of the four coordination functions under this heading.

Present procedures for in-flight coordination are based upon mutual visual contact after gaining proximity. Proximity is usually gained through cooperative navigation and voice communication between the elements. Where coverage is available, this may be aided by a surveillance or control element which observes both parties and directs them along coordinated paths. Information sources are thus the actual physical entity during contact flight, and a command/control element such as the CRC/CRP during indirect coordination. Information distribution is by voice radio, supplemented by visual observation during contact flying.

JTIDS could improve these functions in one of three ways. First, the digital voice feature could be used as a direct substitute for the present UHF voice radio. This would give resistance to jamming and other countermeasures, thus making the voice links available more of the time. Second, flight directions generated by a command and control element could be transferred by data link and displayed or printed as flight guidance cues. Third, the current positions, altitudes, identities, and history trails of the coordinating elements could be depicted in a common frame of reference on a graphical display of the flight situation. The display would be updated from JTIDS position/status and track report generated by other mission elements and command/control elements. In effect, such a display would extend the regime of contact flying by providing visibility of relative positions to a much greater range.

3.3 Get to Objective

The purpose of this activity is to reach the assigned objective and be positioned at the beginning of the proper approach course for execution or delivery. All three third-level functions involve finding fixed locations, paths, or areas, and the information needs associated with these functions include the coordinates and identifying parameters of these locations. It should be noted that the satisfaction of these information needs is considerably more difficult for a target in hostile (and possibly unfamiliar) territory than it is for locations such as the recovery airfield in friendly territory. In one case, the enemy will actively intervene to deny the needed information; in the other case, multiple cooperative friendly systems are designed to satisfy the information needs with high probability and accuracy.

At the present time, this operational function is performed by on-board navigation systems plus direct observation of the terrain and other features. Except for self-contained capabilities such as inertial navigation systems, the viability of these information sources will deteriorate or vanish in hostile airspace. This leaves the burden upon visual contact flying, particularly near or behind the enemy lines.

The primary method for improving these functions with JTIDS involves the relative navigation feature, by which own-aircraft location in a common frame of reference is estimated from received JTIDS messages. With appropriate messages, an appropriate display of the objective location relative to own-aircraft location can be provided. Under good visibility conditions, the flight-following function is improved, but accuracy requirements for this function can be satisfied reasonably well by dead reckoning. Also under good visibility conditions, the approach course is usually defined with respect to terrain or other ground features for ground strike, so JTIDS only provides marginal help there. However, under poor visibility conditions, such as night, bad weather, or battlefield smoke and haze, the JTIDS situation display may provide significant benefits.

3.4 Deliver Weapon Onto Target

For purposes of this discussion, the weapons delivery activity is considered to begin at the point on the approach where target acquisition is first accomplished by the weapons system. It is considered to end at the point where the last weapon release occurs for independent weapons, or the weapon guidance interaction is completed for controlled weapons. The primary information needs implied by these three functions are target location data plus weapon and/or aircraft location and dynamic data. If links between the weapon and aircraft after weapon release are included, additional needs related to weapon guidance, control, and sensor feedback must be included. It is important to separate the information needs for this activity into those of the crew/aircraft/weapons system combination (to or from other elements) and those between the crew/aircraft and its weapons systems. At present, the primary focus is upon the first category since most weapon systems incorporate direct links to satisfy information needs between the weapon and its launch platform. This is operationally appropriate since these information needs do not involve connectivity with other elements as provided by JTIDS. The major contribution of JTIDS in this area would be to provide better data on relative locations and attitudes between the aircraft, weapon, and target.

3.5 Know and Avoid Combat Hazards

A combat hazard is defined as anything which might be directed toward destruction of the tactical aircraft, whether by hostile intent or by mistake. There is clearly a difference between hostile combat hazards and those controlled by friendly forces. Information bases for these two cases differ accordingly, with more data being available from the friendly system to establish location, lethal volume, firing status, coupled target, and other factors. For hostile combat hazards, distinctions are made among three types of threat: surface-to-air missiles (SAM), anti-aircraft artillery (AAA), and aircraft. A certain amount of threat characterization is implicit in each category, and the information needs for avoiding each type of threat are slightly different.

Knowledge of combat hazards reaches a pilot from several sources. The pre-flight briefing usually includes relevant excerpts from a body of information based on intelligence data and observations. This is usually limited to ground-based threats such as SAM and AAA batteries. In flight, he may receive updates to this information from command/control elements or other mission elements. These are transmitted by voice radio, and reach the pilot through the same listening interface which is used for many other information transfers. Warnings of hostile aircraft are provided in the same manner. For friendly anti-aircraft facilities, the pre-flight briefing will indicate location while the burden of avoiding false engagements is placed on the ground system.

These functions are clearly improved by providing a cockpit display of the threat situation, based on position/status and track messages received through JTIDS. The ambiguities, time delays, interference, and lack of cross-registration which are inherent in voice communications are explicitly solved by using JTIDS data to drive a plan-position display of threat locations relative to the aircraft. Hostile SAM avoidance is not improved as much because the uncertainty in lethal area or volume associated with any given site is much larger. For the hostile aircraft case, the existence of a surveillance and tracking capability such as E-3A is assumed.

3.6 Maintain Safety of Flight

This heading is subdivided into three major categories of non-combat hazard to safe flight. The objective of all three operational functions is simply to avoid the indicated hazards. Information needs to accomplish this objective vary with the type of hazard, where the major differences result from the hazard's mobility and whether it is an area or a discrete point. In all cases, information is primarily needed to indicate whether the tactical aircraft and a hazard are likely to attempt to occupy the same airspace in the near future.

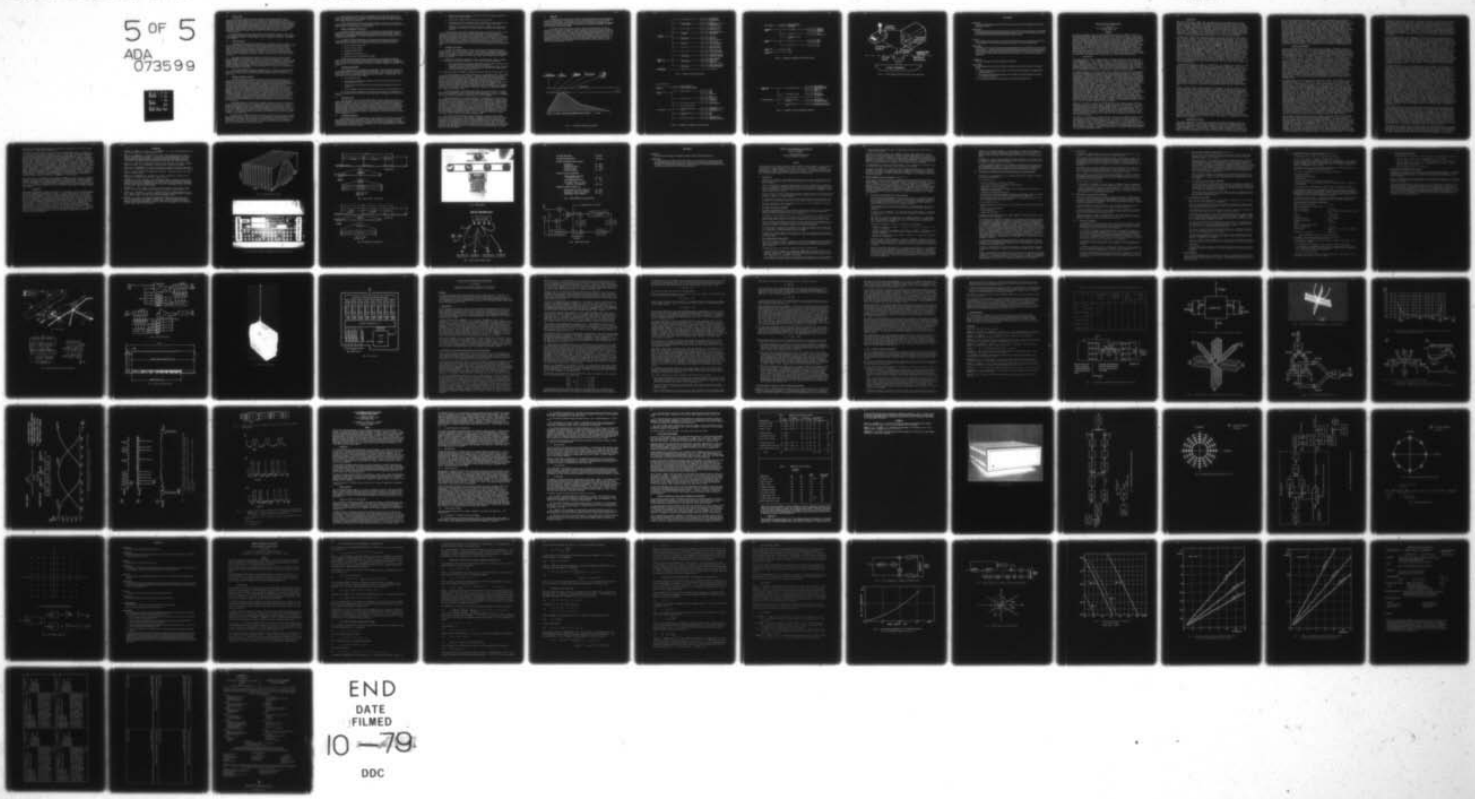
This function area is another case where knowledge of relative position is the key. Avoidance of terrain requires knowledge of terrain characteristics which are obtained from pre-flight briefings and charts, plus knowledge of own-aircraft position relative to terrain hazards based on direct observation or on-board navigation systems. Avoidance of severe weather is a similar situation, except that weather information can change during a mission and must be updated by command/control elements or other mission elements. Separation of air traffic is presently the responsibility of ground-based command and control elements in friendly airspace. Information required by the aircraft is generated at these elements and communicated by voice radio. In hostile airspace where surveillance and control coverage is not available, this function is accomplished through see-and-avoid and the indirect technique of following flight plans which are designed to insure separation.

The keys to achieving improvement through JTIDS are the position location and reporting capability of JTIDS-equipped aircraft and the track reporting capability of surveillance elements which monitor non-JTIDS-equipped air traffic. When a plan-position type of display in the aircraft cockpit shows other aircraft and potential hazards relative to own-aircraft position, the display effectively extends the radius of visual contact for see-and-avoid flying. The greatest benefit will be realized for air traffic separation due to the dynamic nature of those hazards, particularly in hostile airspace where ground-based surveillance and control may not be available.

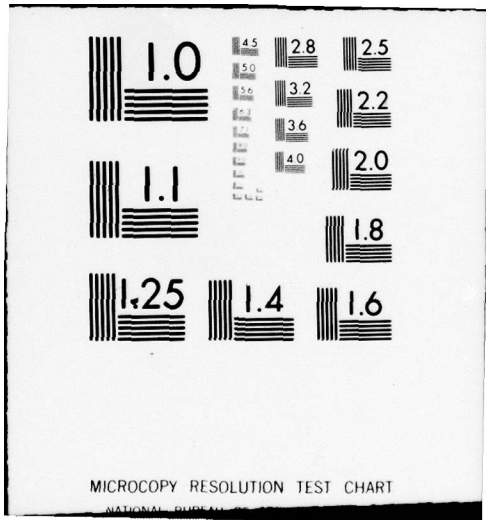
AD-A073 599 ADVISORY GROUP FOR AEROSPACE RESEARCH AND DEVELOPMENT--ETC F/G 17/2
DIGITAL COMMUNICATIONS IN AVIONICS.(U)
JUN 79 H LUEG

UNCLASSIFIED AGARD-CP-239 NL

5 OF 5
ADA 073599



END
DATE FILMED
10-79
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
7.1
8.0
9.0
MICROCOPY RESOLUTION TEST CHART

3.7 Return to Base

This activity includes all functions performed during the mission phase bounded by completion of the last assigned objective and initiation of the landing approach. The basic functions are to leave the objective area, follow the flight plan back toward the base, and find the beginning of the final approach course at the recovery base. These functions are entirely similar to those required to "Get to Objective" (paragraph 3.3) and the information needs are nearly identical. Distinction is made between these two operational functions in order to account for the significant differences in satisfying the information needs--one case being predominately in hostile airspace where enemy systems are geared to defeat these information needs, and the return to base being predominately in friendly airspace (once the aircraft clears enemy airspace) with multiple cooperative systems.

If suitable navigation aids exist in the region covering the return to base flight path, JTIDS does not appear to provide major improvements as a source of navigation information. JTIDS benefits could be significant, however, if other navigation aids are inadequate, do not exist, are disabled, or are destroyed.

3.8 Launch and Recovery

This heading includes those functions performed on and immediately around a friendly airfield. No distinctions is made between takeoff and landing functions, since the information needs are very similar at this level of detail. Precision flight guidance involves very accurate, detailed, and timely data on aircraft position and dynamics relative to the takeoff or landing runway. The other two functional headings involve data which is provided in batches, with the difference being the generality of the data. Clearances and instructions are highly perishable, relating to immediate operations of individual aircraft. "Facility Conditions Information" pertains to the airfield or its equipment rather than to individual aircraft, and it tends to be much less perishable.

The clearance and instructions function is nearly identical to the problem of getting mission requirements to the airborne crew, and this is presently accomplished by voice radio. Facility information is not customized to each aircraft, but is still repeated on an individual basis for many aircraft. For precision flight guidance, the primary Air Force system is the ICAO standard VHF/UHF Instrument Landing System. Backup is provided by Precision Approach Radar (PAR) at some locations, and PAR can be the primary service at forward bases with no ILS. With PAR service, guidance information is communicated from the landing controller to the pilot over a dedicated voice radio channel during the "talk-down".

Precise flight guidance with accuracies comparable to ILS or PAR can be achieved by JTIDS, but only with special processing and dedicated terminals arranged in suitable geometries on the airfield. At the present time, this does not appear more effective than the existing PAR and ILS capabilities.

3.9 Information Generation Functions

All remaining functions identified in Figure 2 involve the output of information from a tactical aircraft. Distinction was made at the top level between the objectives of supporting command/control needs and supporting other mission elements, although the information characteristics overlap totally from the tactical aircraft viewpoint. The basis for defining information needs is, by virtue of the support nature of these functions, determined by the ultimate uses of the information which is generated by the aircraft systems and crew. From the aircraft point of view, data items implied by these support functions can be categorized according to their source and handling. Some data items are generated by avionics and weapon systems without crew intervention. Examples are the position, kinematic, and system status items required for position and status messages on a data link. Other items are derived from crew observations, and cannot be supplied to the outside world without explicit crew involvement. Examples include weather, terrain, and target or threat observations. Still other items could represent a combination of these two, as when the crew actively designates a target whose position is reported on the basis of current aircraft location. The distinction among types of information according to the level of crew involvement in its origination is extremely important when considering the man-machine interface implied by each type of information.

The biggest advantage of JTIDS relative to command and control support is the availability of position data every time the relative navigation function is executed. This data is available within the JTIDS terminal, where it can be automatically combined with other data into a position and status report message and broadcast to all JTIDS-equipped elements. Depending on the aircraft, certain status information can be automatically monitored and included in this message. This could include fuel, ordnance remaining, and various failure or emergency indicators. No cockpit interface is required for this automatic capability, whereas complex interactions are required for most other information output capabilities.

The reporting of targets and threats by voice radio is a simple interaction for the aircraft crew, whereas the data entry required to compose a digital message with the same information becomes complex and time-consuming. Some advantage will accrue from the utilization of JTIDS position estimates with a very limited data entry action to initiate a message which reports something near the current aircraft position. Most information other than the current aircraft location would require a complex man-machine interface for data entry.

Similar observations can be made about the "Report Actual Flight Conditions" function. Limited meteorological data such as outside air temperature, wind velocity and direction can be derived from navigation and air data systems. However, manual data entry into JTIDS will generally require complex man-machine interaction.

Finally, the "Support Other Mission Elements" functions are identical to functions just discussed, as indicated in Figure 2. The effectiveness of JTIDS relative to present capabilities for these functions is therefore the same as indicated above.

3.10 Summary of Implementation Priorities

JTIDS offers significant improvements in effectiveness over present systems to meet the information needs for several operational functions. The previous discussions showed that these improvements are realized only with the proper man-machine interface in the cockpit. Another prerequisite is the implementation of appropriate capabilities in the tactical elements serving as information sources or sinks.

Reviewing the relative effectiveness of JTIDS and existing systems to satisfy the information needs of tactical aircraft operational functions, it is concluded that five basic capabilities should be provided for an operational JTIDS installation in tactical aircraft:

a. Position Location and Reporting,

b. Tactical Situation Display,

c. Jam-Resistant Secure Voice,

d. Command/Control Data Output, and

e. Automatic Interface with Weapon/Sensor System.

Perhaps the most conclusive result of this evaluation is the determination that any information needs which would require complex data entry interactions should not be addressed by JTIDS. In short, the JTIDS installation in tactical aircraft must be designed with a combination of automatic functions and simple man-machine interfaces to be optimally responsive to the needs of the flight crew.

4. IMPLEMENTATION CHALLENGES

Once the capabilities to be implemented through JTIDS on tactical aircraft are identified, the problem of accomplishing the installation must be addressed. Work is currently underway to implement JTIDS on two U. S. Air Force aircraft, the F-15 and the F-16; and preparatory studies have been initiated toward a pod-mounted terminal and installations aboard U. S. Navy aircraft. The installation involves these four major elements:

a. The JTIDS Terminal Set,

b. New or modified avionics whose capabilities are directly associated with JTIDS (the JTIDS Applications Group),

c. Installation of the JTIDS Terminal Set and Applications Group aboard the tactical aircraft, and

d. Support equipment, especially new items which are peculiar to the JTIDS Terminal Set.

The following paragraphs highlight key attributes of the terminal and the applications group, then indicate the basic alternatives toward accomplishing the integration of JTIDS aboard the aircraft.

4.1 JTIDS Terminal Set

The JTIDS Terminal Set is the device which enables a tactical aircraft and its crew to become a JTIDS subscriber. The Terminal Set enables the host platform subsystems and crew to obtain information distributed by JTIDS in a form which is both technically compatible and operationally useful. The terminal also enables the host platform subsystems and crew to use the information distribution capabilities of JTIDS to broadcast operational information which they have generated. Terminal hardware and software elements perform all functions required to extract information from JTIDS and broadcast information via JTIDS while satisfying system standards for signal structure, information format, and net operation protocols. Exploitation of the information extracted from JTIDS, and generation of operational information to be broadcast on JTIDS, are functions of the host platform subsystems and crew.

4.2 JTIDS Applications Group

All hardware and software functions which are not JTIDS terminal functions but which provide new or modified capabilities to generate or exploit JTIDS information are included in the JTIDS Applications Group. Some of these functions may be implemented by modifying hardware or software items previously incorporated within the host platform. They could also be implemented with hardware or software which are either new items or replacements for old items. Examples which can be found in typical JTIDS Application Groups include:

a. Modified or new software modules within the existing central computer program and/or weapon system computer programs.

b. Modified and/or added multiplex bus interface units and cables.

c. Modifications to display processing hardware and software to accept JTIDS data within an existing format (e.g., fire control radar display, flight director, HUD).

d. Replacement of an existing input/output device (e.g., armament control panel) with a display which incorporates the existing functions and additional JTIDS application functions.

A heavy interplay exists between the JTIDS Applications Group (platform) functions and the terminal functions for each installation. Functional requirements for these two groups are established in consonance, and stringent interface controls must be exercised throughout the development phase to insure that all hardware and software items will be compatible. This becomes particularly challenging when the number of unique interfaces is considered along with the fact that JTIDS is being planned for installation in a wide range of aircraft types. The latter point implies that the JTIDS implementation is constrained by existing characteristics of the airframe and avionics, rather than driving those factors. These constraints can become severe, as in the case of some aircraft which have limited space for the installation of a JTIDS Terminal Set and Applications Group.

4.3 Integration Alternatives

The JTIDS Joint Program Office has found that the problem of implementing JTIDS aboard tactical aircraft is not just one problem. Instead, JTIDS integration is actually a separate and unique problem for each aircraft being considered. Since JTIDS is being implemented aboard existing aircraft, the challenge is to exploit the compatible features of existing avionics while minimizing the cost and risk imposed by constraints such as space, weight, power, multiplex bus or computer capacity, and displays.

Based upon a comprehensive assessment of several separate problems, a context is proposed for the consideration of integration alternatives. This context has two major dimensions:

a. The "installation" dimension, which differentiates between internal installation of the JTIDS Terminal Set and a pod-mounted (outboard) installation.
b. A dimension based upon the JTIDS Applications Group, where differentiation is made only between those cases where none is required and where some modified or added avionics are involved with the terminal installation.

This context established four fundamental integration alternatives which are briefly discussed next.

The first alternative is an outboard installation with no applications group. This requires a pod-mounted JTIDS Terminal Set which includes all functional elements required to interface to displays and controls already aboard the aircraft. Since many aircraft have a Maverick missile interface which includes a cockpit display and associated controls, this is a realistic alternative. Figure 6 illustrates the mechanization of a JTIDS/Maverick man-machine interface, with the pod-mounted electronics producing a composite video signal for the existing electro/optical display while the joystick and pushbutton provide the means for pilot input to the terminal. This integration approach is being utilized for the ADM terminal described in paragraph 1.3.c, enabling the pod to be utilized aboard a wide variety of operational aircraft without any special equipment or modifications to the aircraft. This approach is particularly applicable to ground strike aircraft which normally carry external stores.

The second alternative is an internal installation with no applications group. This case applies only if the aircraft is initially designed for JTIDS installation. In the U. S. inventory, the only current possibilities for this alternative involve design changes to aircraft in development (such as the F-18/A-18 or the AV-8B) or future aircraft not yet in development.

The third alternative is an internal installation with some modified or added avionics in the JTIDS Applications Group. Early in the JTIDS program, it was assumed that all tactical aircraft installations would involve this alternative. The high cost and long lead-time to accomplish these installations has forced a reappraisal by the U. S. Air Force. At least for now, the Air Force has determined that internal installations will be made only on prime high-performance fighters -- the F-15 and F-16. A conceptual baseline for the F-15 installation has been developed, and it involves replacement of the existing armaments panel with a new multi-mode display which will support the new JTIDS requirements along with the existing armaments control needs. Tentative concepts have been identified for the F-16, and design studies are being initiated. Current plans call for these two applications groups to be developed in parallel with the pre-production JTIDS Terminal Set, so that the initial operational capability for both aircraft will be achieved in 1984 with an internal installation. Installations for both aircraft can begin in 1984.

The fourth alternative involves an outboard (pod-mounted) installation with a JTIDS Applications Group on the aircraft. This approach is a compromise between the internal installation of the previous alternative and the totally outboard installation of the first alternative. The advantage of this "integrated pod" is that it reduces the magnitude of the aircraft modification problem while permitting maximum utilization of the capabilities offered by JTIDS. This approach could apply to a large number of aircraft types which do not have a prime air superiority role. This alternative would amortize some of the non-recurring costs over a larger population than the internal alternative, since the same pod would fit all aircraft with only minor modifications in interface modules and associated software.

5. CONCLUSION

Implementing JTIDS aboard tactical aircraft is a major challenge to both the developers and the users. One reason for this challenge is that key JTIDS capabilities have no direct precedent in current operational systems. Another reason involves the large population of aircraft, and the variety of aircraft types in which JTIDS will be integrated. By pursuing the implementation program discussed above, the JTIDS Joint Program Office expects to begin operational deployment of JTIDS in tactical aircraft during the mid-1980's.

Once JTIDS is deployed in tactical aircraft, immediate and dramatic benefits should be realized. The first benefit is enhanced survivability of the aircraft, since any real-time data on threats will be available in the aircraft. This is comparable to extending the range, accuracy, and reliability of the crew's vision. The second benefit is enhanced mission effectiveness, since the surveillance and command/control data required for mission execution is available in real time. This data is referenced to the common grid used for JTIDS relative navigation, so the probability of first-pass acquisition will be improved. The combined result of enhanced survivability and mission effectiveness is that more tactical objectives can be satisified per aircraft available at the beginning of the conflict, and these objectives can be satisfied for a longer duration of conflict with more resources remaining at the end.

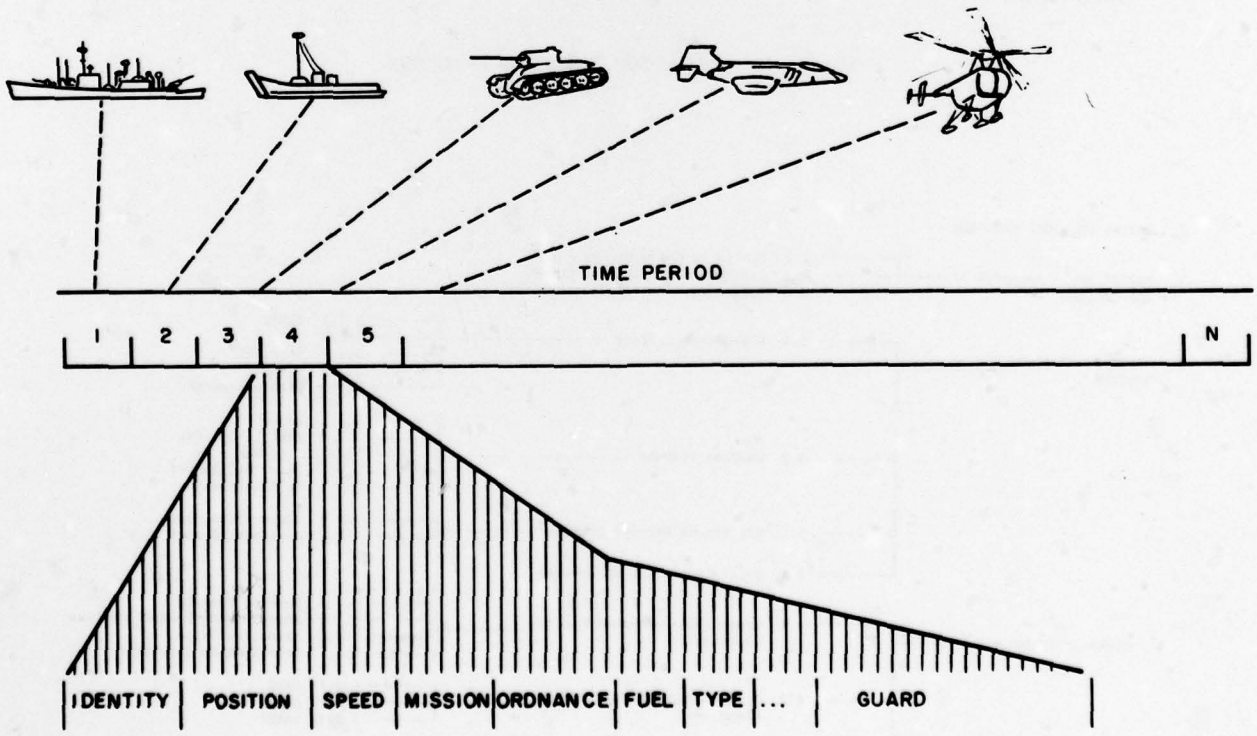


Figure 1. TIME-ORDERED COMMUNICATION SCHEMATIC

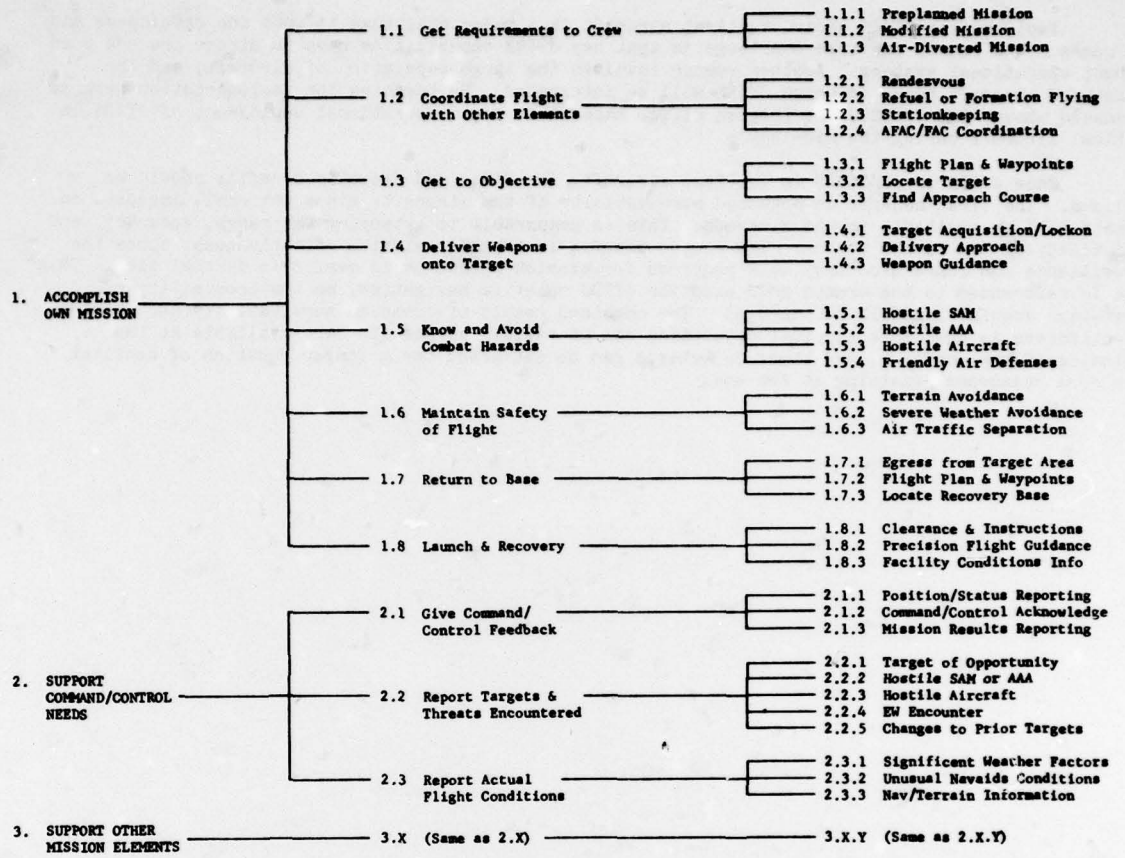


Figure 2. OPERATIONAL FUNCTIONS BREAKDOWN

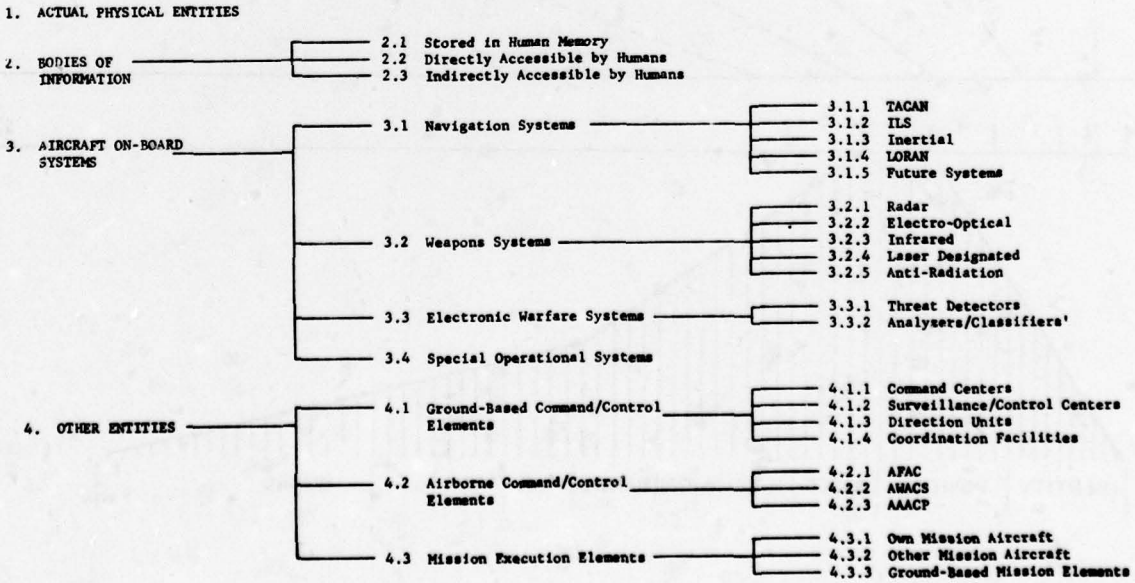


Figure 3. BREAKDOWN OF INFORMATION SOURCES AND SINKS

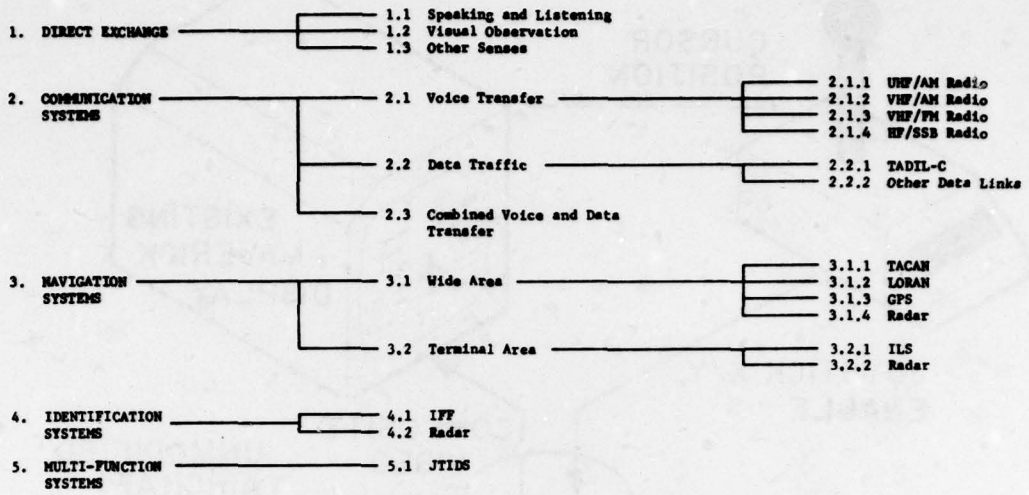


Figure 4. BREAKDOWN OF INFORMATION DISTRIBUTION SYSTEMS

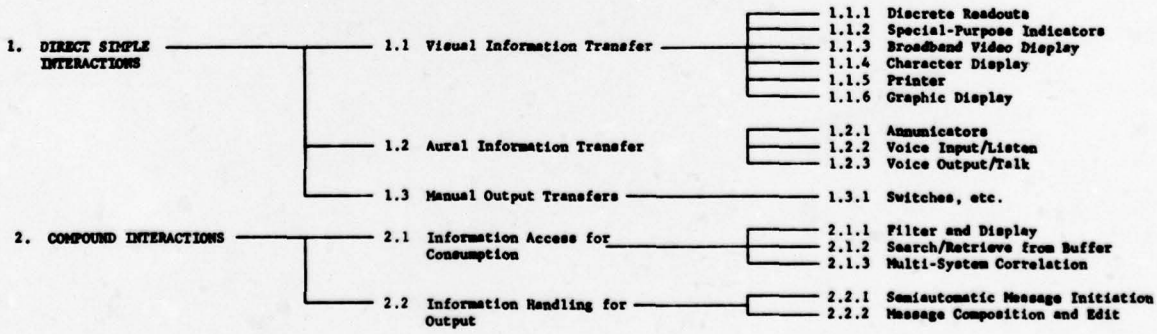


Figure 5. BREAKDOWN OF COCKPIT MAN-MACHINE INTERFACES

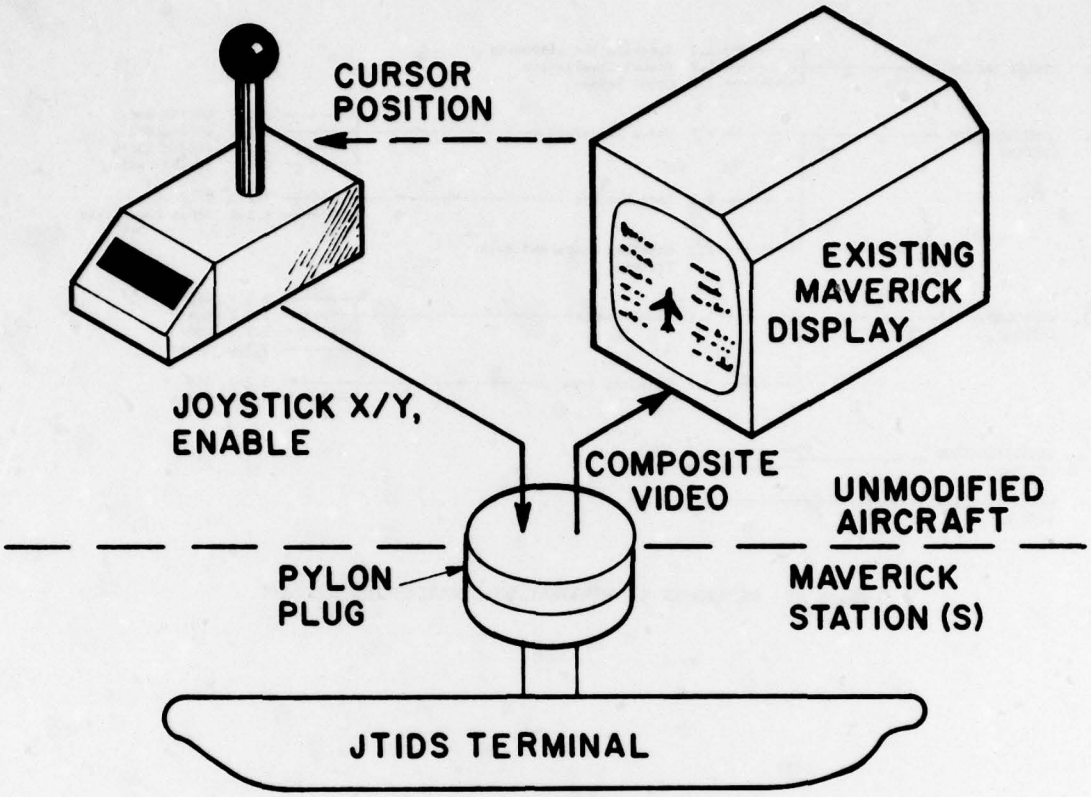


Figure 6. COCKPIT INTERFACE METHOD FOR EXTERNAL JTIDS INSTALLATION

DISCUSSION

C.E.Tate, UK

In the Dynamic Slot assignment protocol described there appeared to be no means of assigning priority for urgent requirements. Is this so?

Author's Reply

Whenever demand does not exceed capacity, no priority is needed. When demand exceeds capacity, the net management protocol basically distributes the shortfall among all surveillance elements. The priority issue then devolves to each surveillance element to determine which reports to broadcast in its slots.

E.Ante, Ge

Can you please give for a typical scenario (1) a typical number of subscribers and (2) a minimun reply rate to fully use the capabilities of communications, navigation and identification, e.g. for a tactical fighter aircraft.

Author's Reply

The JTIDS terminal can remain synchronized by receiving valid position reports or RTT replies from other terminals at very low rates – i.e., intervals of many seconds between replies (exact value classified). Successful operation of Relative Navigation requires at least 3 terminals reporting with good position quality at rates of a few per second (all sources combined). Typical number of subscribers depends on the scenario – how many fighters, surveillance elements, etc.

J.T.Martin, UK

Would you please expand on the dynamic allocation of JTIDS slots.

Author's Reply

As this is the subject of a rather extensive briefing, it is impossible to outline the total protocol. Key elements include:

(1) Subdivision of the time domain into "groups" of time slots, each with a specified use (e.g. fighter position reporting, JTIDS voice).

(2) Use of definite protocols within each group, such as contention (vs. static slot assignment) and dynamic (vs. static) reservation of slots.

(3) Each terminal is programmed with the protocols, so no central net manager is needed.

TDMA FOR RELAYED COMMUNICATIONS

D. L. Baerwald
Rome Air Development Center
Griffiss AFB, NY

SUMMARY

This paper discusses the use of Time Division Multiple Access (TDMA) as a means of sharing the available communication resources of a communication relay, either orbital or otherwise elevated. Work has been done for over ten years at Rome Air Development Center on Demand Access TDMA systems and processing relays to provide solutions to the inherent problems of TDMA and protection from intentional or unintentional interference. The program, conducted primarily by the Ohio State University Research Foundation resulted in the development of both experimental modems and an Adaptive Null Steering Array relay. The relay provides both temporal and spatial adaptive processing to maximize the ratio of recognized desired energy to unrecognized interfering energy. It performs the processing on a pulse-to-pulse basis as each user accesses it during an assigned time slot. In the system developed, each modem synchronizes its clock with that of the relay. This is critical to the rapid processing required of the relay. Other aspects of TDMA such as high peak-to-average power ratio tubes, and network control system software have also been developed. A viable technology for tactical demand assigned TDMA, AJ communications systems has been developed and demonstrated.

1. INTRODUCTION

The newest mode of military communications in everyday operational use today is satellite communications. It is, however, simply a logical extension of other forms of relayed communication, some of which have been in use for many years. The need for relays or repeaters was a direct result of increasing communication traffic requirements. As the traffic increased, the required bandwidth also increased and, thus, the radio frequency required to carry it. Soon only line-of-sight and scatter modes of transmission were available because of the portion of the electromagnetic spectrum being used and its propagation characteristics through the ionosphere and troposphere.

To extend the length of the line-of-sight path, relays or repeaters were soon employed. The elevated repeater, now epitomized by the geosynchronous equatorial communication satellite, was originally treated as every other transmission medium and partitioned by frequency; the multiple users were segregated and assigned portions of the available communication resource on the basis of frequency. The first communication satellites were purely passive reflectors. When the active, hard limiting frequency translating communication satellite was developed, another of the relay's basic resources required allocation consideration: that of available power. At this point, one of the relay transponder's basic resources - time - was still not being considered in terms of allocating communication capability, except in a very gross sense. The total communications resource of a given communication satellite transponder can be expressed in terms of these three axes: Power, Bandwidth, and Time (See Fig. 1). All multiple access systems deal with their allocation, singly or in various combinations. Frequency Division Multiple Access (FDMA), obviously divides users by frequency allocation; each user using all the time axis and sharing the available power. Time Division Multiple Access (TDMA), systems allocate all of the available power and bandwidth to a single user, but only for a fraction of the available time epoch, or Frame. Code Division Multiple Access, (CDMA), segregates users by means of pseudo-random signal coding and bandwidth spreading such that each user can occupy the complete time and frequency axes and shares only the power. Frequency hopping systems are FDMA in nature since they allow a user complete access to neither total instantaneous bandwidth nor power, and are constant envelope signalling schemes.

There is, of course, a virtual, or conceptual fourth axis/dimension which should be addressed; that is the spatial domain. Frequency re-use can be provided by a single transponder through the use of narrow coverage antennas on the satellite. This technique can multiply the useable time-bandwidth product, and increase the available aggregate EIRP as well, due to the higher gain of the narrow beam antennas. It is not considered a primary resource, however, since there are fixed physical limits on its use, those being the location of the earth terminals involved (in particular, their angular separation as seen from the satellite), and the maximum spatial resolution of the satellite antennas.

The object of this paper is to relate the state of the art in TDMA as applied to satellite or other relayed communications systems. The different forms of time-shared communications will be briefly addressed, and the form to be discussed defined. It will address the impact of TDMA operation on other aspects of the communication system's performance such as flexibility, interoperability, anti-jam capability, and equipment economics. In addition, the marriage of TDMA and the adaptively phased array antenna will be discussed. The peculiar advantages of this combination of technologies will be covered along with a detailed description of a breadboard TDMA/ANSAR)Adaptive Null Steering Array Relay) system.

2. DEFINITIONS

The broad definitions of TDMA, FDMA, and CDMA have been established above. Further definition is required, however, to establish the form of time shared communications under discussion. TDMA differs from Time Division Multiplex, TDM, in that the latter converts constant envelope incoming signals to burst or pulse-type signals at a common point. The burst-type signals, therefore, all have their source at one place for timing purposes. TDMA signals, on the other hand, begin their radiated existence as pulses. They, therefore, originate from a distributed source -- spatially, and more importantly, temporally -- with respect to any single point in the system. The difference may seem slight at first consideration but it represents, of course, the most fundamental difficulty in establishing operational TDMA systems, that of system synchronization.

Another point of note is the difference between relayed and non-relayed TDMA systems. The latter are typified by any time-shared use of any radio frequency. The timing may be on a very gross basis, such as one station using it during daylight hours and the other using it at night, or it may be on a very exacting basis such as that used in the JTIDS, (Joint Tactical Information Distribution System), where sharing time is measured in micro-seconds. In either case, there is no absolute system timing reference since the geographical dispersion of the stations prevents it. Obviously in the gross timing example cited this fact doesn't matter since the timing accuracy involved is much looser than the "errors" due to geographical separation. In the JTIDS example, however, it is clear that relative geographical position of the station must be accounted for since even in a line-of-sight network, sufficient variations in terminal-to-terminal separations can exist to cause hundreds of micro-seconds of variation in arrival time of two signals transmitted simultaneously. This problem can only really be solved by allowing sufficient guard time between frames to cover the greatest variation expected. Its effects can, however, be reduced by the use of sufficiently powerful error detecting/correcting codes and controlling the accesses more tightly. These means do not do away with the problem but merely trade off the use of otherwise lost time (guard time) against the probability of intra-system interference and the reduced data throughout due to the coding process.

A relayed TDMA system has a common, or nodal point. It, therefore, has a point at which timing can be established as absolute, (within equipment limitations). While this allows the use of TDMA over much greater distances than the guard time solution could, it is not without penalty. Such a system depends for synchronization upon each station measuring its time separation from the node, or relay. How this information is used varies but it must be done, either actively, by the station concerned, or passively, by the relay. Another disadvantage incurred is the effect of signal overlap when it does occur. In the non-relayed case, signal overlap may not cause a communication problem since the geography involved may greatly favor the desired over the undesired signals at both ends of the problem (i.e., short range desired comm and long range interference). In the relayed case, however, every instance of signal overlap (between equal signal level sources) will cause the same amount of interuption since they not only represent equivalent signal strengths, but also compete for transponder EIRP. These points are made not to imply that relayed TDMA systems are overly burdened with difficulties, but instead to establish the requirements which must be, and have been, met.

Another variable in TDMA systems in general is the means by which the resource is allocated to the system users. Random access is the simplest system and the most common. The commercial (or PTT) telephone system is a random access, time-shared communication system. It depends, as all random access systems do, on the statistical randomness of discrete calls and the relatively low percentage utilization of each subscriber. It is quite efficient, and highly successful during "normal" conditions. Long periods of peak traffic can render it almost useless. A second type of allocation is referred to as dedicated access. This system provides a given number of time slots to a user on a dedicated basis for use whenever he needs them. For certain high percentage utilization circuits, such as trunk lines, etc., where system configuration rarely changes, this type of allocation is quite effective. It requires time frame/epoch establishment and perhaps some form of monitoring or policing as well. Its major limitation, however, is its inability to efficiently use the system's resources under highly dynamic traffic conditions. The third major allocation system, demand assignment, is specifically tailored to this type of traffic scenario. It is perhaps the most complex of all three, and requires expending some communication resource for "overhead" to manage the allocation procedure, but is the most efficient scheme when the traffic loading is heavy and rapidly changing. It also allows a system wherein there are more users than "channels" (like the random access system) and wherein not all users have the same priority or urgency (like the dedicated system). The rest of this paper will deal primarily with the demand assigned, relayed type of TDMA system, as configured for tactical communications networks.

3. SYSTEM IMPACT OF TDMA

Time-shared communication systems allow "sharing" of terminal equipments as well as transponder resources. In a multiply connected tactical network, simultaneous comm links can be established with a single transmitter/receiver and modem/processor. The operator interface can either be serial, through a single I/O device or in parallel, through multiple I/O's. In either case, a single processor sequentially receives, decodes, and buffers multiple input pulse streams and either stores the resultant information for serial presentation, or routes it directly to an available printer, vocoder,

etc. The economy provided is evident. In addition, interoperability is enhanced as various codes (ASCII, etc.) unique to specific I/O devices can be handled simultaneously. This aspect extends much further when considering the interoperability problem of the satellite or other relay transponder. The time-ordered nature of TDMA allows extreme system interoperability and intercompatibility. Since only one user is "on" the transponder at any given instant, his power, bandwidth, modulation waveform, etc., are not at all constrained by other users' characteristics. This means that no "power control" system is needed, as with hard-limiting FDMA, and small (EIRP and/or G/T) users can coexist quite readily with very "large" users. The extremely flexible demand assigned TDMA system merely allocates the number of basic time quanta, (slots), per frame necessary for the data rate that the terminal is capable of. If the terminal has sufficient EIRP, it can transmit multiple data streams, at different bit rates, from separate I/O's, completely independently. This allows a single terminal to "simultaneously" communicate with several terminals of different G/T simply by adjusting the data rate of each comm link. Since the two signals are not really present at the same instant in either the ground terminal transmitter or the relay transponder, the intermodulation problem is avoided. Other beneficial system effects of TDMA operation will be discussed in connection with the description of the experimental TDMA/ANSAR system.

4. STATUS OF DEVELOPMENT

Rome Air Development Center has been pursuing the development of TDMA for satellite communication since 1967. Most of the basic system analysis and design has been done under contract, by the Ohio State University Research Foundation at Columbus, Ohio, under the direction of Dr. Ronald Huff. The preliminary work centered around the network synchronization problem and the performance of various phase locked loops when subjected to pulsed envelope signals. (REINHARD, K.L., 1967; WALLACE, K.A., 1968; HUFF, R.J., et al, 1969; HUFF, R.J., 1969; HUFF, R.J., REINHARD, K.L., 1971) After about four years of work, the OSURF team did an analysis of the state of the art and concluded that ". . . a TDMA system which is efficient, effective, and flexible can be instrumented." (HUFF, R.J., 1971) At this point, it was decided to implement a first generation, breadboard TDMA system. The result of this decision was the construction of 4 modems such as shown in figure 2. The modems, which were completed in 1974, were designed to demonstrate as many of the operationally desirable features as possible within the constraints of technology and financial resources. As a result, many of the parameter values chosen are neither the best that technology could support, nor the optimum operational choice. Instead, they were chosen primarily to demonstrate what could be done. As an example, the modem's internal processor could have accomodated up to seven separate, addressable I/O's; it was implemented with capacity for only two to save memory and software while yet demonstrating multiple I/O capability. In like manner, only two levels of precedence were implemented, and only one incoming message storage buffer. In both cases the multiplicity capability was thus proven, but at minimum expense.

The modem operates at 70 MHz with two basic data input rates, 75 and 2400 bps, and two pseudo-noise coding rates. (TAYLOR, R.C., HUFF, R.J., 1976) The "low rate" format operates at a burst data rate of 11 or 87 Kbps, representing 16 or 2 chips per bit respectively, at a PN rate of 175 Kch/s. The "high rate" format operates at 87 or 700 Kb/s data burst rate, representing again 16 or 2 chips/bit at a PN rate of 1.4 mch/s. The two formats represent two different basic data transmission rates (and therefore network capacity). They are implemented in the modems to demonstrate different types of service available, i.e., to fewer, disadvantaged users (i.e., aircraft, small mobile terminals, etc.), or to many more "large" terminals able to provide the greater EIRP and G/T required for higher burst rate transmission. Figures 3 and 4 show the basic frame formats for the low and high rate formats respectively. Network control is implemented by use of overhead time slots to request a link/slot assignment from the net control station. Any station can perform the net control function if the modem's memory has been properly loaded, and it has the necessary processor. The network control management is done by an H.P. 2100 processor which is also used to load the necessary instructions into the modem's memory. Upon request from a net member, the NCS determines if the called station is busy and if so, the precedence of his traffic, and the location of available time slots. If all is in order, the NCS calls the called station and gives it the address of the calling station and a slot assignment. It then informs the calling station and the link is established. Timing is provided by a clock in each modem synchronized to the 'net master' clock. The source of the master clock can be any of the user terminals and can be transferred during operation. Ideally, the clock would originate at the repeater. The modems were given the capability to provide the clock through the satellite to allow demonstration of TDMA operation over simple communication satellites which were not designed to accomodate TDMA operation.

The individual users synchronize their own internal clocks with the satellite-relayed master clock by transmitting during their linking/ranging (L/R) slots. A coarse estimate of one way range is set into the modem by the operator. The modem then automatically adjusts its clock so that its transmissions reach the satellite exactly during the assigned slot time. The accuracy with which this is done is a key point in the development of a sophisticated TDMA/ANSAR system. It was decided at the outset that the modem should synchronize its own clock to the net master clock within the accuracy of a PN chip, rather than just to within a fraction of a slot. This was done for several reasons. 1) It allows a single modem to receive and correlate multiple incoming messages with only one clock and no lengthy receive synch procedure, as long as its own clock is also synched to the master PN code. 2) It allows future processing satellites to process multiple uplink signals without any search and synch procedure. 3) It allows

an adaptively phased array on-board a satellite to begin its spatial processing without requiring a temporal search as well. The modems are therefore configured to provide bit-synchronous operation with the relay clock, the burden of synchronism being completely handled by the transmitting station. This timing technique, plus the use of differentially coherent PSK for data modulation and detection allows the use of an extremely short (one bit) preamble at the start of each burst. This keeps the preamble overhead low enough that short bursts can be used without a significant percentage overhead resulting. (TAYLOR, R.C., HUFF, R.J., 1976) The experimental modems have been successfully operated over both the DSCS-II and LES-6 satellites. The system worked as designed and the network clock source location was transferred between terminals without incident.

As mentioned previously, the combination of TDMA with an adaptive antenna provides a uniquely capable relay system. (REINHARD, K.L., 1973) Since the separate users are not transmitting at the same instant, and are all sending synchronized PN codes, the array can sequentially form a beam on each user as his time slot occurs. This not only enhances the signal to thermal noise ratio of the uplink signals but also provides discrimination against undesired signals and interference. Figure 5 is a picture of an experimental ANSAR, also designed and developed by OSURF. (MILLER, T.W., et al, 1976) It not only incorporates the beam forming capability but actively forms nulls in its pattern toward any source of in-band energy not possessing the proper PN coding, as shown conceptually in figure 6. The processing algorithm actually maximizes the ratio of desired to undesired signal in a least-mean-square (LMS) sense. This type of algorithm, which produces the best S/I ratio under the signal and geometry conditions extant, is not useable with FDMA or CDMA networks, due to the multiplicity of constant envelope users' signals present. If the desired and interference signals have sufficient angular separation as seen from the repeater, an adaptive array which only forms nulls on interference would provide satisfactory service. When the separation decreases to near the limit of the array's resolution, however, the deepest null may not provide the highest S/I ratio. The ability to process with the signal identifier, LMS algorithm which is only available to TDMA systems, is thus beneficial under geometrically difficult conditions.

The ANSAR shown was built for an RF frequency in L-Band, simply for convenience. The processing is done at 70 MHz IF. The array receives at 1650 MHz and re-transmits at 1550 MHz. Its other characteristics are shown in figure 7. It has a four element receive array and uses a single element for re-transmission. The elements are helices with a gain of approximately 10 dB. They are mounted so that the array spacing can be changed easily to experimentally verify predicted array performance with varying geometry. Each element is followed by circuitry which divides its signal into in-phase and quadrature components. By appropriately "weighting" the output of each, any desired phase and amplitude can be produced in each element's output signal. The weighting circuits are driven by the error voltage obtained from comparing the combined processor output signal with the internally generated reference code, (See figure 8). The loops thus are closed and drive the weighting circuits to minimize the non-correlated output and maximize the ratio of desired (correlated) to undesired signals. By independently changing the phase and amplitude weights of each element, the processor has effectively changed the pattern of the receiving array. This helps to overcome one of the primary drawbacks of pulsed envelope communication systems, i.e., that the instantaneous or burst rate must be higher than the base-band data rate by a burst factor of at least 1/R where R = the effective "duty cycle" of a user's transmission. For the experimental modems, a 75 Bps input data stream is transmitted at a burst rate of 11 KBps in the low rate format. This represents a burst factor of about 22 dB. This means that the signal to noise and interference ratio at the receiver must be 22 dB higher than that required for a constant envelope 75 Bps transmission under otherwise equivalent conditions. Another disadvantage of a relayed TDMA system is the fact that each user must compete individually with receiver noise and any intentional interference (jamming) present at the satellite input. In FDMA or CDMA systems, the jammer must compete with the aggregate input power in order to gain a suppression effect from the hard limiter. Because of these two weaknesses of TDMA, the adaptive array is particularly valuable, to increase the received signal level and to provide signal to interference discrimination.

Another means by which these two problem areas can be ameliorated deals with the type of final amplifier used in the ground terminal transmitters. Since the TDMA signals are essentially pulses, with a duty cycle of 2 - 5%, a transmitter with a higher peak power than average power capability can be used to advantage. This is similar to the concept of radar transmitters where performance is determined more by peak power available than by average power. Work has been sponsored by RADC to develop high peak-to-average power ratio tubes for TDMA communications in both the UHF and X-Bands. In both cases, the tube developments were basically modifications of already available radar-type tubes, to provide longer pulses and somewhat greater duty cycles. The UHF tubes were developed by Eimac Division of Varian and a transmitter using them was built by General Electric. It provides a peak output power of 20 Kw at a duty cycle of up to 3%. Its average power is 600 watts. Development of the SHF version was stopped when it was learned that the U.S. Army Satcom Agency was pursuing a similar program.

The technique of raising the peak-to-average power ratio is aimed primarily at smaller terminals since for larger, high power transmitters one would quickly exceed reasonable peak powers. The smaller terminal, however, both needs higher power more urgently, and can least afford the size and weight of high average power equipment. It, therefore, can benefit significantly from this means to allow it more signal at the satellite --

both to help overcome thermal noise and saturate the transponder, and to help compete with natural or intentional interference.

In order to allow the efficiency which a Demand Assigned TDMA network is capable of, time slots must be controlled, allocated, etc., at high speed. The network management is, therefore, as important to efficient operation as the hardware itself. The net control software for the experimental system was implemented on an HP 2116 originally, and later re-configured for the 2100. It was developed jointly by OSURF and Computer Sciences Corp. Most of its features have been discussed above. The total link establishment time, assuming available resources and terminals, is on the order of 8 seconds. This is about the time necessary to dial a telephone. It should be remembered that four round trip circuits through the synchronous satellite are included in this time. The controller is as flexible as the modems; it can address I/Os directly, through a common or different modems. It can perform priority/precedence over-ride operations, and direct a message to a "busy" modem's storage buffer for later printing if the buffer is empty. The software on the 2100 also includes a network simulation model. This allows exercising of a few modems with a simulated network of other users. The simulated traffic model can either be randomly generated by the processor or an input determined by the operator. This permits the demonstration of modem performance under known, simulated traffic scenario conditions. (FRENKEL, G., et al, 1973)

As a final step in the exploratory development of TDMA relay comm systems, a completely software, non-real-time simulator was developed. The purpose was to simulate timing formats, data rates, etc., beyond those of the experimental modems. This simulation was written in FORTRAN IV for the RADC Honeywell 6180 computer and was just delivered in 1977. It is presently being installed and checked out and plans are underway to modify it to make it even more flexible so it can also handle non-relayed, non-demand-assigned TDMA networks.

5. CONCLUSION

The technology required to produce an efficient, interference resistant Demand-Assigned TDMA relay communication system is available. In its present form, it can accomodate bit rates of several megabits/second. Work is underway to extend this to rates of several tens of megabits. For most tactical communication needs this will be sufficient. Although the work has been done primarily for satellite relays, the technology is equally applicable to any non-orbital elevated platform, be it airborne, balloon-borne, etc. The processor requirements are essentially identical. The antenna array faces a somewhat different geometric situation in the case of a non-orbital relay platform, however. In this situation, the array must provide a nearly hemispherical coverage pattern. The situation in regard to intentional interference is somewhat eased, however, since for reasonable earth-surface separations of desired and undesired sources, the resolution required of the array is much more easily achieved. Since the non-satellite system would be much less costly to implement and deploy, it is expected that the first experimental fielded tactical TDMA relay communication system will utilize airborne transponders.

REFERENCES

1. FRENKEL, G., KINAL, G.V., GAN, D.G., and DOGGETT, J.E., 1973, "TDMA Network Control Study," RADC-TR-73-270, Computer Science Corp.

2. HUFF, R.J., REINHARD, K.L., and UPP, D.C., 1969, "The Synchronization of Time Division Multiple Access Systems -- An Analytical and Experimental Study," Report 2358-9, ElectroScience Laboratory, Electrical Engineering Department, The Ohio State University Research Foundation; prepared under Contract F30602-67-C-0119 for Rome Air Development Center, Griffiss Air Force Base, New York (AD 689 223)

3. HUFF, R.J., 1969, "An Investigation of Time Division Multiple Access Space Communications Systems," Ph.D. Dissertation, The Ohio State University, Columbus

4. HUFF, R.J., and REINHARD, K.L., 1971, "A Delay-Lock Loop for Tracking Pulsed-Envelope Signals," IEEE Trans. Aerospace and Electronic Systems, Vol. AES-7, pp. 478-485

5. HUFF, R.J., 1971, "TDMA Space Communication Systems: Concepts and Practical Techniques," RADC-TR-71-255

6. MILLER, T.W., CALDECOTT, R., and HUFF, R.J., 1976, "A Satellite Simulator with a TDMA-System Compatible Adaptive Array," RADC-TR-76-98

7. REINHARD, K.L., 1967, "Analysis of a Pseudo-Random Network Timing System for Time-Division Multiple Access Communications," Report 2358-2, ElectroScience Laboratory, Department of Electrical Engineering, The Ohio State University Research Foundation; prepared under Contract F30602-67-C-0119 for Rome Air Development Center, Griffiss Air Force Base, New York (AD 820-800L)

8. REINHARD, K.L., 1973, "Adaptive Antennas for Coded Communication Systems," Report 3364-2, The Ohio State University ElectroScience Laboratory, RADC-TR-74-102

9. TAYLOR, R.C., and HUFF, R.J., 1976, "A Modem/Controller for TDMA Communication Systems: Report 3364-5, The Ohio State University ElectroScience Laboratory; prepared under Contract F30602-77-C-0162 for Rome Air Development Center

10. WALLACE, L.A., 1968, "The Tracking Performance of Sampled-Data-Delay Lock Loops with Pulsed Envelope Input Signals," Report 2358-7, ElectroScience Laboratory, Electrical Engineering Department, The Ohio State University Research Foundation: prepared under Contract F30602-67-C-0119 for Rome Air Development Center, Griffiss Air Force Base, New York (AD 821 657)

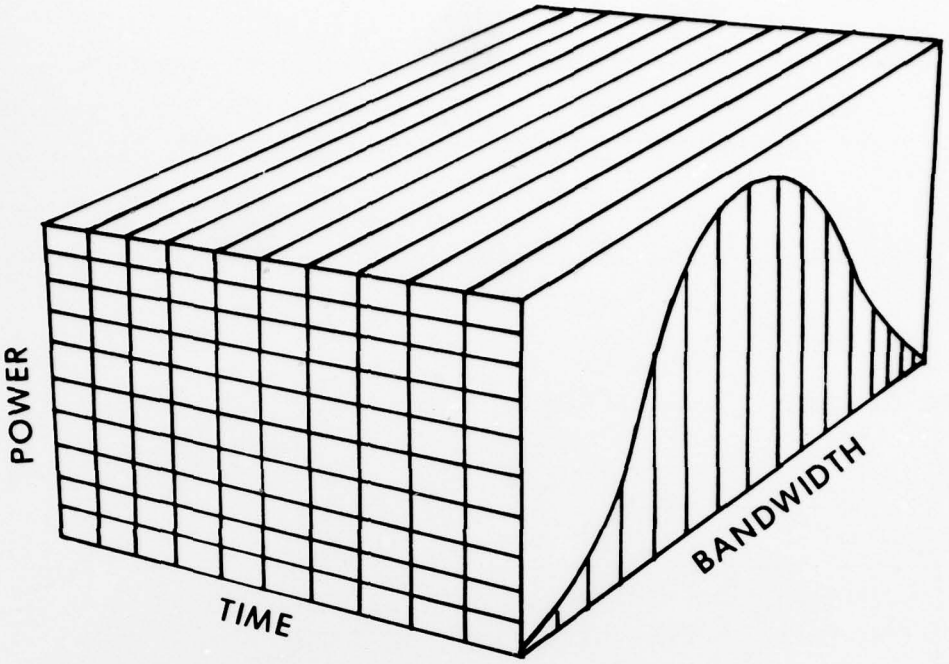


Fig.1 Communications relay resources

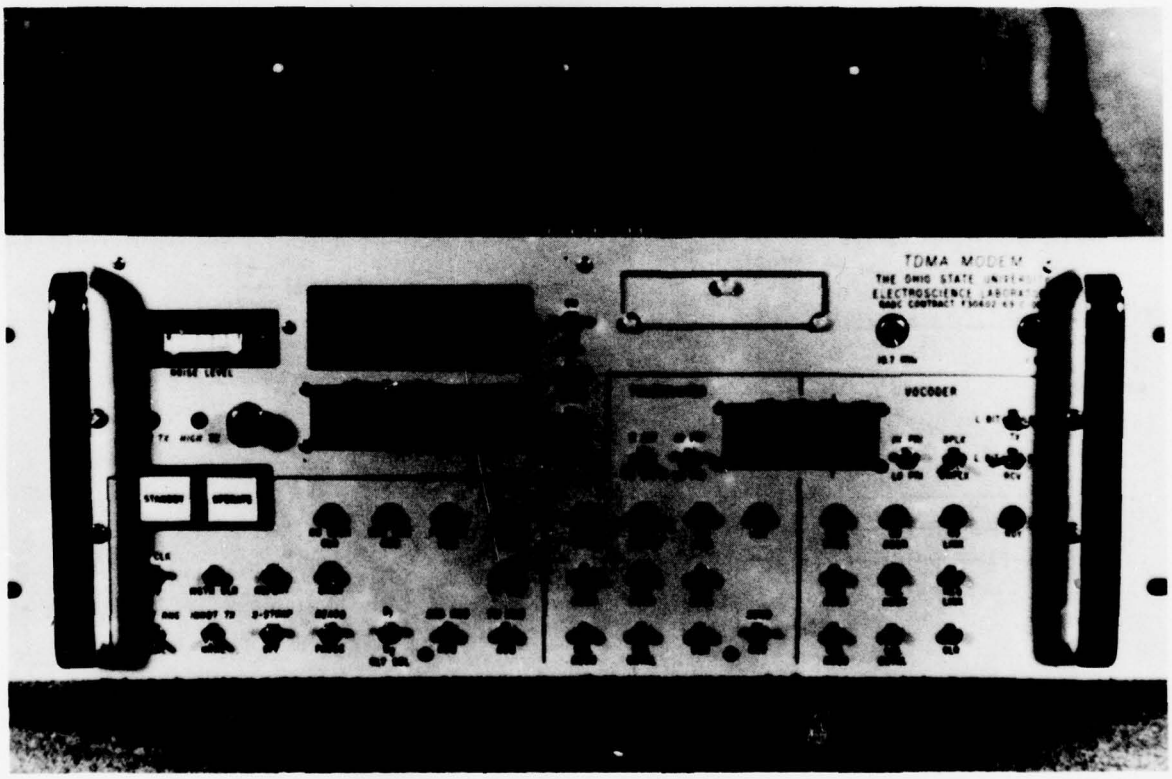


Fig.2 Time division multiple access (TDMA) modem/synchronizer/controller

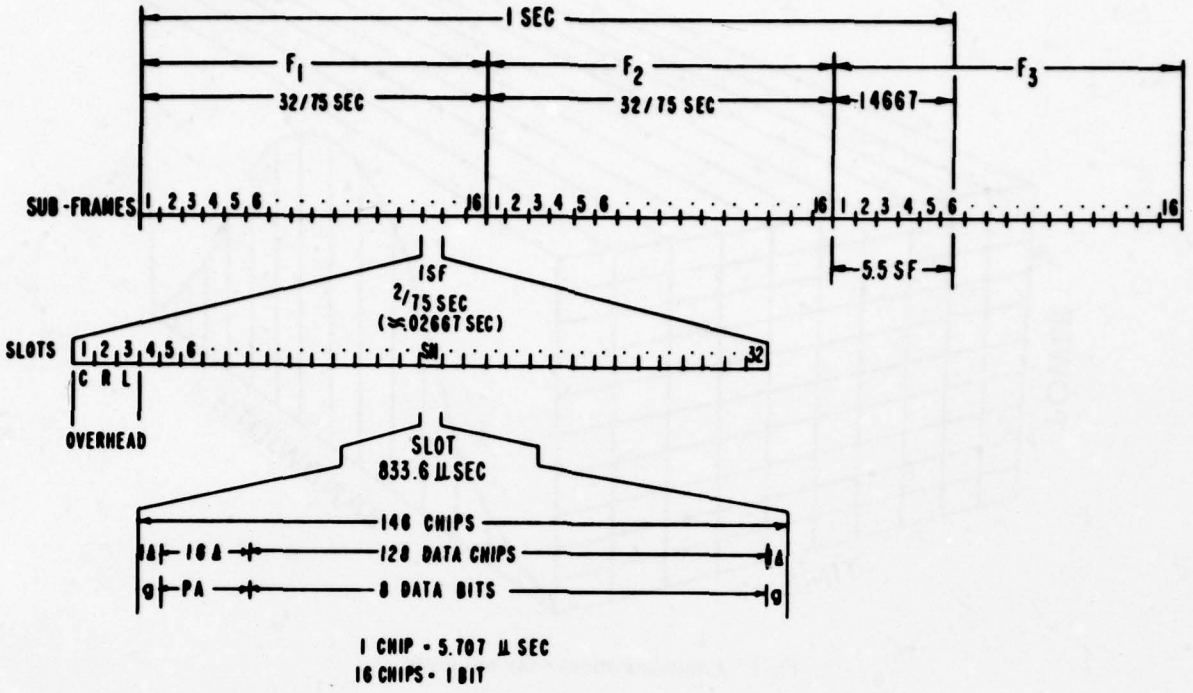


Fig.3 Low rate format – low rate mode

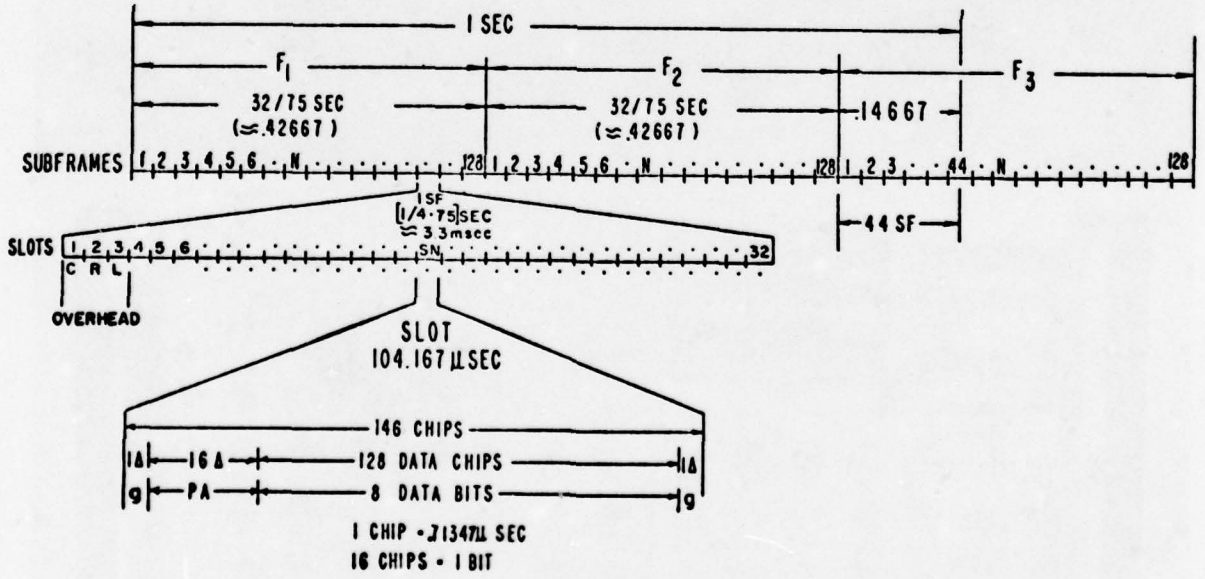


Fig.4 Hi rate format – low rate mode

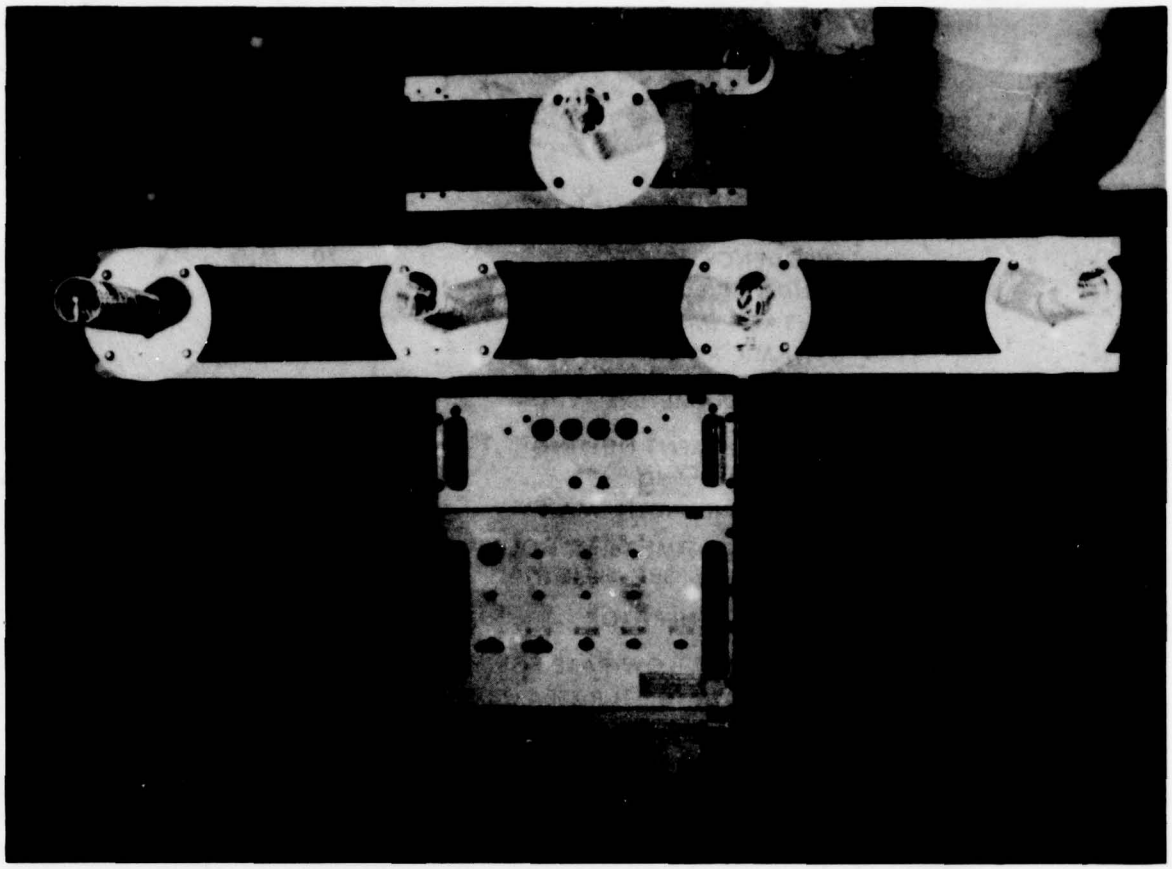


Fig.5 TDMA ANSAR

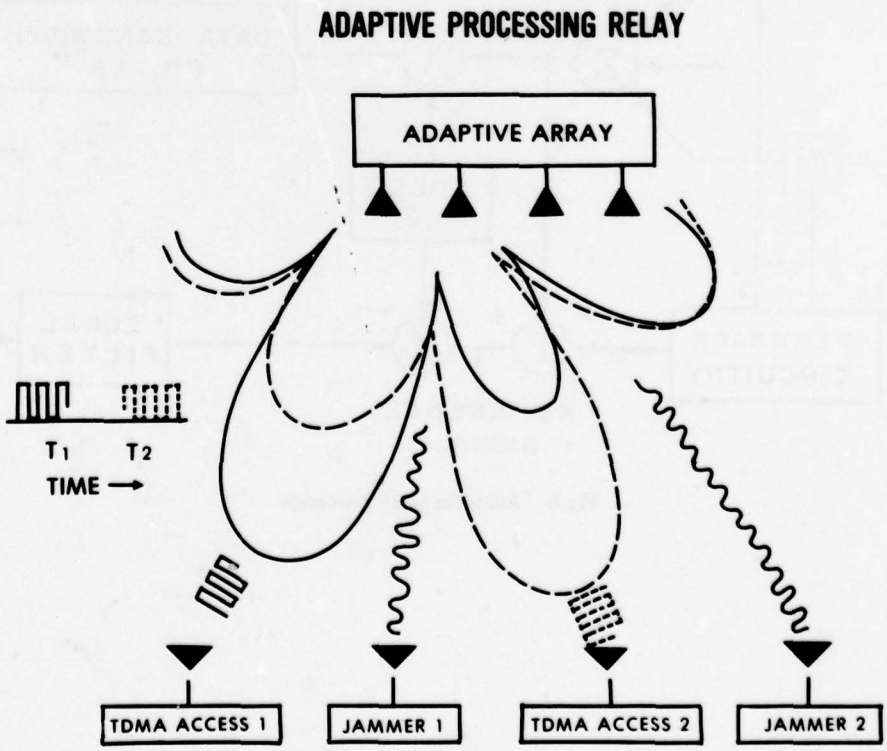


Fig.6 Adaptive antenna nulling for TDMA

RECEIVER FREQUENCY	**1650**	**MHz**
TRANSMITTER FREQUENCY	**1550**	**MHz**
WEIGHT CONTROLLER INPUT SIGNAL		
FREQUENCY	**70**	**MHz**
BANDWIDTH (MAX)	**13.5**	**MHz**
VOLTAGE (MAX)	**1**	**V p-p**
DYNAMIC RANGE	**50**	**dB**
ADAPTIVE LOOP PARAMETERS		
SINGLE ELEMENT NULLING CAPABILITY (MAX)	**65**	**dB**
LOOP BANDWIDTH (MAX)	**2**	**MHz**
SWITCHING TIME (e.g. "HOLD" WEIGHTS, "SET" WEIGHTS)	**500**	**nsec**
REFERENCE SIGNAL GENERATOR		
BANDWIDTH (LOW RATE FORMAT)	**44**	**KHz**
BANDWIDTH (HIGH RATE FORMAT)	**350**	**KHz**
PROCESSING GAIN (APPROX)	**12**	**dB**

Fig.7 TDMA ANSAR operating characteristics

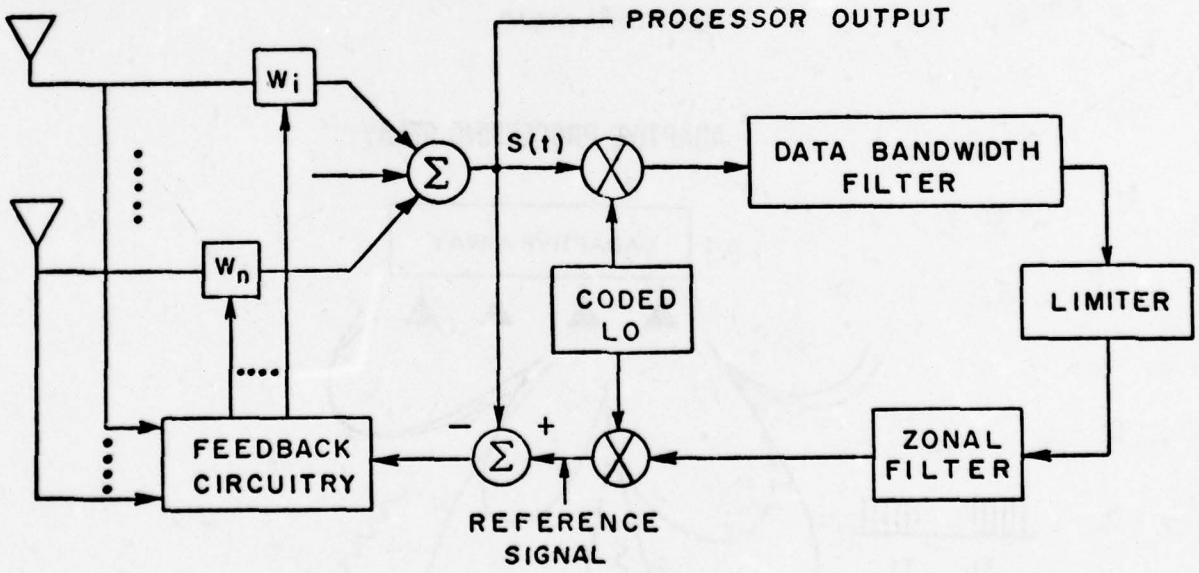


Fig.8 Adaptive array processor

DISCUSSION

J.Buchau, US

Don't secondary nulls produced while nulling out jammers exclude involuntarily possible users?

Author's Reply

No; uncontrollable lobes or nulls are only of concern when the antenna elements are *periodically* spaced greater than half-lambda (when grating lobes come into the "visible region") we carefully avoid periodic spacing to allow the required array resolution without grating lobes. Actually the problem of a grating lobe presenting a null is more serious than an unwanted null, due to the S/I maximizing algorithm.

RESEAU DE RADIOCOMMUNICATION NUMERIQUE EN DUPLEX TEMPOREL

par J. Lautier
Société LE MATERIEL TELEPHONIQUE
92103 Boulogne-Billancourt

RESUME

L'établissement de liaisons numériques duplex radioélectriques demandait, jusqu'à présent, l'utilisation de deux fréquences, l'une pour l'émission, l'autre pour la réception. Aujourd'hui, le duplex temporel permet d'utiliser une seule fréquence porteuse affectée automatiquement tantôt à l'émission, tantôt à la réception, l'utilisateur pouvant cependant parler et écouter simultanément. La modulation Delta utilisée pour cette numérisation se prête particulièrement au chiffrement de la communication.

1 - INTRODUCTION

L'étude de ce réseau de radiotéléphone automatique a été faite pour équiper les terrains de déploiement des Forces Aériennes Tactiques de l'armée française (Fig. N° 1, Plan du Réseau).

C'est un réseau radio raccordé automatiquement au réseau téléphonique de l'Armée de l'Air. Il s'adresse à des usagers mobiles répartis sur un terrain de dimensions limitées et pratiquement à portée optique de la station d'émission. L'Equipement Terminal d'Usager doit être de dimensions très réduites, avec alimentation incorporée et facilement transportable à la main.

L'originalité de ce réseau radiotéléphonique de dimension modeste, réside essentiellement dans l'emploi d'un mode d'émission-réception n'ayant pas encore eu d'application pratique en France : le Duplex Temporel.

Ce mode d'émission-réception peut être appliqué dans d'autres types de réseaux, car il présente d'intéressantes possibilités liées comme toujours à quelques servitudes.

Nous nous efforcerons, dans la description qui suit, de faire apparaître les possibilités intéressantes de ce système.

2 - LES TYPES DE LIAISONS DEMANDEES

Le réseau radiotéléphonique étant composé d'équipements portables (E.T.U. Equipement Terminal d'Usager) reliés par radio à une station centrale, plusieurs types de liaisons peuvent être établies entre ces équipements.

D'une part : des liaisons entre ETU, avec ou sans la présence de la station centrale,

d'autre part : des liaisons entre ETU et station centrale (et vice versa) vers un abonné téléphonique lointain, d'un réseau téléphonique fixe couvrant tout le pays.

Les utilisateurs d'équipements portables ETU (32 au maximum) sont dispersés sur un terrain de dimensions réduites (cercle de 3 km de rayon), mais dont la nature (terrain d'aviation) proscrit les simples liaisons par câbles téléphoniques.

La liaison radio s'impose donc en duplex, et la gamme attribuée a été la gamme UHF (410 - 470 MHz). L'importance présumée du trafic requiert l'utilisation d'un groupe de 8 canaux radio, situés n'importe où dans la gamme.

Ajoutons que certains usagers sont prioritaires, que les connexions doivent pouvoir se réaliser sans intervention d'un opérateur, et que la discrétion des liaisons radio doit être assurée.

3 - LES DIVERS MODES DE REALISATION

Dans un des types de liaisons, l'usager du réseau radiotéléphonique est en communication avec un abonné du réseau téléphonique. Il est donc nécessaire que la liaison radio soit en duplex. Le choix est alors possible entre le duplex bi-fréquences et le duplex temporel.

Le duplex bi-fréquences présente d'emblée plusieurs inconvénients :

- l'utilisation de 2 groupes de 8 fréquences, séparés par un intervalle fixe, par exemple 10 MHz, restreint la possibilité de déplacement des canaux utiles à l'intérieur de la bande 410 - 470 MHz, où seulement certaines bandes de 2 MHz de large sont disponibles aux utilisateurs militaires,
- l'utilisation de filtres Duplex accordables électroniquement sur une large gamme en UHF, délicats à réaliser surtout sur des portables et sensibles aux brouilleurs,

- Les protections nécessitées dans les équipements portables par l'émission et la réception simultanées.

La solution en Duplex Temporel, qui consiste, dans un réseau synchrone, à faire émettre à la Station Centrale les 8 émetteurs en même temps pendant que tous les mobiles sont en réception, pendant un court intervalle de temps, puis à inverser la situation au cours de l'intervalle de temps suivant, apparaît assez séduisante. Elle apporte, en contrepartie, une certaine complexité supplémentaire, mais uniquement par adjonction de circuits logiques en B.F., peu coûteux, fiables et facilement intégrables en microélectronique, et nécessite l'utilisation de parole numérique.

Le fonctionnement du réseau en Duplex Temporel, et les avantages et inconvénients de ce système soit décrits en détail dans la suite de cet exposé.

Le besoin de discrétion des liaisons conduit presque obligatoirement à utiliser une modulation numérique, ici le Delta à 16 kbits, compromis actuel entre une qualité minimum et une largeur de bande occupée trop importante. Il est possible, en effet, sur la parole codée, d'introduire facilement un dispositif de discrétion simple, peu coûteux et efficace.

4 - RESEAU SYNCHRONE EN DUPLEX TEMPOREL (figure N° 2)

La figure N° 2 montre le schéma d'un équipement émetteur-récepteur fonctionnant en Duplex Temporel. Comme dans un équipement Simplex à modulation analogique, il y a un circuit émetteur et un circuit récepteur reliés à une antenne, à travers un dispositif de commutation d'antenne d'émission en réception. Comme dans un équipement Simplex, il n'y a pas de filtre duplexeur, et la fréquence de fonctionnement est la même en émission et en réception. Les circuits du récepteur n'ont pas à être particulièrement protégés puisqu'ils sont déconnectés lorsque l'on émet. Par contre, des circuits nouveaux destinés au traitement de la modulation, à savoir : les circuits d'émission-réception en temps partagé et les circuits de numérisation de la parole, sont ajoutés.

4.1 - Principe de fonctionnement en Duplex Temporel

Le principe du Duplex Temporel n'est pas très nouveau, et a déjà été proposé par certains expérimentateurs, utilisant des moyens de stockage électromécanique d'un signal de parole analogique. Ils se heurtaient cependant jusqu'ici à des problèmes technologiques de réalisation qui sont maintenant résolus en parole numérisée, avec les circuits logiques modernes.

Pour réaliser la fonction "émission-réception" en temps partagé, sur une seule fréquence porteuse, il suffit de traiter les signaux numériques modulés en Delta par :

1) deux cellules de mémoire ; 2) un ensemble de portes permettant l'écoulement du signal ; 3) un dispositif d'horloge et de commutation contrôlant l'écoulement du signal.

La figure N° 2 montre que chaque cellule de mémoire (mémoire 1 et mémoire 2) peut être reliée, soit à l'émetteur, soit au récepteur, soit au codeur numérique, soit au décodeur numérique. Ces liaisons sont établies alternativement avec la mémoire 1 et la mémoire 2.

Considérons la mémoire 1 - Elle a 3 régimes de fonctionnement :

- pendant un intervalle de temps Δt, (niveau 1) l'entrée de la mémoire 1 est reliée au récepteur,
- pendant un intervalle de temps $2\ \Delta t$, qui suit, (niveau 2), l'entrée et la sortie de la mémoire 1 sont respectivement reliées au codeur numérique et au décodeur numérique,
- pendant un nouvel intervalle de temps Δt (niveau 3), la sortie de la mémoire est reliée à l'émetteur.

Analysons chaque régime. Supposons, au début, la mémoire 1 vide de toute information. Aussitôt que l'entrée de la mémoire est reliée au récepteur (niveau 1), cette dernière se remplit en Δt = 64 ms de 2048 bits de la modulation reçue sur l'antenne. Le débit binaire des signaux entrant dans la mémoire 1 est voisin de 32 kbits/s.

Pendant l'intervalle de temps $2\ \Delta t$ (= 128 ms) suivant (niveau 2), ces 2048 bits sont extraits de le mémoire pour être appliqués à l'écouteur, via le décodeur numérique. Le débit binaire des signaux sortant est de 16 kbits/s. Simultanément, l'entrée de la mémoire 1 étant reliée au codeur numérique, cette dernière se remplit de 2048 bits d'informations à émettre. Au bout des 128 ms, la mémoire 1 se sera vidée des informations reçues précédemment et sera remplie d'informations à émettre.

Pendant l'intervalle de temps Δt = 64 ms suivant, la sortie de la mémoire 1 (niveau 3) est reliée à l'émetteur. Cette dernière se vide en 64 ms des 2048 bits contenus. Le débit binaire des signaux sortant est de 32 kbits/s. La mémoire 1 est alors vide. On est revenu à l'état initial, et le cycle recommence.

La mémoire N° 2 joue un rôle complémentaire de la mémoire N° 1 puisque cette dernière est reliée à l'émetteur ou au récepteur quand la première est reliée au codeur et au décodeur, et vice-versa.

A chaque changement d'Emission vers Réception et Réception vers Emission, un petit intervalle de temps ε = 1 ms est réservé pour permettre la commutation de l'équipement d'Emission en Réception, et vice-versa, mais également pour permettre l'écoulement des signaux de supervision définis par la suite.

4.2 - Avantages et inconvénients de l'émission-réception en Duplex Temporel

Faire subir à la modulation un tel traitement présente de nombreux avantages, et quelques inconvénients.

Les avantages sont :

- une seule fréquence utilisée au lieu de 2,
- utilisation d'émetteurs récepteurs de type Simplex,
- suppression du filtre duplexeur,
- suppression des problèmes de saturation des récepteurs par les émissions locales à la station fixe,
- changement de fréquences simplifié,
- meilleure utilisation des bandes de fréquences disponibles, sans la servitude d'un écart fixe entre les deux groupes Emission et Réception,
- introduction facile, en cours de conversation, de messages de service tels que la demande de liaison prioritaire.

Les inconvénients sont :

- bandes passantes multipliées par 2 (mais occupation totale de spectre identique),
- synchronisation nécessaire,
- parole numérique,
- commutation E/R rapide.

Analysons ces différents points et voyons comment on peut rentabiliser les inconvénients cités ci-dessus.

On ne perd rien en spectre de fréquence utile puisque la bande totale occupée par une liaison en Duplex à 2 fréquences ou par une liaison en Duplex Temporel à 1 fréquence est la même. Par contre, la suppression du filtre duplexeur, l'utilisation d'équipements simplex, et surtout la suppression des problèmes de co-location des émetteurs et récepteurs fixes, sont des avantages importants. En effet, si de nombreuses liaisons simultanées sont établies en Duplex classique à 2 fréquences, l'ensemble des puissances émises à la Station Centrale peut provoquer la saturation des récepteurs de cette même Station Centrale. Au contraire, en Duplex Temporel, il suffit de synchroniser la commutation E/R et R/E de tous les équipements de la station fixe pour qu'ils soient tous en émission simultanément, ou tous en réception, de façon à supprimer le risque de saturation. De plus, l'utilisation de la même fréquence à l'émission et à la réception simplifie le synthétiseur de l'équipement qui n'a qu'un seul plan de fréquence à fournir.

On voit donc que la synchronisation présente un premier avantage : celui de supprimer les problèmes de saturation. On verra, par la suite, que l'on peut exploiter cette synchronisation nécessaire pour l'établissement d'appels prioritaires dans un réseau saturé.

De même, la numérisation de la parole exigée par le Duplex Temporel simplifie et rend plus performant le dispositif de discrétion protégeant les conversations.

Enfin, la commutation E/R rapide exige l'utilisation d'un commutateur d'antenne à diode PIN à la place du relais d'antenne que l'on trouve dans les équipements simplex. Si l'on remarque que dans la plupart des cas, le MTBF des équipements est voisin de celui du relais d'antenne, on comprend l'avantage apporté par le remplacement de ce dernier par un dispositif électronique très fiable.

4.3 - La priorité

L'écoulement d'appel prioritaire, en cas de saturation du réseau, est un autre avantage du système d'émission-réception en Duplex Temporel.

Dans un réseau multifréquences analogique de type conventionnel, lorsque tous les canaux du réseau sont occupés par des liaisons du type Mobile-Mobile, il est impossible d'interrompre une communication établie à moins de disposer d'un équipement d'émission à forte puissance fournissant sur l'antenne de chaque équipement mobile un champ H.F. supérieur au champ de chaque équipement en communication.

On préfère, dans ce cas là, réserver des voies aux abonnés prioritaires, mais ceci entraine une mauvaise utilisation des canaux. En effet, si 3 voies sur 5 sont réservées aux abonnés prioritaires, trois appels prioritaires pourront être écoulés en cas de saturation du réseau, mais pas un de plus.

Dans un réseau synchrone fonctionnant en Duplex Temporel, il suffit de prévoir une période très courte de silence total (située au même instant sur tous les équipements, en début de période d'émission, par exemple) pour écouler toutes sortes de messages prioritaires.

De cette façon, il est possible :

- d'interrompre brutalement une conversation en empêchant chaque correspondant d'émettre, et en utilisant le canal libéré pour écouler l'appel prioritaire,
- d'avertir deux abonnés en cours de conversation par un signal de sonnerie pour qu'ils raccrochent et libèrent le canal au plus vite : soit pour écouler un appel prioritaire ne les concernant pas, soit pour les informer que l'un d'eux est appelé par un correspondant prioritaire.

4.4 - Modulation et signalisation numérique

Pour tirer le meilleur parti du système d'émission-réception en temps partagé, il faut utiliser une modulation de type numérique. Elle peut, en effet, être chiffrée ou codée plus facilement et efficacement qu'une modulation analogique. Elle permet, en outre, d'introduire facilement les messages concernant la signalisation, également en numérique.

La bonne immunité au bruit est fonction (pour une bande passante donnée) du choix de l'indice de modulation et de la forme du signal modulant.

L'indice de modulation retenu, $m = 2/\pi$, permet d'avoir une répartition spectrale la plus uniforme avec une largeur de bande utilisée égale à la moitié du débit binaire des deux côtés de la porteuse.

L'étude des formes à donner au signal modulant (pente des transitions) nous a amenés à préférer un signal trapézoïdal filtré, aux signaux en $\sin^2$, $\cos^4$, ou gaussien tronqué, pour sa facilité de mise en oeuvre.

4.4.1 - Modulation numérique

La numérisation de la parole fait appel à des méthodes de codage ou des méthodes d'analyse. Si les méthodes d'analyse sont encore dans une phase expérimentale, les méthodes de codage sont utilisées en téléphonie classique, et apparaissent de plus en plus en radiotéléphonie.

Les procédés de codage les plus employés sont les codages par impulsions (M.I.C.),ou les codages différentiels (M.I.C. différentielle, modulation Δ). La modulation Δ est une modulation par impulsions codées, différentielle à 1 bit. Pour des quantités d'informations transmises inférieures à environ 20 kbits/s., la qualité de la parole est meilleure en modulation Δ qu'en modulation par impulsions codées. C'est pour cette raison, et pour sa simplicité de mise en oeuvre, que la modulation Δ a été préférée.

Le diagramme fonctionnel du codeur et du décodeur utilisés est donné par la figure N° 3. C'est un dispositif à pente variable, à loi de compression à 5 digits.

L'information numérique issue du comparateur traverse un registre à décalage, et l'information contenue dans ce registre est dosée pour modifier l'amplitude du signal de comparaison, et asservir ainsi la pente.

Ce type de codeur comporte un circuit logique dont l'action est de diminuer les effets d'une impulsion perturbatrice ou d'une surmodulation. L'intelligibilité de la parole est ainsi améliorée pour des débits binaires voisins de 16 kbits/s.

4.4.2 - Signalisation numérique d'appel radio (figure N° 4)

Dans un réseau utilisant une modulation analogique, l'appel sélectif des abonnés se fait par l'émission successive de tonalités basses fréquences cadencées suivant une loi connue.

Chaque fréquence B.F. doit être émise pendant plusieurs dizaines de millisecondes pour être correctement décodée. Si le signal d'appel comporte un grand nombre de chiffres, l'émission du signal est longue. Au contraire, dans un réseau utilisant une signalisation numérique d'appel radio, la durée de l'émission de l'ensemble des chiffres du numéro d'appel sera courte. Dans ce cas, il est souhaitable d'utiliser les moyens dont on dispose pour s'assurer que le numéro reçu ne présente pas d'erreur (loi majoritaire, code correcteur d'erreur, bit de parité) sans pour celà allonger considérablement la durée du signal d'appel radio.

Dans le réseau considéré, la signalisation d'appel radio est émise en 40 millisecondes, et comporte 10 chiffres transmis. Le fonctionnement en réseau comportant 8 canaux de trafic banalisés impose la répétition du message d'appel.

4.4.3 - Autres types de signalisation numérique

Le fonctionnement en réseau demande l'utilisation de signaux de supervision. Ce sont des signaux de service échangés sans que l'abonné les perçoive ; à savoir signal de synchronisation, signal de liaison prioritaire, signal de "décroché" et de "raccroché" du combiné. Ces signaux sont émis, si nécessaire, au début de chaque période d'émission.

L'exploitation de ces signaux (utilisés même en cours de conversation) permet la gestion contrôlée du réseau, ce qui ne serait pas possible avec une signalisation analogique lente.

4.5 - La commutation des canaux

Elle tire parti des avantages de la signalisation numérique dont l'utilisation facilite le fonctionnement du Duplex Temporel.

Dans les réseaux multifréquences à accès aléatoire sur plusieurs canaux mis en commun entre les abonnés, la liaison s'établit par émission du signal d'appel sur le premier canal reconnu libre.

Les équipements sont toujours en recherche d'appel en écoutant cycliquement sur chacun des canaux dans le but d'y décoder leur propre numéro.

La commutation des canaux doit se faire rapidement pour avoir un temps d'appel court. Ceci est permis par l'utilisation d'une signalisation numérique concentrant l'ensemble du message d'appel (à savoir : identité de l'appelant, numéro de la Station Centrale, numéro de l'appelé demandé, type de liaison, priorité..) en une seule période d'émission de 64 ms.

Ici, dans le but d'éliminer les réceptions parasites dues à l'intermodulation des émetteurs entre eux, un plan de fréquence sera choisi, ne conservant que les 8 canaux non contigus intermodulant peu entre eux (par ex. N° 1 - 3 - 8 - 14 - 18 - 30 - 37 - 39) parmi 58 plans possibles.

Pour celà, le synthétiseur est positionné au départ sur la première fréquence du plan de fréquence, et explore successivement, par un commutateur électronique de code, les fréquences du plan choisi.

Ce commutateur positionne le synthétiseur par des sauts de fréquences n'excédant pas 12 espacements de canaux, ce qui permet un accrochage en phase très rapide du synthétiseur de fréquence.

De plus, l'envoi du message complet en une seule période d'émission permet également de transmettre tout ordre de service nécessaire à la gestion du réseau.

4.6 - Conclusions

Les caractéristiques signalées ci-dessus montrent bien que c'est au niveau du système, pris dans son ensemble, qu'il faut considérer les avantages du Duplex Temporel.

5 - LES EQUIPEMENTS

Trois sortes d'équipements sont utilisés dans ce réseau : l'équipement E/R portable, appelé Equipement Terminal d'Usager LMT 3453-A ; les équipements d'E/R fixes qui sont placés dans une baie LMT 3485-A ; et l'organe d'interface placé également dans une baie LMT 3872-A.

5.1 - Equipement Terminal d'Usager LMT 3453-A (figure N° 5)

Son volume est de 9 litres, et le poids de 5 kg environ.

La face avant comporte l'ensemble des commandes, à savoir :

un clavier permettant de composer le numéro de l'abonné appelé, trois commutateurs permettant d'afficher les fréquences de fonctionnement, un commutateur "Arrêt" - "Marche avec fixe" - "Marche sans fixe". Elle comporte également un combiné téléphonique et une antenne.

Le coffret contient 5 circuits imprimés montés dans un coffret solidaire de la face avant, et l'ensemble batterie-alimentation.

Les circuits imprimés remplissent les fonctions suivantes :

- Emetteur-Récepteur
- Synthétiseur de fréquence
- Codeur-Décodeur
- Emetteur et Récepteur de signalisation numérique
- Base de temps, circuit d'émission-réception en temps partagé, et dispositif de discrétion.

L'ensemble batterie rechargeable-alimentation peut être séparé du coffret.

Les problèmes de poids et de consommation exigent la recherche de la meilleure technologie dans chaque sous-ensemble. C'est pourquoi les circuits logiques fonctionnant à faible vitesse sont en "C-MOS", certains circuits du synthétiseur de fréquence sont en "Low Power Schottky", et un petit nombre seulement fonctionnant en diviseur de 400 à 100 MHz sont en "ECL".

Le découpage des fonctions et le minimum d'éléments ajustables permettent d'assurer une maintenance aisée, compatible avec les besoins des utilisateurs.

Caractéristiques générales de l'équipement

Nombre d'abonnés par réseau	:	32 max.
Nombre de communications simultanées	:	8
Type de modulation	:	Δ 16 kbits
Type de transmission	:	Duplex Temporel sur une seule fréquence
Poids	:	5 kg environ
Autonomie	:	5 heures min.
Bande de fréquence	:	410 - 470 MHz
Espacement des canaux	:	50 kHz
Portée	:	6 km max.
Nombre d'usagers prioritaires	:	5
Gamme de température	:	-25 +55°C
Spécification	:	AIR 7303 II B (sauf pour la température)
Dispositif de discrétion	:	Numérique

Déport possible de l'antenne à 10 mètres du poste.

5.2 - Equipement d'émission-réception fixe 3485-A (figure N° 6)

Cet équipement, logé dans une baie du standard 19 pouces comportant 4 tiroirs de 6 unités, se compose :

- d'un dispositif de couplage d'aérien et de deux amplis répartiteurs,
- d'un ensemble de 8 équipements identiques aux équipements portables 3453-A sur le plan des circuits (ils sont disposés en coffret, mais l'alimentation ne se fait pas par batterie individuelle)
- d'un tiroir d'alimentation.

Cette baie, fixée à la cloison à l'aide d'amortisseurs, peut pivoter sur des charnières pour dégager la forme de câbles.

5.3 - Organe d'interface 3872-A

Cet équipement est logé dans une baie identique à la précédente, qui comporte :

- dans un tiroir 6 unités, les 16 équipements de ligne d'abonné,
- dans une 2ème tiroir 6 unités, 16 autres équipements de ligne d'abonné,
- dans une 3ème tiroir 6 unités, les 8 équipements de voie radio et les éléments assurant la connexion,
- dans le dernier tiroir, l'alimentation.

6 - AMELIORATIONS ENVISAGEES DANS LE FUTUR

Aujourd'hui, l'utilisation d'une modulation numérique en Duplex Temporel ou en Duplex bi-fréquence entraîne par canal une occupation d'une bande de fréquence plus importante que celle nécessitée par l'utilisation d'une modulation de fréquence ou de phase en Simplex.

Aussi, les améliorations des deux systèmes, et plus particulièrement celles du Duplex Temporel, sont fonction des progrès qui seront réalisés dans le domaine de la numérisation de la parole.

Les études menées dans tous les laboratoires, dans le domaine de l'analyse et la synthèse de la parole, laissent espérer, dans un avenir assez proche, l'apparition de vocoders adaptatifs fournissant une parole de qualité acceptable pour les débits binaires de l'ordre de 2,4 kbits/s. Dans ce cas, les espacements de canaux inférieurs à 12,5 kHz seront possibles pour l'écoulement de liaisons en Duplex Temporel. A condition toutefois qu'une microminiaturisation poussée permette d'envisager, pour ces codeurs, les faibles dimensions et le prix de revient modeste compatibles avec leur utilisation dans un poste portable.

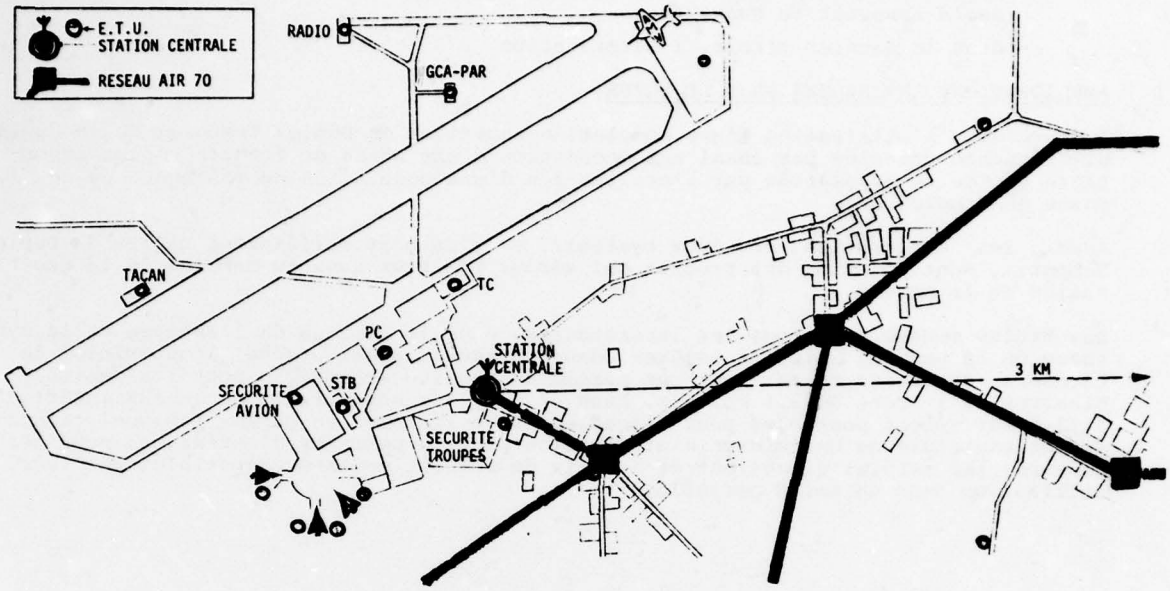


Fig.1 Plan du reseau

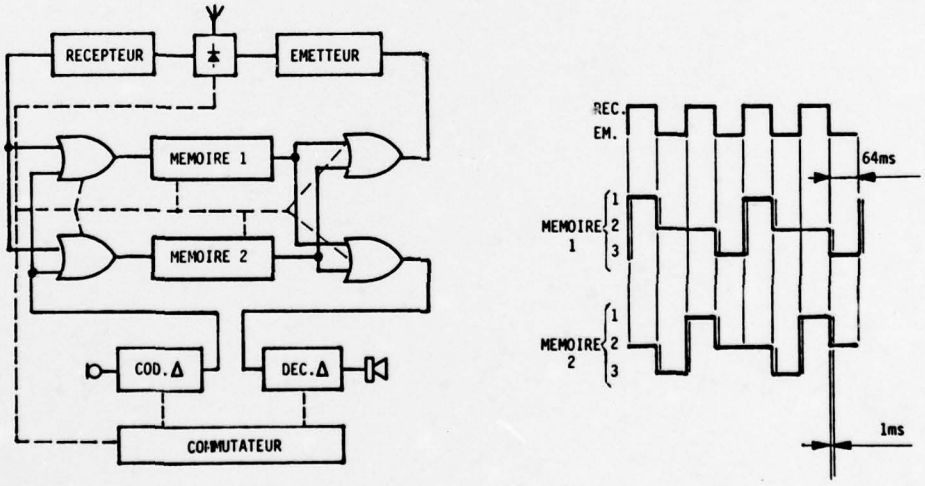


Fig.2 Principe de fonctionnement du duplex temporel

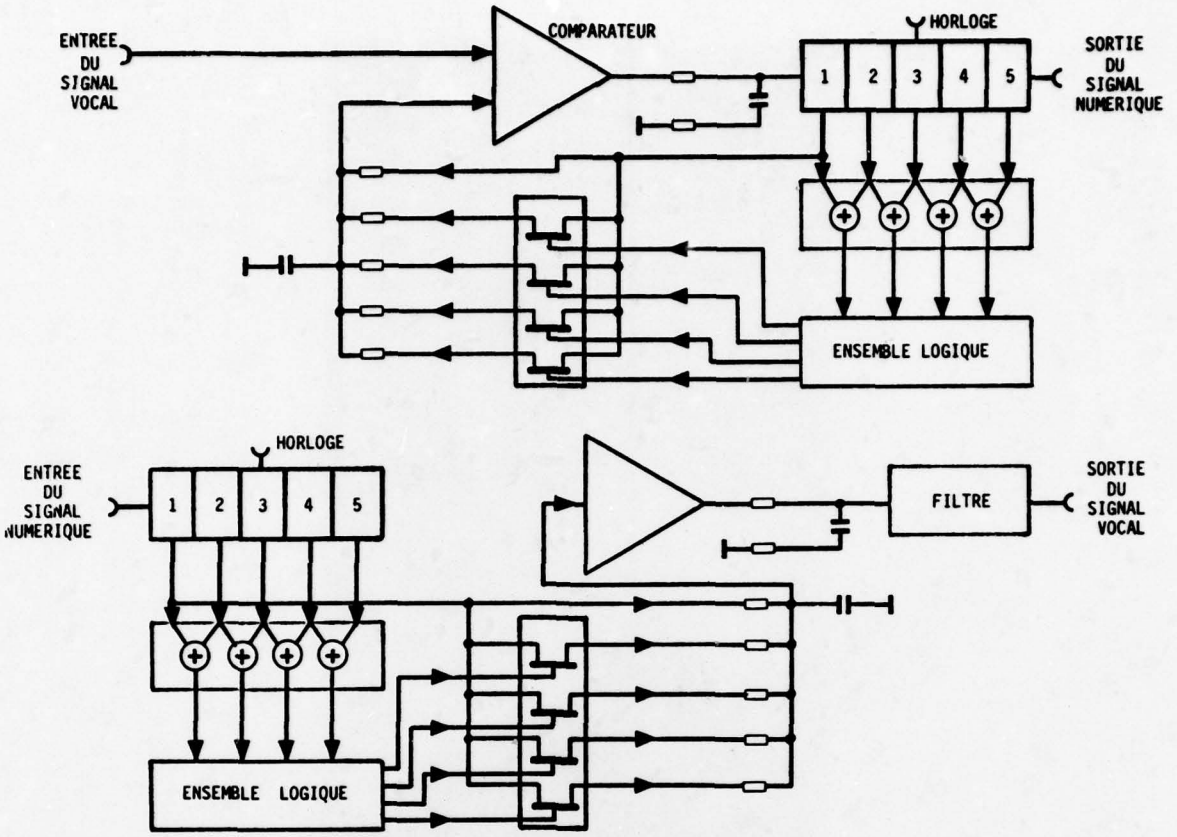


Fig.3 Codeur decodeur delta

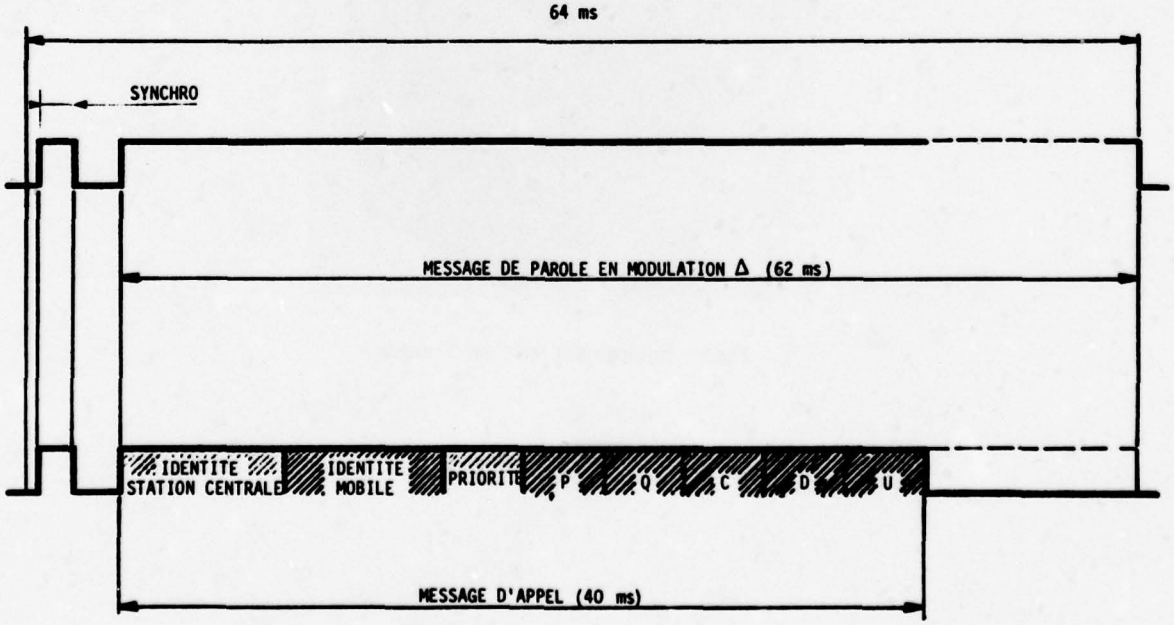


Fig.4 Exemples de messages transmis

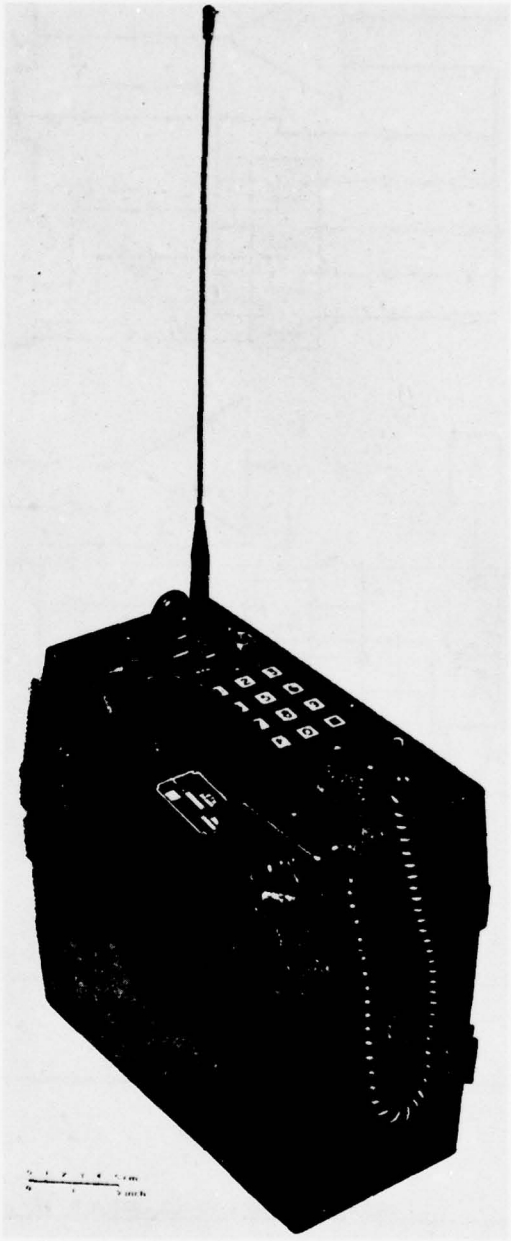


Fig.5 Equipement terminal d'usager

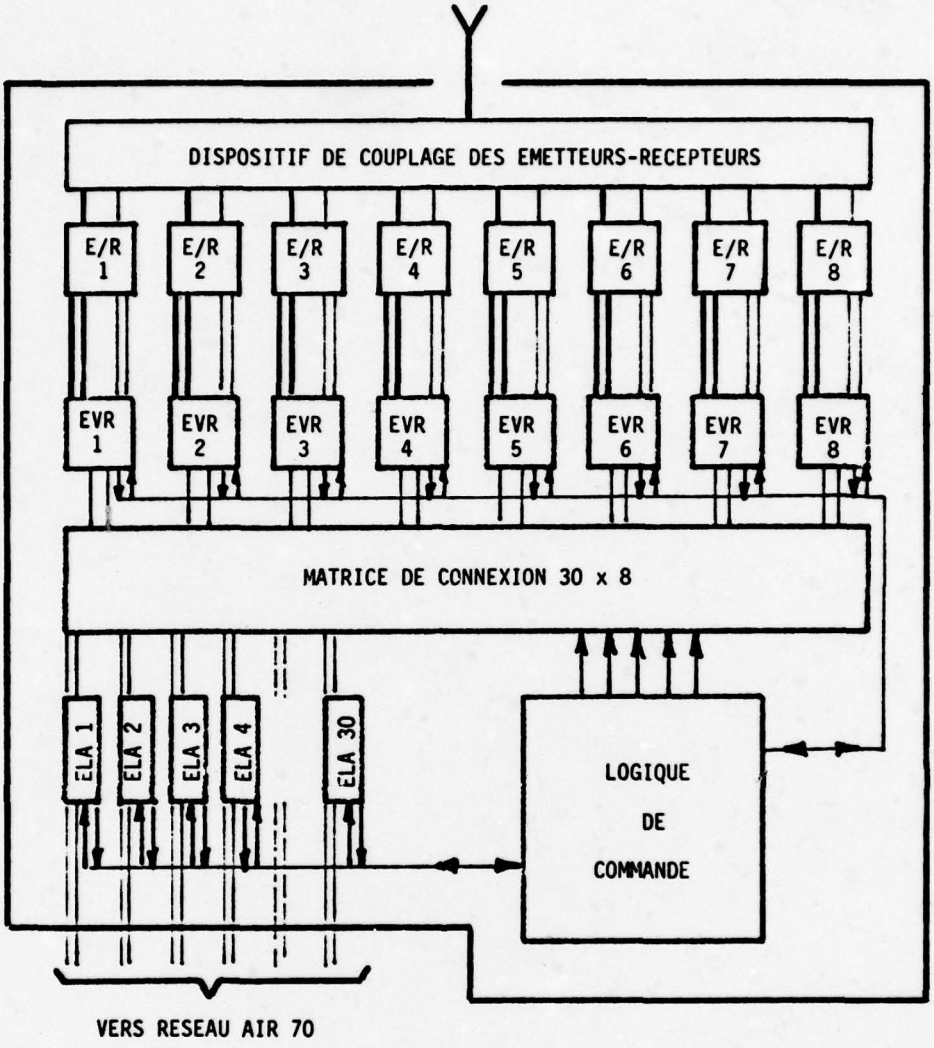


Fig.6 Station centrale

A MULTI-GBIT/S RZ-FORMAT DIODE MULTIPLEXER

Udo Barabas

Institut für Elektronik, Ruhr-Universität Bochum,
Postfach 10 21 48, D-4630 Bochum, F. R. Germany

ABSTRACT:

A clocked multiplexer circuit was realized which provided 4.48 Gbit/s, 5 Gbit/s, and 7.84 Gbit/s output-pulse streams for PCM-type input tributaries at 1.12 Gbit/s, 0.25 Gbit/s, and 1.12 Gbit/s, respectively. The circuit employed essentially modified ultra-broadband hybrid tees, step-recovery diodes, and GaAs Schottky-barrier diodes. Output voltages up to 2 V were obtained across a load of 50 Ω. The pulse width of the output pulses was approximately 100 ps.

1. INTRODUCTION:

In anticipated high-resolution radar and in broadband PCM communication systems, information streams with extremely high bit rates will be processed. Essential circuits in such systems are multiplexers which combine several data streams at a low bit rate to a higher bit-rate output bit stream. In this contribution, a multiplexer for RZ-format output signals operating up to 7.84 Gbit/s is reported. The multiplexer basically relies on step-recovery diodes, modified ultra-broadband hybrid tees (ANZAC INC.), and GaAs Schottky-barrier diodes.

It was principally shown (BARABAS, U., et al, 1976), (BARABAS, U., et al, 1977), (BARABAS, U., 1977) that clocked circuits employing both modified hybrid tees and step-recovery diodes (SRDs) can be used for both the regeneration of baseband pulses at ultra-high bit rates and for multiplexer operation. As regenerators they simultaneously provide pulse shaping (shortening), amplification, and retiming, whereas as multiplexers (BARABAS, U., et al, 1978) they combine several low bit-rate input signals into one common high bit-rate output. The output pulse width of the multiplexer described here is about 100 ps and is nearly independent of the shape of the input pulses. If the width of the input pulses is much greater than that of the output pulses, essential reduction of the pulse width is reached. Previously (RUSSER, P., et al, 1975), (BARABAS, U., et al, 1976) multiplexers were described which contained step-recovery diodes but no modified hybrid tees, obtaining bit rates of RZ-format output pulses up to 2 Gbit/s.

Here, two multiplexers will be investigated, the first operated from 0.25 Gbit/s (input) to 5 Gbit/s (output), the other ranged from 1.12 Gbit/s (input) to 4.48 Gbit/s (output). The latter was modified in a feasibility study and then operated from 1.12 Gbit/s (input) to 7.84 Gbit/s (output). In those two versions output voltages of about 2 V at a load of 50 Ω were obtained. The maximum of the output pulse amplitudes experimentally reached was 3.2 V. In all multiplexers the theoretical maximum output pulse amplitude equals half of the breakdown voltage of about 20 V of the employed step-recovery diodes.

2. DESCRIPTION OF THE PRINCIPLE BEHAVIOUR OF THE MULTIPLEXERS:

Fig. 1 is the schematic representation of both multiplexer circuits. They consist of clocked pulse shortening stages and an adding circuit. Four input bit streams produced by bit pattern generators were multiplexed into one common output channel feeding the load. Shortened by 4 pulse shortening stages the pulse streams were properly delayed (by the lines τ_0 to τ_3) and finally combined in the adding circuit.

In the 250 Mbit/s (input) to 5 Gbit/s (output) multiplexer, the potential number of tributaries which could be combined was 20 since this was the amount of the available time slots. Because of their number being limited to only 4, the channels were arbitrarily distributed within the output time interval. The complete output word contained 320 bits. The four input bit streams were made up of pseudo-random-like words of 16 bits generated by MECL-III-ICs (SCHWEIZER, L., 1970). The output pulse width of this bit pattern generator was approximately 2.25 ns, thus a reduction of the pulse width of the factor of about 20.5 was obtained. This multiplexer was carried out in order to demonstrate its capability to transform pulses generated by commercial ICs into a higher Gbit/s range.

In future PCM communication systems a bit rate of approximately 1.12 Gbit/s might be of some interest. This bit rate was chosen for the input bit streams of the other multiplexer. Since the next higher bit rate (hierarchy) of PCM systems is usually generated by multiplexing of four input channels, the output bit rate of the multiplexer results in about 4.48 Gbit/s. In a feasibility study, the highest possible output bit rate was determined for an input bit rate of 1.12 Gbit/s, yielding an output at a bit rate of 7.84 Gbit/s. In this case, four out of a potential maximum of seven bit streams were multiplexed. The input pulses of the multiplexer were produced by a bit pattern generator containing bipolar transistors. The pulse shortening stages were the same as in the 250 Mbit/s (input) to 5 Gbit/s (output) multiplexer. Only the adding circuit was changed by using GaAs Schottky-barrier diodes. In the 250 Mbit/s (input) to 5 Gbit/s (output) multiplexer a resistive network matching the four input lines of the adding circuit was employed.

As shown in Fig. 1, a modified hybrid tee is in the center of a pulse shortening stage. Fig. 2 shows the principle circuit of a pulse shortening stage. The energy for such a stage is supplied by a sinusoidal pump (clock). In the 1.12 Gbit/s (input) multiplexer the pulse shortening stages were operated in a push-pull mode by the input signal sources $V_s(t)$ and $-V_s(t)$, whereas in the 250 Mbit/s (input) multiplexer these stages were fed only by one input signal source $V_s(t)$. The lines for the input signals were shunted with the step-recovery diodes SRD_1 and SRD_2. The output signal was $V_a(t)$.

Together with the step-recovery diodes, the hybrid tee decouples the output line from the pump line as long as no input signals $V_s(t)$ and $-V_s(t)$ occur. In this case the two reference lines containing the step-recovery diodes are ballanced. If input signals arrive at the step-recovery diodes, the latter are unequally charged by the signals, thus unballancing the reference lines. The pump is then connected with the output line and an output signal $V_a(t)$ occurs.

At first the modified version of a hybrid tee, an essential part of a pulse shortening stage, will be reported briefly (BARABAS, U., to be publ.). A hybrid junction can be defined as a "waveguide arrangement (including coaxial transmission lines) with four branches which, when branches are properly terminated, has the property that energy can be transferred from any one branch into only two of the remaining three" (IRE STANDARDS ON ANTENNAS AND WAVEGUIDES: DEFINITIONS FOR WAVEGUIDE COMPONENTS, 1955). In common usage of such junctions, this energy is equally divided between the two branches.

The hybrid tee described here consists of both a lumped branching point, constituting a bridge circuit, and two (transmission line) broadband transformers. The resistances of the bridge circuit are formed by 6 lines (4 striplines and 2 coaxial lines) and their load resistors. All lines have the same characteristic impedance. The branching point is presented in Fig. 3. The dielectrics between the lines (3), (4), (5), and (6) are not shown. The two coaxial lines (1) and (2), going diagonally through the branching point are arranged side by side, their inner conductors having no contact with one another.

The circuit of the branching point is shown in Fig. 4. The substrate between the striplines consists of glass-fibre reinforced teflon with a thickness of 1 mm. The coaxial lines are micro-semirigid cables (UT 47-Sp) with an outer diameter of 1.22 mm. Two lines each of the 6 lines (lines (1) and (2), (3) and (4), (5) and (6), respectively) are combined to three pairs of lines. Two of the three pairs are joined and then connected to the broadband transformers (Fig. 5).

The resulting circuit is a hybrid tee employing no transmission lines with defined lengths related to a particular transmitted wavelength. Therefore, this circuit is finely suited for broadband (pulse) applications. The operation frequency range is limited by the broadband transformers (RUTHROFF, C. L., 1959), (HILBERG, W., 1965), (WELLENS, U., 1977). The upper cut-off frequency of the hybrid tee was measured at 5 GHz between input (a) or (b) and line (3) or (4) and the lower cut-off frequency at 2.4 MHz. The insertion loss was approximately 0.6 dB (Fig. 6).

In a pulse shortening stage the hybrid tee is operated only in one direction. Therefore, the transformer in the output line can be omitted, thus causing two output lines. This operation has the advantage that the limitation in frequency bandwidth, which is caused by the transformer in the output line, is removed both between the lines (3) and (4) (shunted by the step-recovery diodes) and the output. The branching point itself shows to have an operating range of about 12 GHz between the lines (3), (4), (5), and (6), and approximately 7 GHz from the lines (1) and (2) to the lines (3), (4), (5), and (6). At the upper cut-off frequency of about 12 GHz the dimensions of the branching point (approximately 1 mm in the direction of the electric fields) are not negligible, thus the branching point can no longer be regarded as lumped. The wavelength at 12 GHz in a teflon substrate is 17.25 mm. The other upper cut-off frequency of 7 GHz is caused by the non-negligible dimensions of the branching point and also by the self-inductances of the inner conductors of the lines (1) and (2) inside the branching point (see Fig. 3) showing total (equivalent) inductivities of about 2 nH.

Using of the transformer in the lines (1) and (2), the conductors of these lines are connected to one another thus causing short circuits. This is the reason for another lower cut-off frequency of the hybrid tee. However, this frequency is not significant because by providing the lines (1) and (2) with ferrite beads it can be made much lower than the lower cut-off frequency of the transformers.

The behaviour of the modified hybrid tee employed in these multiplexers can be explained with the scattering matrix of a hybrid tee.

$$\begin{bmatrix} b_3 \\ b_4 \\ b_a \\ b_b \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_3 \\ a_4 \\ a_a \\ a_b \end{bmatrix}$$

In this equation system a_a is the squareroot of the pump power. The pump power gets, equally divided, only into the lines (3) and (4) being shunted with the step-recovery diodes SRD_1 and SRD_2. Therefore, the reflexion coefficients occuring in these lines are

determined both by the impedances of the step-recovery diodes and the internal resistances of the signal sources $V_s(t)$ and $-V_s(t)$. When these reflexion coefficients are equal to one another, the reflected powers a_3^2 and a_4^2 are also equal to one another. For equal distances between the branching point and the step-recovery diodes is $b_b = 0$ and

$$b_a = \frac{1}{\sqrt{2}} (a_3 + a_4).$$

The whole reflected power returns to the pump source.

When the reflexion coefficients in the lines (3) and (4) are unequal to one another there is $a_3 \neq a_4$. Therefore, an output signal

$$b_b = \frac{1}{\sqrt{2}} (a_3 - a_4)$$

exists. Since the hybrid tee is used here in its modified version, (with no transformer in the output line (b)), there exist two output lines (5) and (6). Correspondingly there are two output signals

$$b_5 = b_6 = (a_3 - a_4).$$

The principle behaviour of the pulse shortening stage is demonstrated in Fig. 7. Fig. 7a shows the circuit of the stage with the branching point of the hybrid tee in the center of the circuit. The pump, equally divided by the transformer in the line (a), feeds the lines (1) and (2). To maintain the clearance of Fig. 7a this transformer is not shown. The step-recovery diodes SRD_1 and SRD_2, shunting the lines (3) and (4), are biased with the d.c. voltages V_1 and V_2.

The internal behaviour of the pulse shortening stage can be explained by Fig. 7b. The sinusoidal pump voltage, shown at the upper trace, is measured in the lines (3) and (4), respectively. These voltages are measurable if both step-recovery diodes are permanently in the cut-off state, which is caused by negative bias voltages V_1 and V_2. With both diodes properly biased, V_3 and V_4 are the voltages in the lines (3) and (4), respectively. A negative pump halfwave in the lines (3) and (4) charges the diodes and a positive one discharges them. If the IV - and especially the CV - characteristic curves of both diodes differ a little from one another both diodes are not identically charged and discharged. However, an unequal charging and discharging of the step-recovery diodes can be compensated by a proper choice of the bias voltages V_1 and V_2. Then both diodes go into the cut-off states at the same time. During the whole period of the pump wave the voltages in the lines (3) and (4) are nearly identical with each other. Therefore, no output signal is in the lines (5) and (6).

When the input signals $V_s(t)$ and $-V_s(t)$ occur during the conducting time of the step-recovery diodes, the diodes are differently charged. A conducting time lasts for the whole charging and discharging period. During this time the internal impedances of the step-recovery diodes are very low. Therefore, nearly the whole input signal currents are led through the diodes.

When SRD_2 is additionally charged by $-V_s(t)$ and SRD_1 necessarily is discharged by $+V_s(t)$, SRD_2 goes a little later into the cut-off state than SRD_1 (see Fig. 7b). Only in the short time of different diode impedances the voltages in the lines (3) and (4) are unequal. The difference of both voltages gets, equally divided, into the output lines (5) and (6). One of these output voltages is shown at the lower trace of Fig. 7b. The leading edge of the output pulse is generated by SRD_1 and the trailing edge is shaped by SRD_2. The risetimes of both step-recovery diodes are about 50 ps, the halfwidth of the output pulses is approximately 100 ps. The short risetimes of the step-recovery diodes are the reason for the removal of the transformer in the output lines; its upper cut-off frequency of 5 GHz would otherwise cause an essential flattening and broadening of the output pulses. With an upper cut-off frequency of the branching point of about 12 GHz from the lines (3) or (4) to (5) or (6), the short output pulses are transferred nearly undisturbed.

By increasing the pump amplitude, the input signal charge, stored in the step-recovery diodes, is removed faster. Therefore, the switching edges of the step-recovery diodes approach each other and the output pulses get shorter and higher. The shortest measured halfwidth of output pulses was 75 ps.

Some basic properties of a pulse shortening stage can now be determined:

1: The pulse shortening stage is a clocked circuit which regenerates the amplitude, shape, and timing of the input pulses. The retiming is determined both by the timing of the pump voltage related to the input signals and the d.c. voltages V_1 and V_2. The output pulses are produced sooner when V_1 and V_2 become smaller.

2: The pulse shortening stage is charge-controlled. The charge of both output pulses is, caused by losses, always smaller than the whole signal charge stored in both step-recovery diodes.

3: With a defined charge of an output pulse the pulse gets shorter with increasing amplitude. If the input and output pulses possess equal shapes but unequal durations and

amplitudes, the insertion charge efficiency is for a push-pull operation

$$v_q = \frac{V_o}{V_i} \frac{T_o}{T_i}$$

with V_o and V_i being the amplitudes and T_o and T_i being the halfwidths of the output and input pulses, respectively. Furthermore, it is presumed that the whole input pulses are stored in the step-recovery diodes. If the pulse shortening stage is not operated in a push-pull mode, the insertion charge efficiency is 2 v_q. The insertion power gain is for a push-pull operation

$$v_p = v_q^2 \frac{T_i}{T_o} .$$

If the pulse shortening stage is operated with only one input signal source, the insertion power gain is also 2 v_p. It can be seen, that v_p increases with T_i/T_o at constant insertion charge efficiency v_q.

4: Input and output signals are decoupled in time. By the integration of the input signal charge disturbances on an input signal, e.g. noise with higher frequencies than the pump frequency, have no influence on the output pulse, if their average value is zero for the conducting time of the step-recovery diodes. A jitter of the input signals is also insignificant, as far as the complete input signals are stored in the step-recovery diodes during the conducting time (about 650 ps at a bit rate of 1.12 Gbit/s). Fig. 8 shows the amplitudes of the output signals at a bit rate of 1 Gbit/s of a pulse shortening stage when the input signals arrive at different times. The duration of the input pulses (50 % to 50 %) is 375 ps. With this input signal no essential influence on its output pulse is determinable when the input signal is shifted within about ΔT = -100 ps and ΔT = +180 ps related to the optimum phase displacement between the input signal and the pump voltage. If the input signal is shifted within this time slot, only one output signal is generated. Its amplitude varies from 98 % (ΔT = -100 ps) via the maximum amplitude of 100 % (ΔT = 0 ps) to 95 % (ΔT = +180 ps).

5: The minimum of needed pump power is obtained under two conditions: The first being that the switching edges of the step-recovery diodes are shifted by V_1 and V_2 into the maximum of the discharging pump halfwave. The second condition is fulfilled if one diode begins its switching process when the other diode has just ended it. Then the amplitude of the pump in the lines (3) and (4) is approximately as high as the sum of both output pulses in the lines (5) and (6). The pump amplitude V_a in line (a) is

$$V_a = \sqrt{2}\ (V_{RL_1} + V_{RL_2}).$$

6: The frequency behaviour of the pulse shortening stage must be separated in both the frequency behaviour related to the signal pulses and to the frequency of the pump.

a: The lower frequency limit, related to the input signals, determines the maximum permissible successive number of logic zeros and is caused by both the capacitive coupling of the input signal sources to the step-recovery diodes and the short circuits produced by the inner and outer conductors of the lines (1) and (2). The short circuits shunt the output lines (5) and (6).

b: The lower limit of the frequency of the pump is determined by the lower cut-off frequency of the broadband transformer in line (a), here 2.4 MHz. It also depends on the signal charge losses caused by both the lifetime of the step-recovery diodes and a discharging of their diffusion capacitances by the outer network. With the employed diodes (HP, 5082-0008) a pulse shortening stage can be well operated at pump frequencies above 100 MHz. For pump frequencies below 100 MHz larger step-recovery diodes should be used in order to obtain longer lifetimes. A disadvantage of larger diodes, however, is their slower switching process, causing longer output pulses. Table 1 gives a general survey over some step-recovery diodes.

c: Employed as multiplexer the upper limit of the pump frequency is determined by the ratio of the width of the discharging pump halfwave to the switching time of the step-recovery diodes. If the maximum of the pump voltage is to be made available for the output pulses, the discharging halfwave of the pump has to be at least twice as long as the risetimes of the step-recovery diodes, here approximately 50 ps. Therewith the maximum of the frequency of the pump is limited to approximately 4 GHz. Furthermore, the conductivity modulation of the step-recovery diodes, which causes a reduced stored signal charge and thus a decreased insertion charge efficiency, is not negligible at these high pump frequencies. With increasing pump frequency the input pulse width T_i decreases while the output pulse width T_o remains constant. Therefore, the ratio of T_i/T_o decreases and both the voltage gain and the power gain decrease, too.

3. THE 250 MBIT/S (INPUT) TO 5 GBIT/S (OUTPUT) MULTIPLEXER:

As shown in Fig. 1, the multiplexer is fed by the four output channels of the bit pattern generator. The bit pattern generator produces pseudo-random like words with a length of 16 bits and essentially consists of a four-stage shift register. The last two Flip-Flop outputs of the shift register are combined with an EXCLUSIVE-OR gate and then fed back to

the input of the first Flip-Flop (MOHRMANN, K. H., 1974). The signals of the four inverted Flip-Flop outputs have an NRZ-structure. These signals are gated in NOR-gates by the clock; thus they are transformed to RZ-format output signals. The bit sequence of one output channel of the bit pattern generator is shown in Fig. 9a. The pulsewidth is about 2.25 ns and the amplitude is attenuated to 80 mV.

The bit pattern generator as well as the pulse shortening stages are operated by the same clock (pump) which has a frequency of 250 MHz. Therefore, the charging and discharging times of the diodes are fixed with regard to the output pulses of the bit pattern generator. The amplitude of the pump in the lines (1) and (2) is 1 V (see Fig. 1 and 7), thus the output pulse amplitude cannot exceed 0.5 V. As it can be seen in Fig. 9b, the output pulse amplitude of one pulse shortening stage is 380 mV. Therefore, the pump voltage has an amplitude, which is higher than its optimum. This results in shorter and higher output pulses than those obtained by a pulse shortening stage being in its optimum operating point. Two of these output pulses are shown in Fig. 9c. They have a pulse halfwidth of 85 ps, a risetime of 67.5 ps, and a falltime of 45 ps.

The trailing edges of the output pulses of a pulse shortening stage are shorter than the leading edges. This is because both the series inductances (bondwires) and the junction capacitances of the step-recovery diodes cause an overshoot at the ends of the switching-off processes (see Fig. 7b). Therefore, V_3 begins already to fall while V_4 is still rising. The difference between both voltages gets faster to zero than without overshoot of V_3. During the overshoot of the voltage at SRD_2, the difference $V_3 - V_4$ changes its polarity (Fig. 7b) causing an overshoot at the end of the output pulses of the pulse shortening stage (see Fig. 9c). These overshoots favour the time domain multiplexing at bit rates of several Gbit/s.

The ripple on the base line after the output pulses is caused by stimulated oscillations between the series inductances and the junction capacitances of the step-recovery diodes, when the latter are in the cut-off state. These oscillations have no identical frequencies and no identical amplitudes. Therefore, they cannot be compensated in the branching point of the hybrid tee, thus reaching the output lines (5) and (6). Because of the bias voltages V_1 and V_2 the switching processes of the diodes are shifted into the maximum of the discharging pump halfwaves. So the cut-off states of the step-recovery diodes last approximately a quarter of the period of the pump voltages, in this case about 1 ns. Fig. 9c indicates that the ripple on the baseline continues for this time. During the conducting state of the diodes their junction capacitances do not exist and the oscillations vanish.

The output pulses of the pulse shortening stages, which have an amplitude of 380 mV, cannot perfectly switch GaAs Schottky-barrier diodes in the adding circuit. Therefore, a resistive network, which matches its four input lines, is used in the adding circuit. The matching network has an important attenuation of 8.9 dB and increases the halfwidth of the multiplexer output pulses to about 110 ps. In Fig. 10 these output pulses are shown with an amplitude of approximately 110 mV at a bit rate of 5 Gbit/s. By increasing both the pump power and the amplitude of the input signals the amplitude of the output pulses of the multiplexer increases, too.

4. THE 1.12 GBIT/S (INPUT) MULTIPLEXER:

In the 1.12 Gbit/s (input) to 4.48 Gbit/s (output) and 7.84 Gbit/s (output) multiplexer the output pulse amplitude was essentially increased. Therefore, both the power of the pump and the amplitude of the input signals of the multiplexer were enlarged. Furthermore, the multiplexer was operated in a push-pull mode, thus doubling the input signal charge. By using Schottky-barrier diodes the attenuation of the adding circuit was essentially reduced.

The bit pattern generator employed bipolar transistors in miniature stripline packages. One output bit stream of the bit pattern generator is shown in Fig. 11a. The signals have a pulse halfwidth of about 475 ps.

The insertion signal charge efficiency is determined by 4 terms. All numerical values are related to a bit rate of 1.12 Gbit/s.

1: A charge efficiency of the input signal charge related to the signal charge, which was computed to 1.89 with the signal charge delivered by the signal source to a load $R_L = R_i$. R_i, is the internal impedance of the signal source. This charge gain occurs during the storage phase of the input signal and is caused by the internal diode impedances which are very low compared to the internal resistances of the signal sources.

2: There is a charge efficiency of about 0.917 of the stored signal charge related to the input signal charge of the pulse shortening stage. This charge efficiency implies two charge losses, firstly a part of the input signal charge flows into the branching point and not through the step-recovery diodes, and secondly a discharging of the diffusion capacitances occurs through the outer network (6.7 % of the input signal charge). A further dissipation is caused by recombination in the diodes (1.7 % of the charge, flowing through the diodes). These charge losses occur also during the storage time of the diodes.

3: Furthermore, there is a charge efficiency of 0.545 of the output signal charges related to the stored signal charge in the step-recovery diodes. These charge losses ap-

pear during that time interval, in which the stored charge is removed and they are caused by the internal resistances of the signal sources shunting the step-recovery diodes together with the branching point.

4: A comparison with the measured output pulses yields an additional loss of approximately 8.8 %. The difference between the computed and the measured results can be explained because in the computation both the branching point and the output lines of a pulse shortening stage are assumed to be ideal and lossless.

With respect to these 4 terms the pulse shortening stage has in total an insertion charge efficiency of about 0.86.

Fig. 11b shows the output signals of the multiplexer. The upper trace presents the output pulses at a bit rate of 4.48 Gbit/s with an amplitude of about 2 V and a halfwidth of approximately 100 ps. Fig. 11c presents the output pulses of the multiplexer at a bit rate of 7.84 Gbit/s with an amplitude of about 2 V. This bit rate was generated by shifting the output pulses of the pulse shortening stages as near as possible to one another. The minimum pulse distance was adjusted by the delay lines τ_0 to τ_3, so that an unobjectionable RZ-structure could just be obtained. Furthermore, the output bit rate should be a whole number higher than the input bit rate.

5. ACKNOWLEDGEMENTS:

The author wishes to thank Prof. B. G. Bosch for guidance and encouragement and Mr. J. Kirchhoff for helpful assistance in both the implementation of the branching points of the hybrid tees and the computation regarding the pulse shortening stages. Financial support of the Deutsche Forschungsgemeinschaft (DFG) is gratefully acknowledged. The author would also like to thank the Hewlett-Packard Co., Palo Alto, for providing the step-recovery diodes employed.

REFERENCES:

ANZAC INC., UHF 180° hybrid (model H9).

BARABAS, U., WELLENS, U., LANGMANN, U., BOSCH, B. G., 1976, "Diode Circuits for Pulse Regeneration and Multiplexing at Ultrahigh Bit Rates", IEEE Trans. MTT-24, pp. 929 - 935.

BARABAS, U., LANGMANN, U., 1977, "Improved Versatile Gigahertz Pulse Generator", Electron. Letters, Vol. 13, pp. 28 - 29.

BARABAS, U., 1977, "A Differential Pulse Regenerator Driven by 1 Gbit/s PCM-type Signals from Avalanche Photodiodes", Electron. Letters, Vol. 13, pp. 536 - 537.

BARABAS, U., LANGMANN, U., BOSCH, B. G., 1978, "Diode Multiplexer in the Multi-Gbit/s Range", Electron. Letters, Vol. 14, pp. 62 - 64.

BARABAS, U., "On an Ultra-Broadband Hybrid Tee", to be publ.

HILBERG, W., 1965, "Die Eignung des Leitungsübertragers für die Impulstechnik", Nachrichtentechn. Zeitschrift, No. 4, pp. 219 - 230.

IRE STANDARDS ON ANTENNAS AND WAVEGUIDES: DEFINITIONS FOR WAVEGUIDE COMPONENTS, 1955, Proc. IRE, Vol. 43, pp. 1073 - 1074.

MOHRMANN, K. H., 1974, "Erzeugung von binären Quasi-Zufallsfolgen hoher Taktfrequenz durch Multiplexen", Siemens Forsch.- und Entwickl.-Ber., Vol. 3, No. 4, pp. 218 - 224.

RUSSER, P., GRUBER, J., 1975, "Hybrid integrierter Multiplexer mit Speicherschaltdioden für den Gbit/s-Bereich", Wiss. Ber. AEG-Telefunken 48, pp. 54 - 59.

RUTHROFF, C. L., 1959, "Some Broad-Band Transformers", Proc. IRE, Vol. 47, pp. 1337-1342.

SCHWEIZER, L., 1970, "Eigenschaften und Anwendungen von binären Quasi-Zufalls-folgen", Frequenz, Vol. 24, pp. 230 - 234.

WELLENS, U., 1977, "Design of HF-UHF Broadband Stripline Transformers", AEÜ, Vol. 31, pp. 130 - 132.

Table 1: List of some step-recovery diodes with different carrier lifetimes and transition times.

Co.	Diode	Junct. Capacitance at V_R = 10 V C_{jR} (pF)	Minimum Breakdown Voltage V_{BR} (V)	Carrier Lifetime T_L (ns) (I_F = 10 mA)	Transition Time T_T (ps) (20 % - 80 %)
NEC	ND 1111	0.1 - 0.24	< 10		
HP	5082-0008	0.25 - 0.5	20	15	50
Aertech	A4S 008	0.25 - 0.5	20	15	50
HP	5082-0020	0.35 - 1	25	40	75
Aertech	A4S 015	1 - 1.5	35	75	125
Aertech	A4S 021	1.6 - 3.2	40	85	150
Aertech	A4S 017	3.2 - 5.2	75	200	400

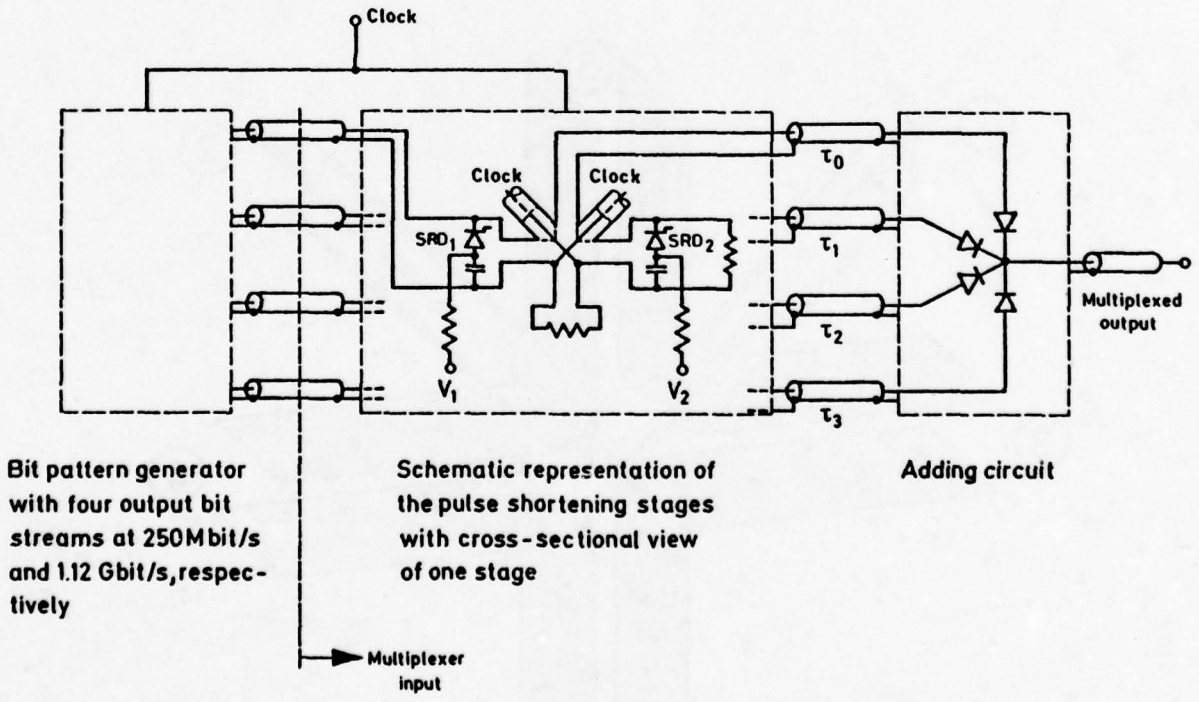


Fig. 1: Schematic diagramm of the multiplexer circuit.

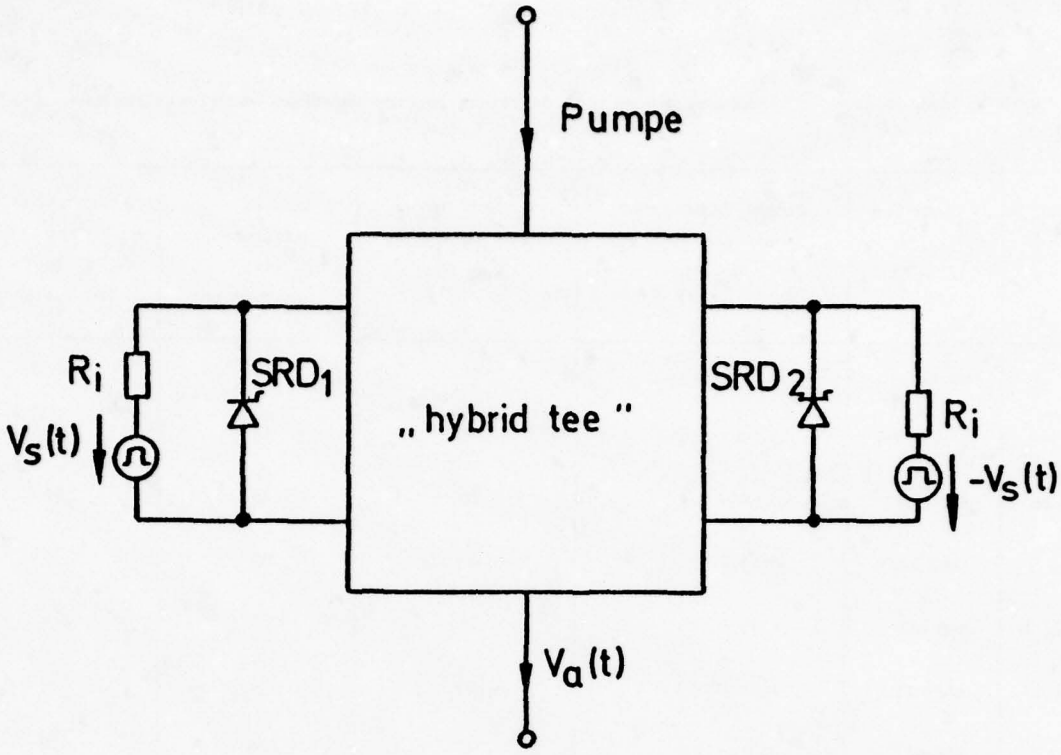


Fig. 2: Principle circuit of a pulse shortening stage.

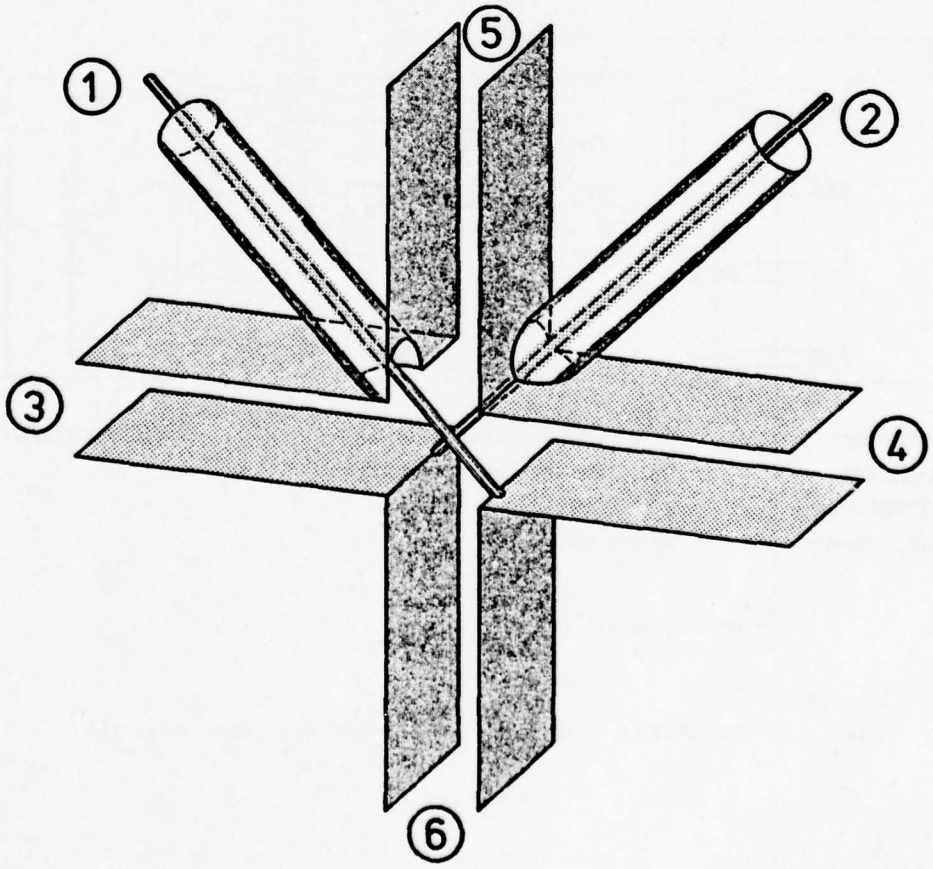


Fig. 3: Illustration of the branching point of the hybrid tee.

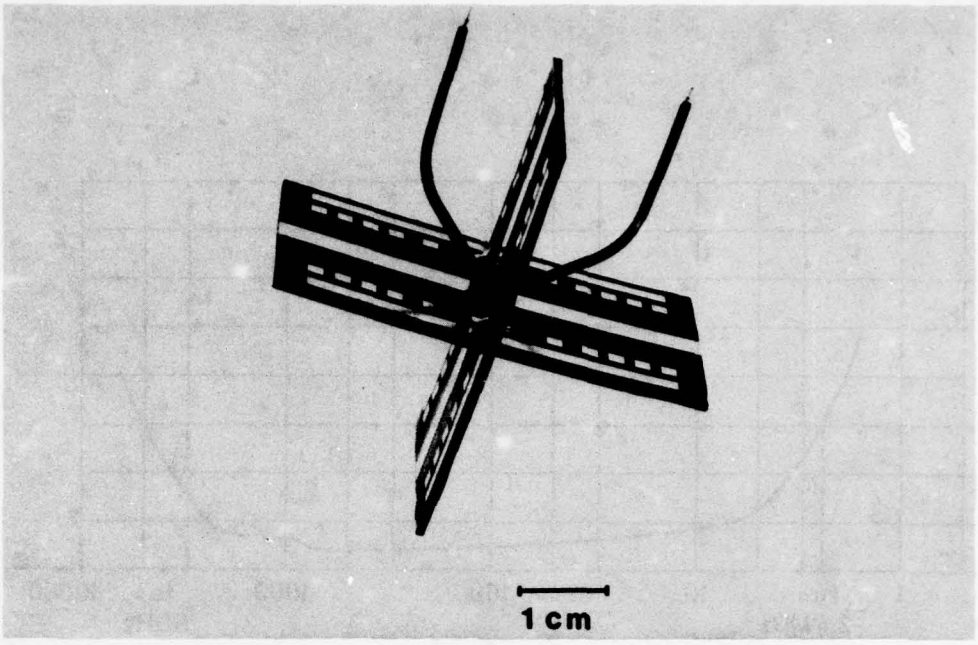


Fig. 4: Photo of the branching point of the hybrid tee.

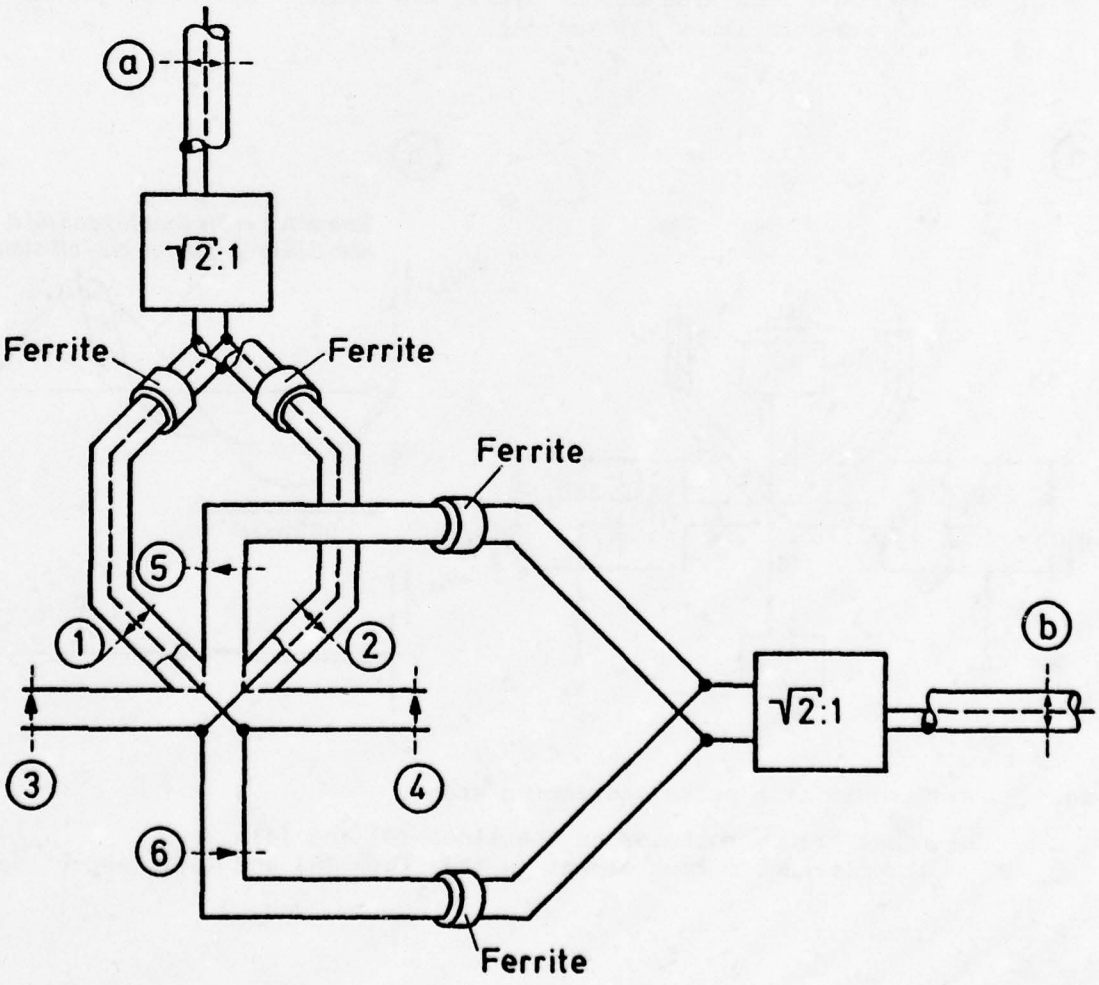


Fig. 5: Circuit of the hybrid tee.

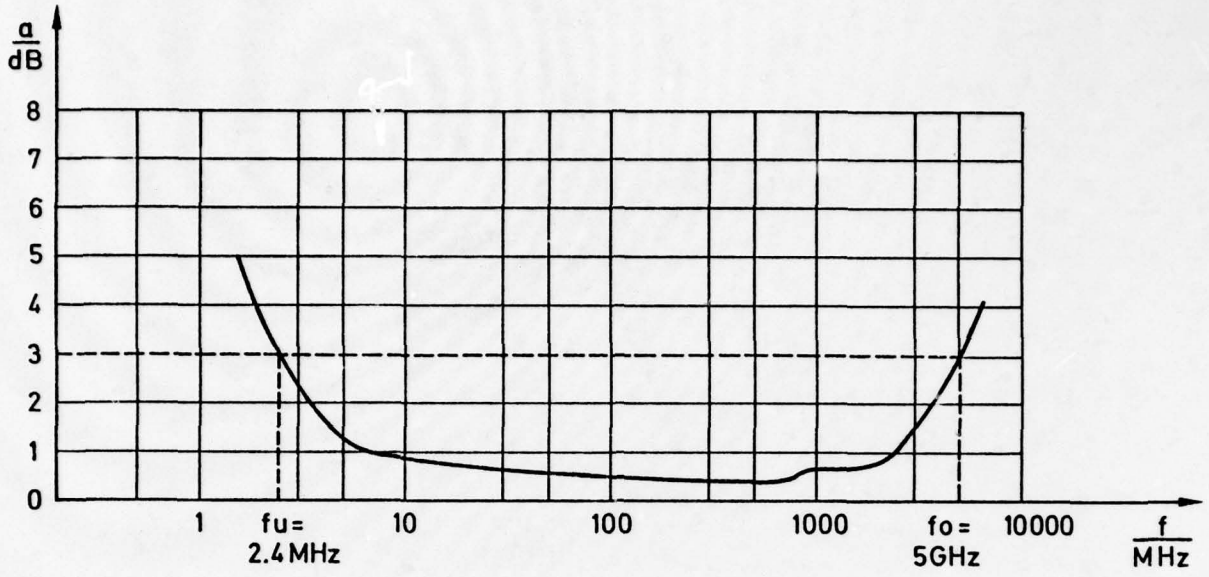


Fig. 6: Insertion loss a/dB of the hybrid tee between the lines (a) or (b) and both lines (3) and (4).

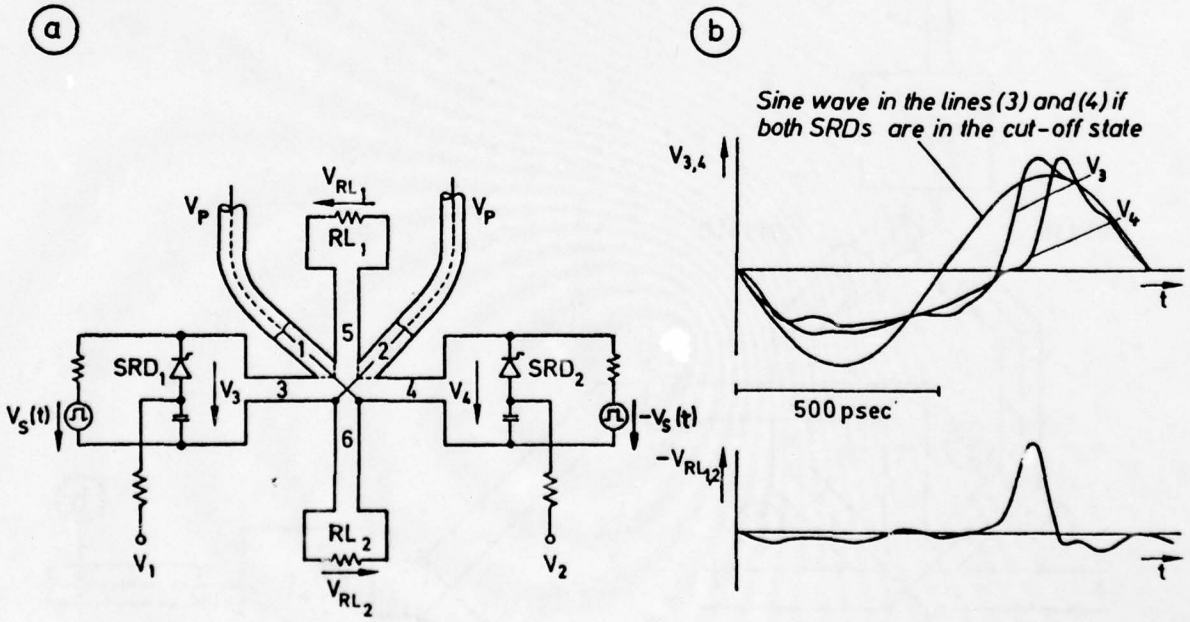


Fig. 7: a: Circuit of a pulse shortening stage.

b: upper trace: Voltages in the lines (3) and (4),
lower trace: Output signal in the lines (5) and (6), respectively.

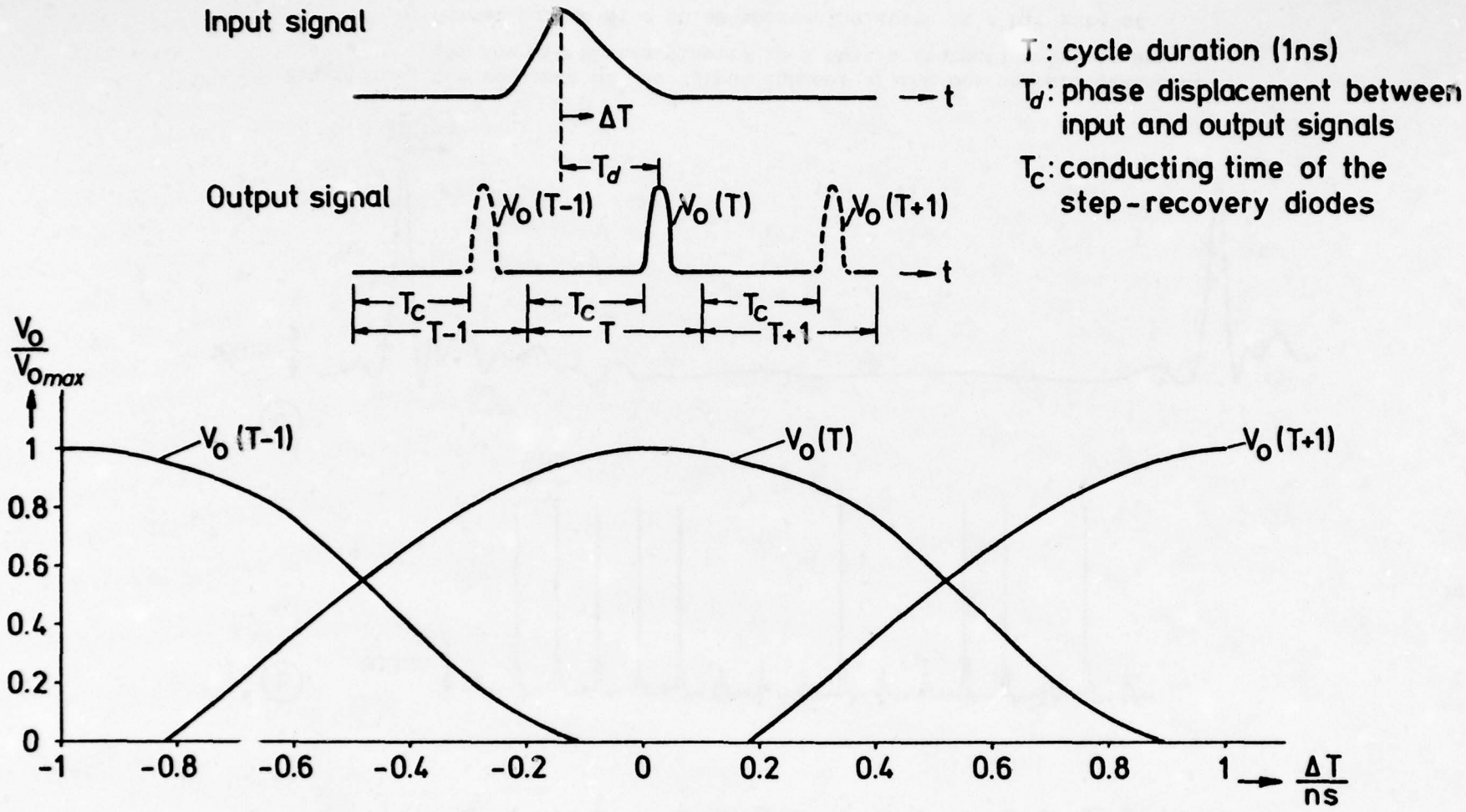


Fig. 8: Amplitudes of the output pulses at a bit rate of 1 Gbit/s when the input signals arrive at different times T.

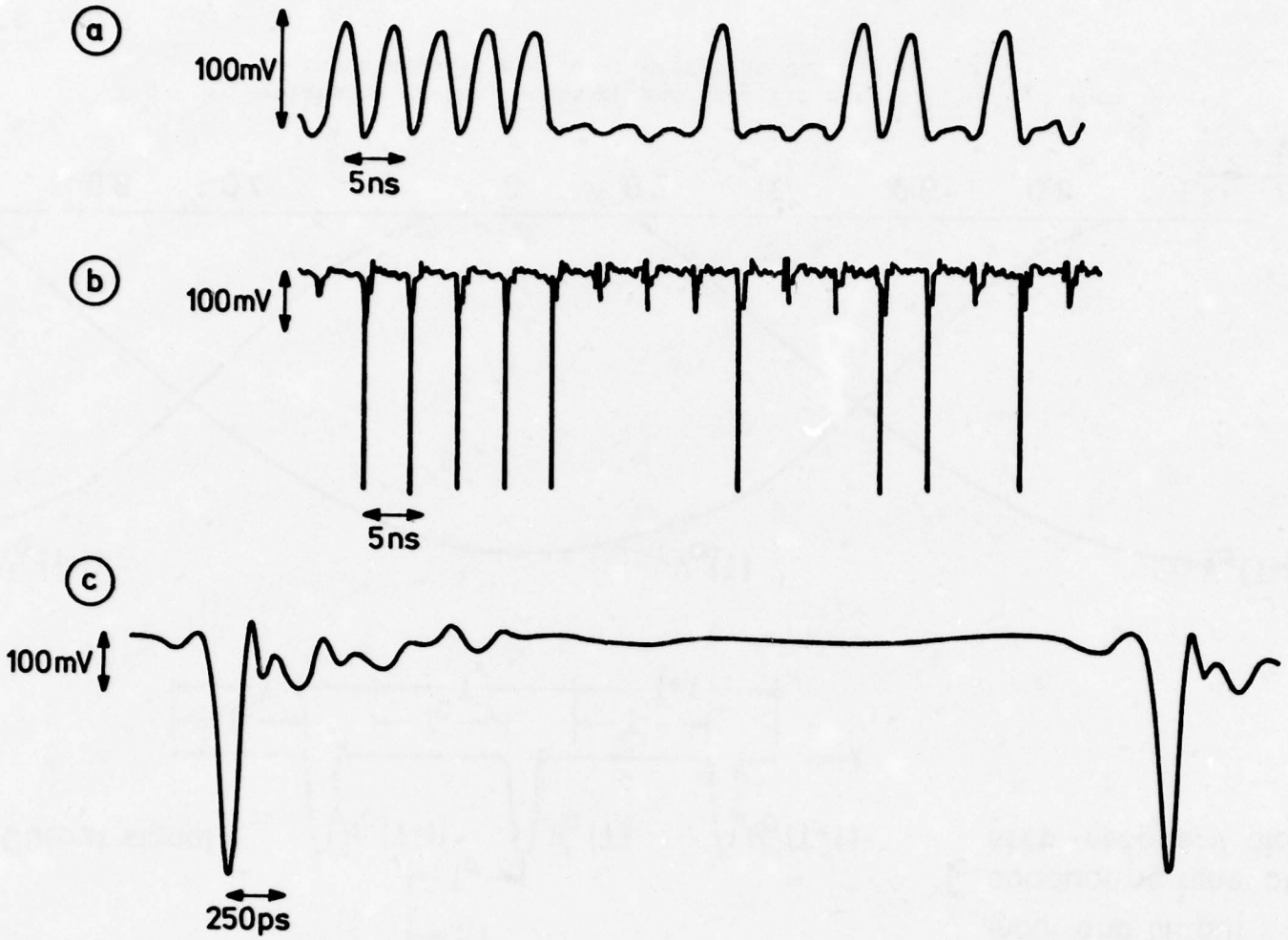


Fig. 9: a: Bit sequence of one output channel of the bit pattern generator for the 250 Mbit/s (input) to 5 Gbit/s (output) multiplexer.

b: Output pulses of a pulse shortening stage at a bit rate of 250 Mbit/s.

c: Output pulses of a pulse shortening stage in a stretched time span.

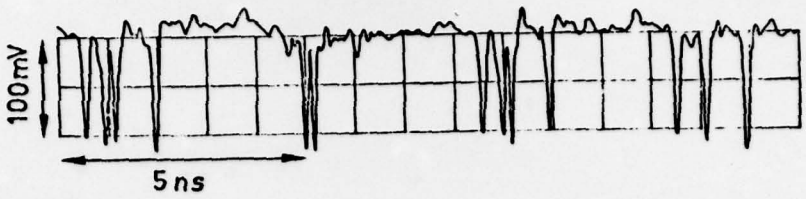


Fig. 10: Output pulses of the 250 Mbit/s (input) to 5 Gbit/s (output) multiplexer.

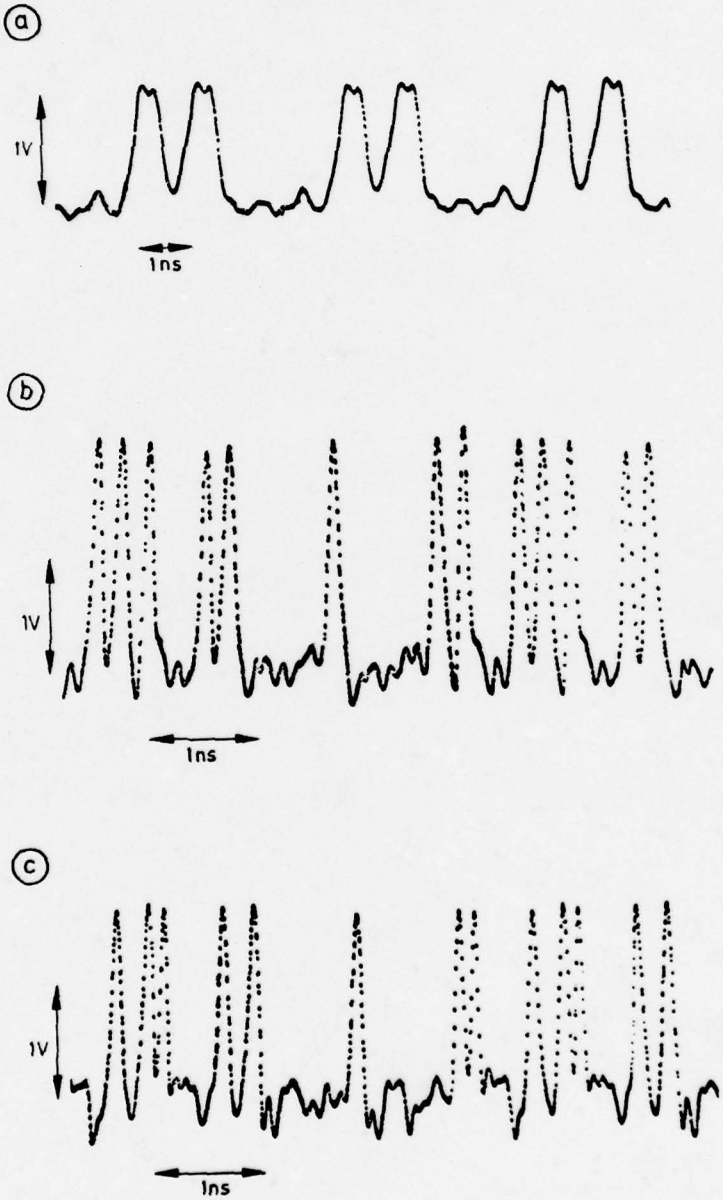


Fig. 11: a: Signals of one output channel of the bit pattern generator feeding the 1.12 Gbit/s (input) to 4.48 Gbit/s (output) and 7.84 Gbit/s (output) multiplexer, respectively.

Output signals of the 1.12 Gbit/s (input) multiplexer at a bit rate of

b: 4.48 Gbit/s and

c: 7.84 Gbit/s.

A 16 KB/S MODEM FOR SECURE VOICE SERVICE OVER NARROWBAND ANALOG CHANNELS

RICHARD A. NORTHRUP
Rome Air Development Center (RADC)
Griffiss AFB NY 13441

T. R. LOSSON, D. D. McRAE and F. A. PERKINS
Harris Electronic Systems Division
Melbourne, FL 32901

SUMMARY

This paper discusses the development, testing, and planned applications of the 16 Kbps Modem which operates over unconditioned 4 kHz voice frequency channels. The modem transmitter, receiver, automatic equalizer, and unique stored program processor are described in detail. Operation of the modem in both the training mode and data transmission mode is presented. A comprehensive test program of the modem operating at both 16 and 8 Kbps over the worldwide dialed-up circuits of the U.S. Automatic Voice Network (AUTOVON) is described along with a summary of the test results. Based upon the successful test performance, planned applications for the modem are inexpensive 16 Kbps Continuous Variable Slope Delta Modulation (CVSD) secure digital voice transmission and the 8 Kbps transmission of high quality data or imagery over dialed-up narrowband circuits. An alternate secure digital voice application is 32 Kbps CVSD, whereby a biplexer device splits the 32 Kbps into two 16 Kbps streams for transmission by the 16 Kbps Modem over two separate channels.

1. INTRODUCTION

A significant portion of avionics communications is voice transmission. For digital secure voice, the use of a 16 Kb/s Continuously Variable Slope Delta-modulation (CVSD) device is attractive from a number of standpoints: it provides good voice quality, intelligibility and speaker recognition; the voice digitizer and digital to analog conversion equipment is extremely simple and inexpensive; and, the tolerance to bit errors is such that performance is comparable to analog voice on power limited channels where adequate bandwidth is available. For these reasons and others, CVSD has been selected for use in tactical systems such as TRI-TAC. The major disadvantage of CVSD in the past has been the inability to transmit 16 Kb/s over analog 4 kHz Voice Frequency (VF) channels. However, in many avionics communications nets, it is highly desirable to utilize existing VF channels for at least some portion of the digital voice transmission.

The USAF Rome Air Development Center (RADC) and the Harris Corporation's Electronic Systems Division have developed and successfully tested a modem capable of transmitting 8 or 16 Kb/s over unconditioned 4 kHz narrowband analog channels. To achieve this performance, developments of a number of new techniques were found to be necessary. Both the signal design and the processing algorithms at the receiver were carefully chosen. This paper describes the basic approach used in the modem design and discusses the special features which have been incorporated.

The modem has received extensive testing on the worldwide U.S. Automatic Voice Network (AUTOVON) and has been proven to transmit good voice quality on a high percentage of the dialed up lines in this network. A description of these tests is presented as well as a summary of the test results. The principal communication media involved in these tests were microwave LOS, commercial and military satellite links, undersea cable and troposcatter.

Finally, a discussion of the modem's potential application to a number of communications requirements is presented.

2. MODEM DESCRIPTION

The 16 Kb/s Modem is shown in Figure 1. This model is a commercial version of a much larger breadboard model of the modem developed by RADC under contract with the Harris Corporation. The modem has two operating modes. These are the data mode and the training mode. During the training mode, the transmitter sends special signals to allow the receiver to properly adapt itself to the particular transmission path in use. We will first describe the modem operation in the data mode and then discuss the training mode.

2.1 Transmitter Operation in the Data Mode

Figure 2 presents a functional block diagram of the transmitter portion of the modem operating in the data mode at 16 Kb/s. The transmitted format is basically that of suppressed carrier quadrature amplitude modulation (QAM) at a symbol rate of 2.667 Kb/s, thus requiring 6 bits per symbol to achieve 16 Kb/s. The 64 symbol possibilities are arranged as combinations of 16 phases and four amplitudes as illustrated in Figure 3 where the X and Y components into the quadrature modulator are represented as orthogonal axes.

To obtain the choice of transmitted symbol from the incoming serial 16 Kb/s data, the data is first randomized by modulo-two addition with another 16 Kb/s signal obtained from an 11 bit maximal length shift register. The result is then converted to six-bit parallel words each symbol time. Four of these bits are used to form the 16-ary phase decisions and two are used to form the 4-ary amplitude decisions. The four phase bits are first grey decoded and then accumulated in a four bit accumulator. The purpose of the grey decoder is to incorporate phase decision at the receiver whereby an error in phase corresponding to mistaking adjacent nodes (the most common type of error) causes only one bit to be in error. The accumulator is used to provide for differential phase transmission. In this manner the phase information bits are associated with the difference between phase nodes used in time-adjacent symbols.

The amplitude bits are also grey decoded (grey to binary conversion) and input along with the four phase bits from the accumulator to the X and Y mapping programmable read-only memories (PROMS). These PROMS select the appropriate values of X and Y as shown in Figure 3 for each of the 64 possible inputs and supply the corresponding eight-bit output to digital-to-analog (D/A) converters. These outputs are then filtered by baseband filters with bandwidth of approximately 1.4 kHz. The resulting waveforms are used to suppress carrier modulate two orthogonal carriers. The carrier frequency is 1.8 kHz. The two modulated outputs are linearly summed and filtered by a bandpass filter before being applied to the telephone coupler with appropriate adjustments to the output level. Both the symbol related timing and the 1.8 kHz carrier signals are derived from a stable oscillator which has a long term stability of one part in 10^8.

2.2 Receiver Operation in the Data Mode

A functional block diagram of the receiver portion of the modem is shown in Figure 4. The receiver can be viewed as containing four separate functions. The first is the QAM demodulator which receives the line signal, demodulates it into X and Y components, and digitizes interleaved samples of these two signals for inputs to the equalizer. The second component is a 240-tap adaptive transversal filter or automatic equalizer which reduces the intersymbol interference on the X and Y estimates to an acceptable level. The third component is a stored program processor which further processes the X and Y estimates from the equalizer to combat phase jitter, makes amplitude and phase decisions, supplies estimates of X and Y error to the equalizer so that it can continuously update its weights, and makes a phase error estimate for the tracking oscillator in the QAM demodulator. The fourth function can be called the output function and consists of grey coders, parallel to serial converters, and a derandomizer to reverse the operations on the serial bit stream which were performed at the transmitter. We will now discuss each of these operations in more detail.

2.2.1 QAM Demodulator

The QAM demodulator contains a bandpass input filter to remove unwanted out-of-band signals. This is followed by a digital AGC circuit which remains at a fixed level during the data mode (it is originally set during the training mode as will be discussed later). The output of the AGC is supplied to two suppressed-carrier demodulators which use a local estimate of the transmitter carrier from a digital phase-locked loop. The loop is initially set in frequency and phase during the training mode and continues to track at a slow rate during the data mode. The demodulated signals are then sampled and held and multiplexed so that sequential digitized samples can be output to the equalizer. Two samples are furnished to the equalizer each symbol time from each of the two QAM demodulated outputs for a total of four new samples per symbol. This is a higher sampling rate than normally used in high rate modem operation and, as will be discussed in the next paragraph, is felt to be one of the major contributors to the superior performance exhibited by the modem.

2.2.2 Adaptive Equalizer

The equalizer is an adaptive transversal filter which weights 240 samples (120 complex samples) from the QAM demodulator to make each estimate. Thus, the equalizer utilizes samples from the input which span 60 symbols (or 22.5 milliseconds) since there are four samples being weighted from each input symbol. Two estimates are performed each symbol, one corresponding to the value of the transmitted X component and one corresponding to the estimate of the transmitted Y component. The equalizer weights are originally established during the training mode, but are continually updated during the data mode. The method of updating these weights is by accumulating a scaled product of the estimated X or Y error and the appropriate input value which was used with that weight to establish the error involved. This correlation technique was suggested by Lucky in 1967 (Lucky, R. W. and Rudin, H. R., 1967). The technique of establishing the X and Y error in the processor will be discussed later.

The use of two samples each pulse is key to the good performance achieved in the modem. Although some recent recognition has been given to the merits of decreasing sample spacing in the equalizer (Vungerboeck, G., 1976 and Qureshi, W. V. H. and Forney, G. D., 1977), available high-rate (9.6 Kb/s) wireline modems that are known to the authors utilize only one sample per symbol. This increased sampling rate not only allows for better crosstalk cancellation, but makes the degree of cancellation essentially independent of the time within the symbol in which the samples are taken. Since the symbol timing of the transmit and receive modems are both derived from a very stable clock, the equalizer can be used to adjust weights as the relative symbol timing changes, which removes the requirement for a symbol timing loop at the receiver. The stability of the transmit-receive clocks is such that approximately 24 hours of operation without resynchronization is available before the slip of the symbol time clocks significantly degrades the equalizer performance. In applications where longer time periods of continued operation (without re-sync) are required, a relatively simple symbol tracking algorithm based upon equalizer weights can be used to keep the relative clock drifts within allowable bounds indefinitely. The elimination of a symbol timing loop in the ordinary sense makes symbol slips essentially impossible, hence removing this as a problem associated with crypto units involved.

2.2.3 Stored Program Processor

The stored program processor performs a number of sequential calculations each symbol time. These operations are:

(a) The equalizer X, Y estimates are converted to phase.

(b) The new phase value is stored with the phase values of the preceding seven received symbols. A phase reference is then adjusted such that the mean-square phase error between these eight phase values and the closest possible transmitted phase nodes (based on this reference) is minimized.

(c) A differential phase decision is then made as the difference between the closest phase nodes to the two phase values corresponding to the input samples that occurred in the center of the time spanned by the samples corresponding to the eight phase values.

(d) The difference between the phase reference and zero is fed to the QAM demodulator as a phase error.

(e) The magnitude of the X and Y estimate is established and used to make an amplitude decision. This decision is delayed four symbols so that it is available at the same time the corresponding differential phase decision is made. Also, an amplitude error is established.

(f) The phase and amplitude errors are used to establish the X, Y errors fed back to the equalizer. The procedure for accomplishing this is to assume that the transmitted node was located at a point corresponding to the actual received X, Y value (from the equalizer) shifted in phase by the phase error and modified in magnitude by the amplitude error. The X, Y components of this error are calculated and supplied to the equalizer. This technique is necessary since the individual phase node decisions may be unreliable when the differential phase decisions are good. Hence, an attempt to use the actual decided nodes as a basis for establishing equalizer error would cause the equalizer to fail under conditions when the amplitude and differential phase decisions (the data decisions) were still good.

It is felt that both the phase tracking algorithm and the technique for maintaining equalizer weights are important to the modem performance.

2.2.4 Output Functions

The final functions in the receiver are very straightforward. The differential phase decisions are grey coded to reverse the grey decoding operation at the transmitter. In this manner, the individual symbol errors which are likely to be in error by only one node in either phase or amplitude contain only one bit error in the output. Since 16 Kb/s CVSD is capable of operating satisfactorily at error rates that are quite high (5%), the reduction in bit error rate provided by this coding is actually quite important since it significantly increases the allowable symbol error rate, and hence, its overall robustness under difficult channel conditions.

After grey coding (binary to grey conversion), the amplitude and phase decision bits are combined, converted to serial and de-randomized by an 11 bit maximal-length pseudo-noise (PN) generator identical in design to that used in the transmitter for randomizing. The synchronization of this register with the randomizing register is accomplished during the training mode.

2.3 8 Kb/s Data Mode

In the 8 Kb/s mode, the transmitter of Figure 2 utilizes only three bits per symbol from the serial-to-parallel converter. These three bits are grey coded, accumulated and used to provide an eight-phase constant-amplitude mapping shown in Figure 5. The receive functions are essentially identical to that associated with 16 Kb/s except that the processor phase tracking and decision algorithm recognizes that eight phases rather than sixteen are being transmitted. The principle of operation, however, is identical.

2.4 Training Mode

The sequence of transmission during the training mode is illustrated in Figure 6. This sequence consists of approximately one second of carrier transmission followed by eight "frames" of sync and training. Each of the eight frames consists of 32 symbols of carrier (sync) with phase 180° different from that used during the carrier transmission, followed by 2015 symbols of random transmission used by the receiver to train the equalizer. The entire sequence takes approximately 7 seconds. The transmitter mapping during the random transmissions is illustrated in Figure 7. Thus, the values for selection of the six-bit symbols are obtained directly from the 11 bit P-N generator shown in Figure 2. The six bits are subdivided into two three-bit words, one determining the X value and the other the Y value. The grey coding and accumulation logic is bypassed in this mode.

The sequence of events at the receiver during training is as follows:

(a) The receiver continuously looks for a carrier burst. It requires that the spacing of zero crossings be within a specified tolerance for a specified length of time. The required time is long enough to make the probability of recognizing carrier during normal transmission negligible. Carrier recognition is required to start the receiver training sequence.

(b) After carrier recognition, the receiver measures the carrier frequency to an accuracy better than one part in 10^5. This is determined from the zero crossings of the carrier. The digital VCO in the receiver is set to this measured frequency. The phase of the VCO is set to that corresponding to the incoming carrier phase.

(c) Following the carrier frequency and phase settings, the digital AGC sets the level of the incoming sinewave to a value determined by a reference following the A/D preceding the equalizer. The AGC level setting is not critical since the adaptive equalizer will determine the final gain setting.

(d) After setting of the AGC, the receiver looks for a phase reversal indicating that the beginning of sync has occurred. When this condition is recognized, the randomizer sequence is started at a prescribed starting state, the equalizer weights are zeroed, and symbol counters to identify frame locations are started.

(e) Gates driven from the counters are used to enable a phase measurement during the sync burst in each of the eight frames. These gates are also used to inhibit equalizer training during the sync bursts.

(f) During the portion of the frame in which random data is transmitted, the equalizer weights are updated by using the receiver randomizer as a perfect reference to establish error for the correlation operation. This comparison is performed in the processor. The initial setting of the receiver randomizer is selected to roughly center the weight pattern in the equalizer.

(g) After the symbol counter determines that the training sequence is over, the equalizer loop gain is reduced by 16 and the basis for equalizer error inputs are changed in the processor to that previously described for operation in the data mode.

The training sequence for the 8 Kb/s mode is the same as that for the 16 Kb/s mode.

3. 16 KB/S MODEM TESTING PROGRAM

In the fall of 1976, extensive tests of the 16 Kb/s Modem were conducted on the worldwide AUTOVON network. These tests were conducted in Europe, the Continental U. S. (CONUS), and in the Pacific. Trans-Atlantic tests from England to the U. S. and trans-Pacific tests from Hawaii to the U. S., the Phillippines and Japan were also conducted. Performance was evaluated at both the 16 and 8 Kb/s transmission rates.

Transmission media within Europe consisted of LOS microwave, troposcatter, some leased PTT circuits, and combinations thereof. Trans-Atlantic and trans-Pacific calls were either submarine cable or satellite. Within the CONUS and Hawaii, circuits are those leased from commercial carriers. It is significant to note that the modem operated satisfactorily on difficult tropo links and on the new submarine cable systems which use sharp 3 kHz channel filters.

Several categories of calls were placed. Calls originated from one AUTOVON switching center to another switching center were designated as interswitch trunk (IST) calls. Those from a switching center to a subscriber of the originating switch were designated as access line calls. Those from a switching center to a subscriber of a different switching center were designated as remote access calls. Calls were dialed up and transmitted over voice quality circuits with no special conditioning whatsoever.

In order to cover as much of the network as possible from a limited number of locations, most of the calls placed were loop-around calls where the transmitted signal was looped back to the originating site from the remote site. On these calls, the portion of the network that was transitioned by the signal is twice that which would be associated with an ordinary call between the two locations. Of course, the loop-around error rate performance of the modem was much poorer than would be experienced in one-way normal operation. To help calibrate the error rates achieved in the loop-around calls relative to those that would be expected from normal one-way calls between the two locations, a separate transmitter was placed at several of the locations used in the loop-around calls and error rates for both loop-around and normal configurations were made on the calls to these locations.

Table 1 presents a summary of the test results for 16 Kb/s operation categorized by general location and the type of call involved. The voice quality of 16 Kb/s CVSD is easily usable at error rates as high as 5% although the conversation is somewhat noisy at that level. Error rates less than 1% allow outstanding digital voice conversation. The percentages of calls in which bit error rates were measured to be less than 5%, 2%, and 1% are shown in Table 1. In the case of loop-around calls, predictions are made as to the percentage of those looped calls that would have been less than 5%, 2%, 1% had they been ordinary one-way calls. These estimates are based upon the calls in which data was taken for both looped and ordinary configurations. Table 2 presents a summary of the test results at 8 Kb/s.

As can be seen from Tables 1 and 2, the tests of the 16 Kb/s Modem at both the 16 and 18 Kb/s rates were successful over all communication media. 16 Kb/s CVSD digital voice can, therefore, be transmitted very reliably over worldwide dialed-up narrowband 4 kHz circuits. At 8 Kb/s, "data quality" transmission can be achieved on a very high percentage of available circuits. In addition, at 16 Kb/s, the first-time synchronization success on all circuits of 96.4% further attests to the capabilities and robustness of the 16 Kb/s Modem.

4. POTENTIAL APPLICATION OF 16 KB/S MODEM TO COMMUNICATIONS REQUIREMENTS

The excellent test results of the 16 Kb/s Modem indicate that it has a promising potential to satisfy a number of communication requirements. The ability of the modem to successfully operate over a variety of worldwide dialed-up 4 kHz voice bandwidth circuits without any special conditioning provides a very flexible communications capability for secure digitized voice, digital data and imagery.

The prime application for the 16 Kb/s Modem is for secure voice transmission over voice bandwidth channels in conjunction with an inexpensive CVSD voice digitizer. With CVSD digitization, 16 Kb/s is commonly considered the lowest rate at which voice quality, intelligibility, and recognition all remain acceptable to a large number of operational users. Before the development of the 16 Kb/s Modem, the 16 Kb/s rate could only be achieved over wide bandwidth analog facilities or on digital PCM/TDM systems. Now this capability is available to any user of the worldwide switched network such as the U. S. AUTOVON system.

Due to standardization and system interoperability considerations, 32 Kb/s digitized voice with CVSD will remain a required configuration for some systems. For 32 Kb/s operation, the 16 Kb/s Modem can provide this transmission rate with a 32 Kb/s biplexer using two (2) modems and two (2) 4 kHz narrowband channels as shown in Figure 8. The biplexer splits the 32 Kb/s stream into two 16 Kb/s streams and recombines to 32 Kb/s at the receiver, taking into account the transmission time differences between the two circuits.

Table 1. Summary of Test Results at 16 Kbps

Type of Call	No. of Calls	Median BER	% of Looped Calls with BER≤ 5%	2%	1%	% of One-Way Calls with BER≤ 5%	2%	1%	Synchronization Performance Tries	Successes	%
European IST Loops	85	1.03E-3	92	80	70	100	98	96	346	330	95.4
Pacific IST Loops	25	4.50E-4	100	99	97	100	100	100	93	93	100
CONUS IST Loops	78	4.70E-3	100	85	80	100	100	100	234	234	100
European One-Way	21	5.40E-4	-	-	-	100	96	90	91	87	95.6
Trans-Atlantic One-Way	32	4.00E-3	-	-	-	100	92	75	161	161	100
Trans-Pacific One-Way	24	1.18E-4	-	-	-	100	100	99	109	109	100
European Access Loops	34	3.80E-5	100	99	97	100	100	100	151	139	92.1
European Remote Access Loops	26	1.03E-2	85	66	46	95	93	88	118	102	86.4
Pacific Remote Access Loops	9	3.80E-3	89	73	67	100	100	100	27	27	100
TOTALS	334								1330	1282	96.4

Table 2. Summary of 8 Kb/s Test Results

Type of Call	% of Calls with BER ≤ 10^{-4}	% of Calls with BER ≤ 10^{-5}	Median BER	Median Block Throughput
European IST Loops	79%	47%	1.5×10^{-5}	99.3
Pacific IST Loops	94%	92%	$< 10^{-5}$	100
CONUS IST Loops	99%	89%	$< 10^{-5}$	100
European IST One-Way	83%	55%	$< 10^{-5}$	95.5
Trans-Atlantic One-Way	97%	76%	$< 10^{-5}$	100
Trans-Pacific One-Way	83%	73%	$< 10^{-5}$	100
European Access Loops	94%	83%	$< 10^{-5}$	100
European Remote Access Loops	59%	35%	5.2×10^{-5}	96.6
Pacific Remote Access Loops	83%	73%	$< 10^{-5}$	100

Finally, the excellent performance of the 16 Kb/s Modem at the 8 Kb/s transmission rate offers several communication alternatives such as record data transmission, imagery communications and higher quality 16 Kb/s data using a biplexer, two modems, and two communications channels. Flexibility and economics at the 8 Kb/s rate are important factors due to the modem's ability to operate on dialed-up circuits. Presently available 9.6 Kb/s modems normally require special conditioned lines which limit flexibility of operation and increase cost of implementation.

5. CONCLUSIONS

The development and successful testing of the 16 Kb/s Modem have proven the feasibility of transmitting 16 Kb/s over 4 kHz narrowband analog channels. With inexpensive CVSD voice digitization, the 16 Kb/s Modem provides a flexible, inexpensive, high quality secure voice capability. In addition the 8 Kb/s

rate can be used for data quality transmission of imagery and record data. The U. S. Air Force plans to continue the development of the 16 Kb/s Modem into the advanced and engineering development phases so that fully supportable production hardware can be available in support of operational requirements by the early 1980's time period.

REFERENCES

LUCKY, R. W. and RUDIN, H. R., "An Automatic Equalizer for General Purpose Communication Channels", Bell System Technical Journal, 46, No. 9, Part 2, November 1967, pp 2179-2208.

QURESHI, S. V. H. and FORNEY, G. D., "Performance and Properties of a T/2 Equalizer", NTC '77 Proceedings, Volume 1, December 1977, pp 11:1-1, 11:1-9.

VUNGERBOECK, G., "Fractional Tap Spacing Equalizer and Consequences for Clock Recovery in Data Modems", Volume Com-24, August 1976, pp 856-863.

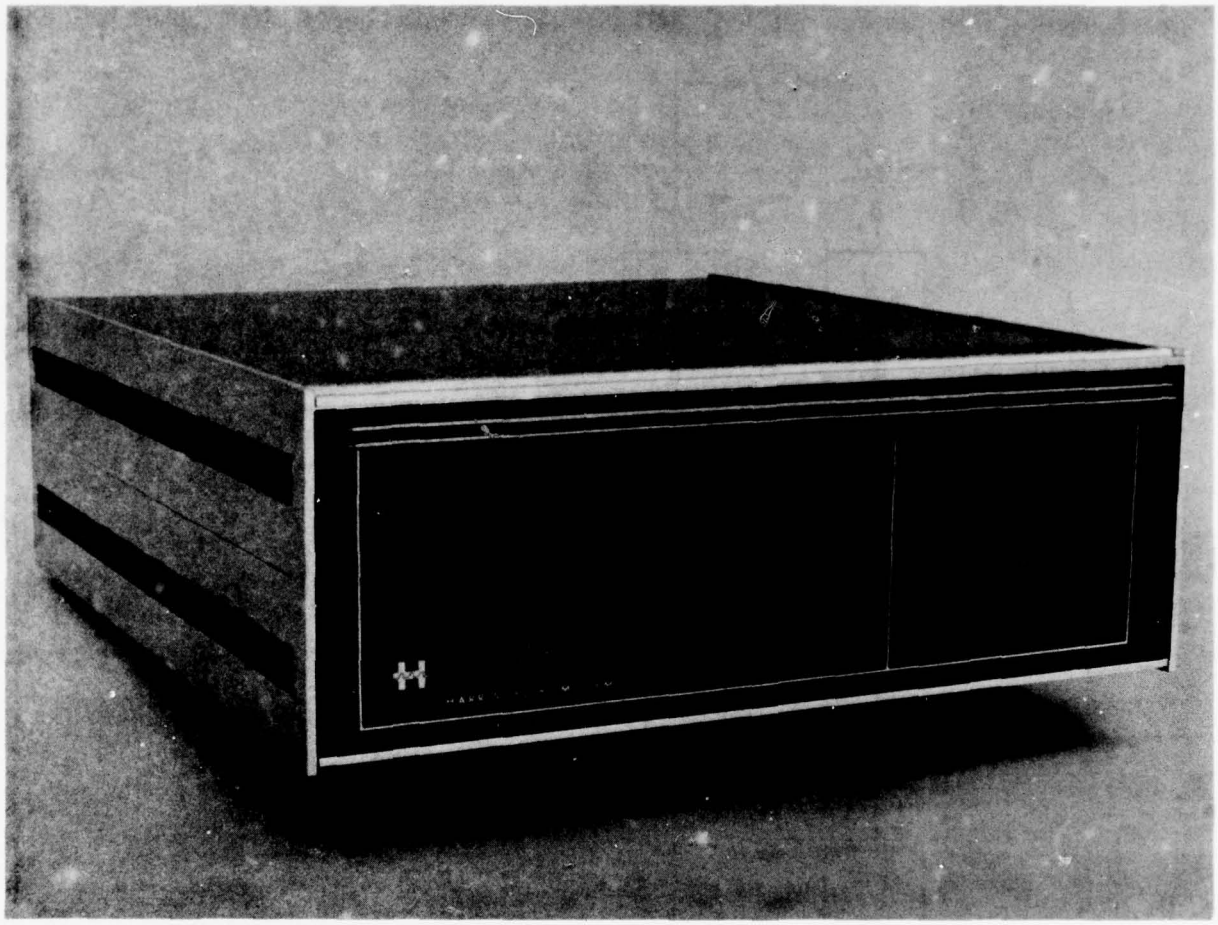


Fig.1 16 Kb/s Modem

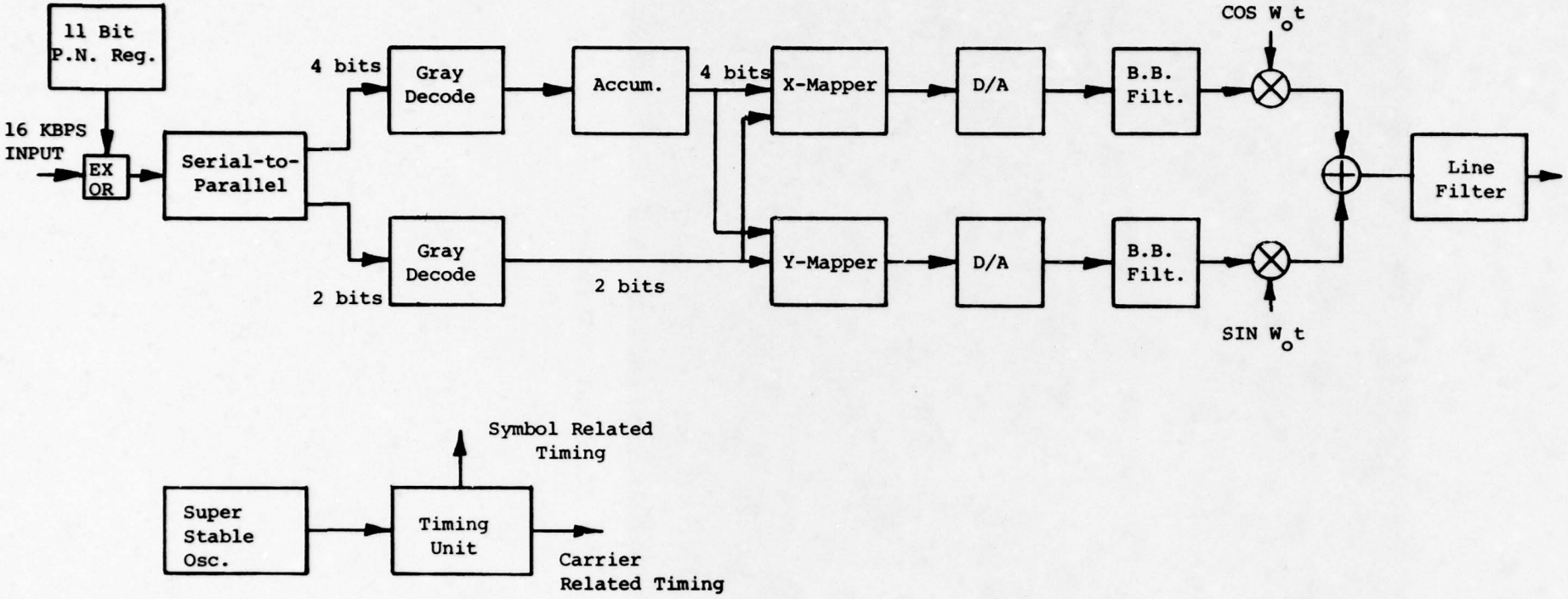


Fig.2 Transmitter block diagram during the data mode at 16 Kb/s

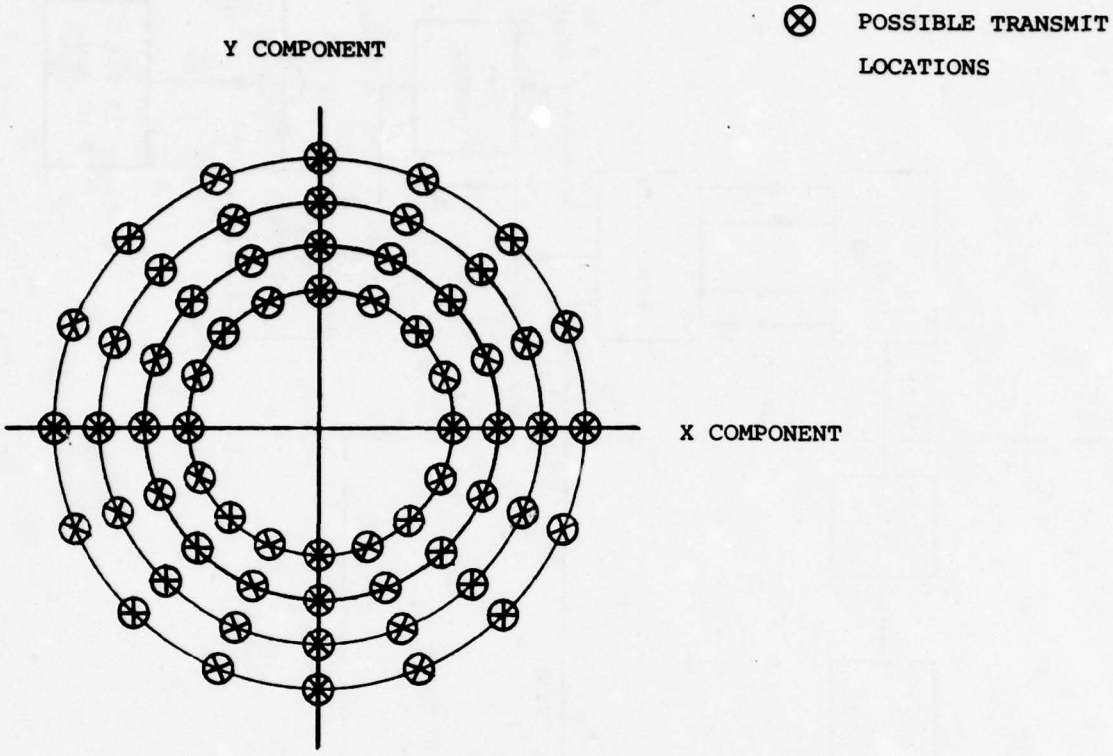


Fig.3 Transmitter mapping during the data mode at 16 Kb/s

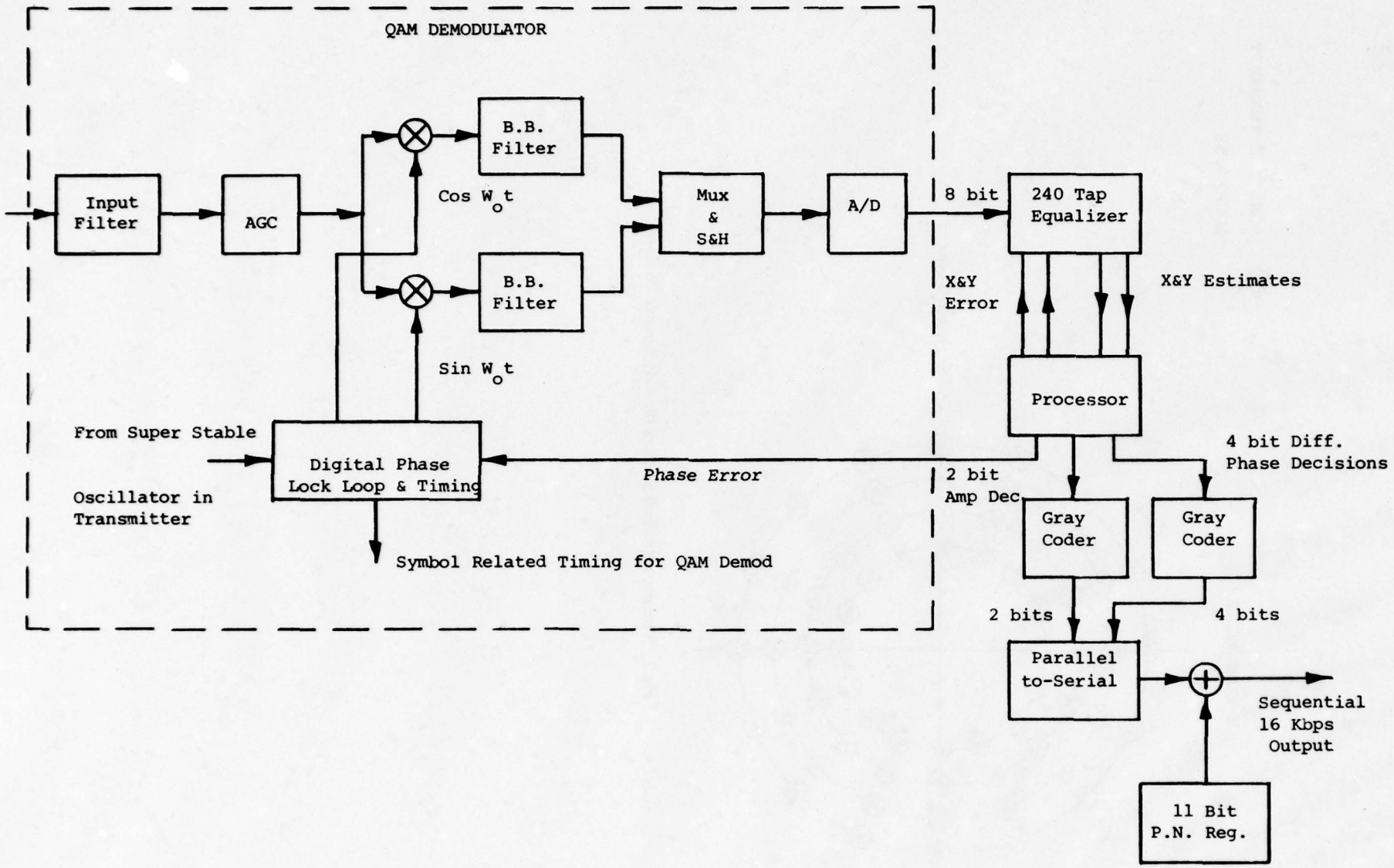


Fig.4 Receiver block diagram during the data mode at 16 Kb/s

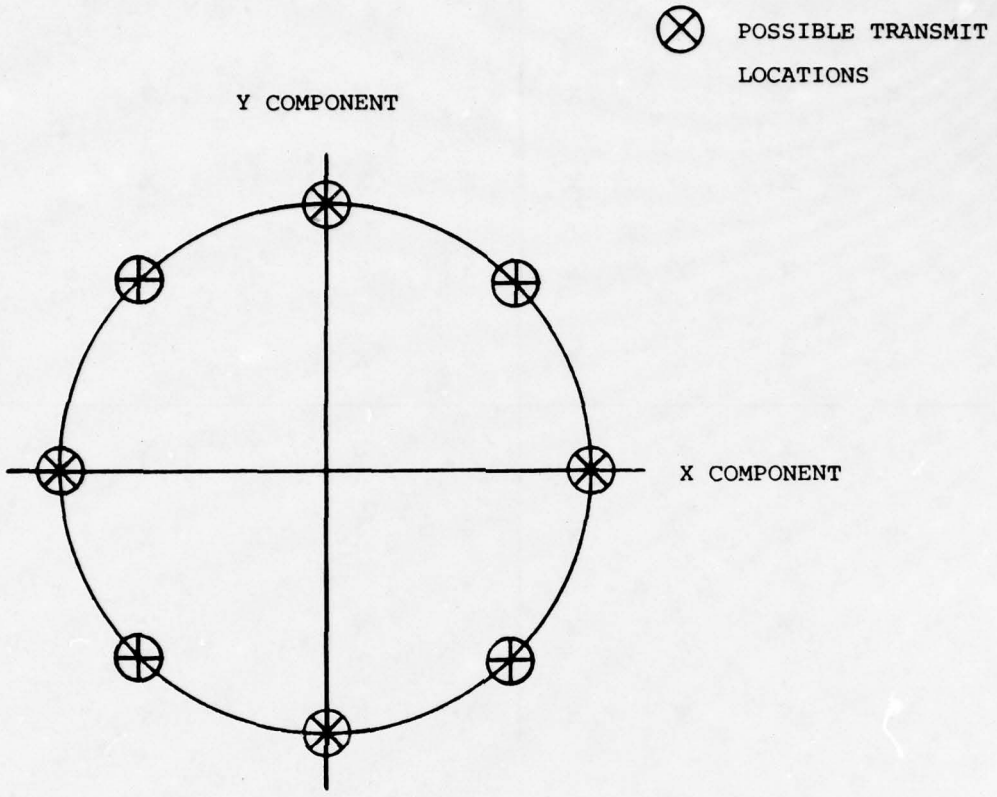


Fig.5 Transmitter mapping during the data mode at 8 Kb/s

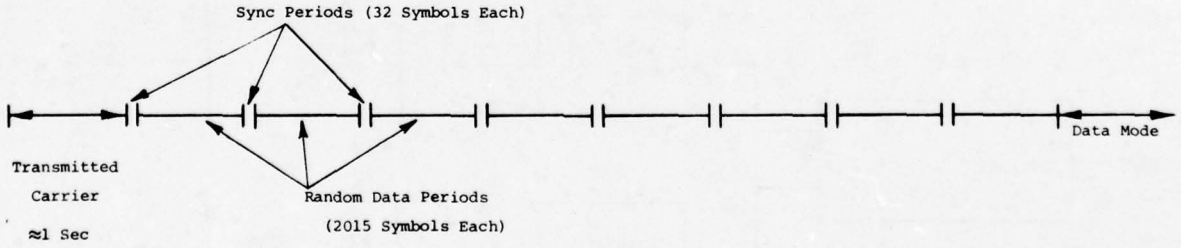


Fig.6 Transmitter training sequence

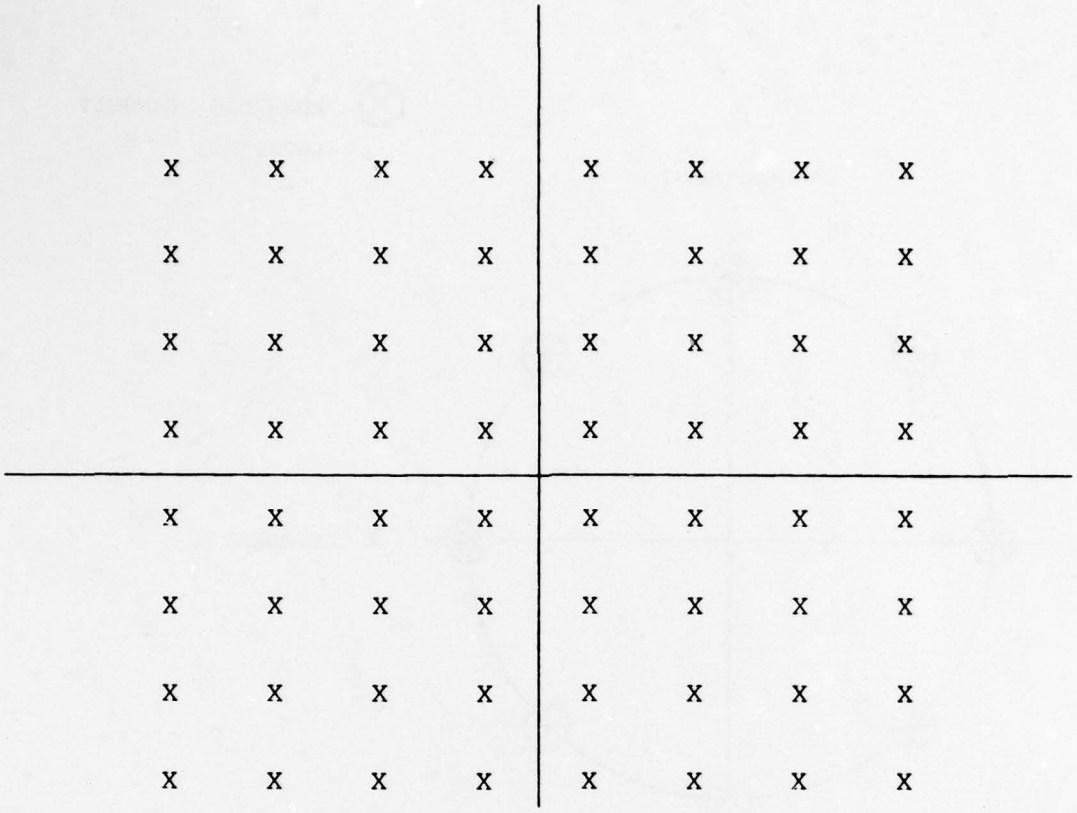


Fig.7 Transmitter mapping during the training mode

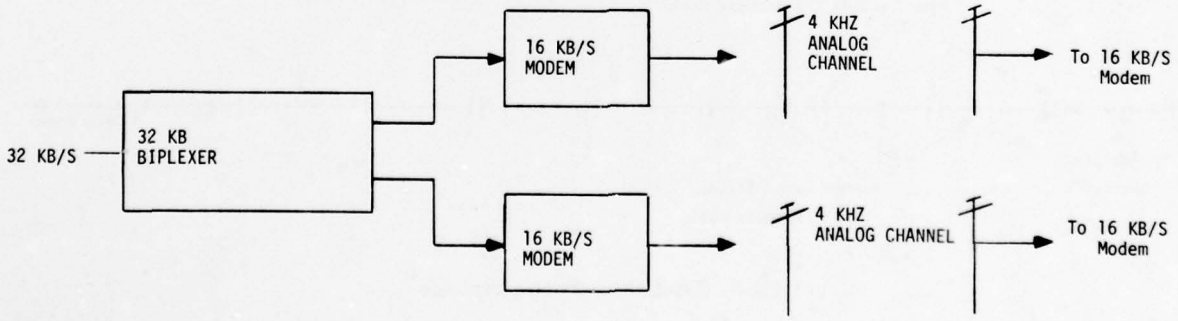


Fig.8 32 Kb/s biplexer configuration

DISCUSSION

J.Buchau, US

How did you actually establish the BER's you reported?

Author's Reply

BER's were measured for each test by counting errors received in a known transmission pattern over a fixed period of time which was usually 10 seconds.

E.Meinel, Ge

What is the manner of operation for the crypto set?

Author's Reply

We have tested the 16 Kbps Modem with standard US crypto sets such as the KG-13 and KG-34. We also plan to use the 16 Kbps Modem with advanced crypto schemes which are currently under design.

M.Alexis, Fr

Avez-vous effectué les tests d'intelligibilité comparés entre le modem à 16 Kbps et une liaison 4 kHz analogique classique?

Author's Reply

We did not perform comparative intelligibility tests between 16 Kbps CVSD and analog voice. However, comparative user tests over the telephone system showed no perceptible difference between 16 Kbps CVSD with an error rate of less than 1% and clear analog voice.

E.Ante, Ge

To what signal to noise value do the bit error rates in the tables correspond?

Author's Reply

Most of the communications channels have signal-to-noise ratios of 30 to 35 db.

N.Tepedelenlioğlu, Tu

What algorithm and criterion are you using in adjusting the transversal equalizer?

Have you made any attempts to shorten the training period?

Author's Reply

The transversal equalizer weights are updated by supplying phase and amplitude error information to the equalizer. The basic algorithm and criterion for accomplishing this function is to:

(1) Obtain a measure of the phase error by minimizing the mean-square phase error measured to the nearest phase node for a number of received phase values.

(2) Obtain a measure of the amplitude error by making an amplitude decision and measuring the difference between the received amplitude and the decided amplitude.

(3) Assure that these phase and amplitude errors appear on the tip of the received signal vector, i.e., project them in line and orthogonal to the actual recived vector.

(4) Convert the resulting vector to a projected X error and a projected Y error for use in updating the equalizer weights.

Considerable effort has been expended on experimenting with different training periods over the various worldwide (commerical/military) telephone networks. It has been found that the present 7 second training sequence can be shortened by 3 or 4 seconds where line impairments can be maintained well within modem design specification. When considering satisfactory modem operation over the general worldwide telephone networks and the marginal conditions plaguing some of those systems, the 7 second training sequence has been found to offer the best modem performance.

DOUBLE DIFFERENTIAL PSK SCHEME IN THE PRESENCE OF DOPPLER SHIFT

Mario Pent
Istituto di Elettronica e Telecomunicazioni
Politecnico - C. Duca degli Abruzzi, 24 - Torino - Italy

SUMMARY

A Double Differentially Coherent demodulation scheme (DDPSK) is proposed for digitally phase modulated signals: its performance is theoretically independent on slow carrier frequency fluctuations, and this result is obtained by means of a very simple and reliable structure.

The theoretical analysis of the ideal DDPSK presented in the paper shows that, in the absence of carrier frequency shift, it suffers a penalty of about 4 dB with respect to the conventional DCPSK demodulation scheme.

Since the potential applications of such a demodulation scheme are in the field of the Air-to-Air or Air-to-Ground data links, a comparative analysis has been made with respect to the conventional DCPSK in the presence of a carrier frequency shift, and the conditions are derived (in terms of carrier frequency, relative speed and information rate) under which the DDPSK performs better.

1. INTRODUCTION

The problems that arise in demodulating digital phase-modulated signals affected by Doppler shift are well-known: if coherent detection is considered, high order PLL's are required in order to properly track the carrier frequency; in the case of differentially coherent detection, a loss in performances is produced because of the phase offset induced by the Doppler shift, and such an impairment becomes particularly significant when high carrier frequency values and slow baud rates are considered.

All these problems are severely enhanced if a burst-mode operation is required: in such a condition the differentially coherent demodulation scheme is generally preferred, because of its simpler and faster symbol timing acquisition (the carrier recovery is obviously not required).

The paper presents a demodulation scheme, derived from the conventional DCPSK, whose performance is (at least in principle) insensitive to a Doppler shift. Other examples of modifications of the conventional DCPSK demodulator have been proposed in the literature (KATO, M; INOSE, H..1970 - LOMBARD, D.; IMBEAUX, J..1975), all aiming at some improvement in the overall performance (in terms of error rate versus signal-to-noise ratio).

If $\Delta\phi_k=\phi_k-\phi_{k-1}$ is the phase difference between two adjacent symbols, the proposed demodulation scheme operates on the double difference $\psi_k=\Delta\phi_k-\Delta\phi_{k-1}$: as it will be shown in sect. 3, such a quantity is unaffected by a carrier frequency shift, and consequently the demodulator is suitable for operating in the presence of Doppler shifts.

The theoretical analysis of the performance of such a demodulator in the presence of additive Gaussian noise is presented in sect. 4, where an ideal distortionless channel is assumed. In sect. 5 a comparison between the proposed demodulation scheme and the conventional DCPSK demodulator is outlined, having in mind the potential applications of the demodulator in the field of the Air-to-Air or Air-to-Ground digital communications.

2. THE DCPSK RECEIVER IN THE PRESENCE OF A DOPPLER SHIFT.

Let us consider a standard 4Φ-DCPSK demodulator, as shown in fig. 1. The received signal can be written as:

$$(1) \qquad r(t) = \sum_k s(t-kT)\cdot e^{j(\omega_o t + \Delta\omega t + \phi_k + \phi_o)} + n(t)\cdot e^{j\omega_o t}$$

where $s(t)$ is the equivalent low-pass response of the channel (including the RX filter) to the elementary waveform produced by the transmitter, ω_o is the nominal angular carrier frequency, $\Delta\omega$ is the Doppler shift, ϕ_k is the information-carrying phase in differential format, such that $\phi_k-\phi_{k-1}$ can assume one of the values $(0,\pi/2,\pi,3\pi/2)$, ϕ_o is the initial phase, and $n(t)$is the low-pass representation of the incoming noise, which is assumed to be a bandpass Gaussian process.

In the absence of noise, at the output of the in-phase branch of the demodulator we obtain a signal $z_R(t)$:

$$(2) \qquad \begin{aligned} z_R(t) &= \mathrm{Re}\{r(t)r^*(t-T)e^{-j\delta}\} = \\ &= \mathrm{Re}\{e^{j[(\omega_o+\Delta\omega)T - \delta]}\sum_{kh} s(t-kT)s^*(t-T-hT)e^{j(\phi_k-\phi_h)}\} \end{aligned}$$

If the channel transfer function is symmetrical around the carrier frequency ω_o, $s(t)$ is a purely real signal; in addition, the phase shift δ is generally chosen in such a way that $\omega_o T-\delta=\pi/4$; under these assumptions, $z_R(t)$ can be written as:

$$(3) \qquad z_R(t) = \sum_{kh} s(t-kT)s(t-T-hT)\cos(\phi_k-\phi_h+\pi/4+\Delta\omega T)$$

Similarly, the signal at the output of the quadrature branch $z_Q(t)$ is easily found to be:

$$(4) \qquad z_Q(t) = \sum_{kh} s(t-kT)s(t-T-hT)\sin(\phi_k-\phi_h+\pi/4+\Delta\omega T)$$

It can be verified that, in the absence of intersymbol interference and without Doppler shift, the digital information can be recovered on the basis of the signs of $z_R(t)$ and $z_Q(t)$ sampled at appropriate instants t_o+mT.

From equations (3) and (4) we can find that the Doppler shift $\Delta\omega$ acts as a rotation of an angle of $\Delta\omega T$ radians on the signal plane. The effects of such a rotation on the performance of the demodulator in the presence of noise are shown in fig. 2, where the increment of the signal-to-noise ratio necessary to achieve an error rate of 10^{-4} is given as function of the normalized frequency shift $\Delta\omega T$.

3. THE DOUBLE DIFFERENTIAL DEMODULATION SCHEME.

Let us consider the phase term of the k-th received signal element in the absence of noise (real distortionless $s(t)$ is assumed):

$$(5) \qquad \theta_o = \omega_o t + \Delta\omega t + \phi_k + \phi_o$$

for the T-delayed replica, we have:

$$(6) \qquad \theta_1 = \omega_o(t-T) + \Delta\omega(t-T) + \phi_{k-1} + \phi_o$$

for a 2T-delayed replica, the phase term becomes:

$$(7) \qquad \theta_2 = \omega_o(t-2T) + \Delta\omega(t-2T) + \phi_{k-2} + \phi_o$$

The second difference

$$(8) \qquad \psi_k = (\theta_o-\theta_1) - (\theta_1-\theta_2) = \theta_o - 2\theta_1 + \theta_2 = \phi_k - 2\phi_{k-1} + \phi_{k-2}$$

is completely independent on the frequency shift $\Delta\omega$; clearly such a result is valid if

$\Delta\omega$ remains constant within a 2T time interval: in other words, if the Doppler shift is slowly varying with respect to the symbol rate.

So, a phase demodulator based on the second difference ψ_k will be independent on a carrier shift; clearly, the transmitted phases ϕ_k must be coded in such a way that ψ_k represents the current information being transmitted. The block diagram of a possible implementation of such a demodulator (4ϕ case) is shown in fig. 3.

4. ERROR RATE EVALUATION OF THE DOUBLE DIFFERENTIAL DEMODULATOR.

If we assume an ideal distortionless noiseless channel, the received signal at the k-th sampling instant t_k has the expression:

$$r'(t_k) = A \cdot e^{j(\omega_o t_k + \phi_k)} \tag{9}$$

where A is the amplitude of the signal and ϕ_k is the current phase. If we consider the same signal with additive noise, its expression becomes:

$$r(t_k) = \rho_k \cdot e^{j(\omega_o t_k + \phi_k + \theta_k)} \tag{10}$$

where ρ_k and θ_k are random variables taking into account the superimposed noise, and their joint probability density function is:

$$W(\rho_k,\theta_k) = \frac{\rho}{2\pi\sigma^2}\, e^{-(A^2+\rho_k^2 - 2A\rho_k\cos\theta_k)/(2\sigma^2)} \qquad \rho\geq 0 \;;\; 0\leq\theta<2\pi \tag{11}$$

where σ^2 is the noise variance. The same statistics can be used to describe the signal samples taken at the time instants $t_{k+1}=t_k+T$ and $t_{k+2}=t_k+2T$. Assuming stationary noise, the variance σ^2 is the same for the three samples; in addition, we assume that the noise samples are statistically independent. This assumption implies some restictions on the noise power spectrum, namely that the autocorrelation function of the noise is negligible for delays equal to T and 2T; such an assumption is well approximated if the channel does not introduce intersymbol interference.

The demodulator is assumed to operate on the double difference $\Delta\phi$ which is given by:

$$\Delta\phi = \angle r(t_k) - 2\angle r(t_{k-1}) + \angle r(t_{k-2}) = \phi_k - 2\phi_{k-1} + \phi_{k-2} + \theta_k - 2\theta_{k-1} + \theta_{k-2} \tag{12}$$

If we consider an M-phase system, the information-bearing term $\psi_k = \phi_k - 2\phi_{k-1} + \phi_{k-2}$ can take one value out of the ensemble $\{0,2\pi/M,2\cdot 2\pi/M,\ldots,(M-1)\cdot 2\pi/M\}$. For a given value of ψ_k, an error occurs when the detected double phase difference $\Delta\phi$ lies outside the interval:

$$\psi_k - \pi/M < \Delta\phi < \psi_k + \pi/M \tag{13}$$

or,equivalently, if:

$$-\pi/M < \theta_k - 2\theta_{k-1} + \theta_{k-2} < \pi/M \tag{14}$$

Denoting with θ the phase error $\theta = \theta_k - 2\theta_{k-1} + \theta_{k-2}$, the error probability then becomes:

$$P(E) = P(\cos\theta < \cos\pi/M) \tag{15}$$

4.1 Error rate evaluation for the binary case.

Let us consider first the binary case (M=2); the error probability then becomes:

$$P(E) = P(\cos\theta < 0) \tag{16}$$

The evaluation of P(E) follows the procedure outlined in (CASTELLANI,PENT. 1977). It requires the definition of an auxiliary random variable $\lambda = \cos\theta$; if $C_\lambda(p)$ is the characte

ristic function associated with such a r.v., the error probability becomes:

$$P(E) = \frac{1}{2} + \sum_{n \text{ odd}} \frac{C_\lambda^*(n\pi)}{jn\pi} \tag{17}$$

and the values of the characteristic function needed in the expression (17) can be evaluated according to the expression:

$$C_\lambda^*(n\pi) = E\{e^{-jn\pi\cos\theta}\} \tag{18}$$

where $E\{\cdot\}$ means the statistical expectation. After heavy algebraic manipulations, the following final expression can be deduced:

$$P(E) = \frac{1}{2} - \frac{\sqrt{\pi}}{4}\,\eta\sqrt{\eta}\,e^{-(3\eta/2)} \sum_{0}^{\infty} \frac{(-1)^m}{2m+1} \{I_m(\eta/2) + I_{m+1}(\eta/2)\}^2 \cdot$$
$$\cdot\{I_{2m+\frac{1}{2}}(\eta/2) + I_{2m+1+\frac{1}{2}}(\eta/2)\} \tag{19}$$

where $\eta = A^2/2\sigma^2$ is the signal-to-noise ratio, and $I_\nu(\cdot)$ is the modified Bessel function of the first kind. It is also easy to prove the convergence of the series (19).

4.2 Extension to the M-ary case.

While the exact definition of the error probability in the general case is given by the expression (15), it is easier to evaluate an upper bound by means of the well-known "union bound" technique. With reference to the fig. 4, we can write:

$$P(E) = P(\theta \in D) \leq P(\theta \in D') + P(\theta \in D'') \tag{20}$$

By rotation of $\pi/2 - \pi/M$, we can also write:

$$P(\theta \in D') = P(\theta + \pi/2 - \pi/M \in D_o)$$
$$P(\theta \in D'') = P(\theta - \pi/2 + \pi/M \in D_o) \tag{21}$$

where the domain D_o is defined also in fig. 4. Defining a correction angle θ_M as:

$$\theta_M = \frac{\pi}{2}\left(\frac{M-2}{M}\right) \tag{22}$$

we can finally write:

$$P(E) \leq P(\theta + \theta_M \in D_o) + P(\theta - \theta_M \in D_o) \tag{23}$$

Notice that, by means of the expression (23), the evaluation of the upper bound to the error probability can be reduced to a modest modification of the method employed for the binary case. In fact, following the same procedure as before, we obtain:

$$P(E) \leq 1 - \frac{\sqrt{\pi}}{2}\,\eta\sqrt{\eta}\,e^{-(3/2)} \sum_{0}^{\infty} \frac{(-1)^m}{2m+1} \{I_m(\eta/2) + I_{m+1}(\eta/2)\}^2 \cdot$$
$$\cdot\{I_{2m+\frac{1}{2}}(\eta/2) + I_{2m+1+\frac{1}{2}}(\eta/2)\} \cdot \cos\{(2m+1)\theta_M\}$$

4.3 Results.

Expressions (19) and (24) can be evaluated by means of a digital computer; care must be taken in truncating the sum to a finite number of terms, in order to reach the best trade-off between the truncation error (due to a finite number of terms) and the approximation error (due to the inaccuracy in the computation of the various terms of the sum). The overall accuracy in the computational procedure can be estimated in the order of magnitude of 10^{-8}, which seems to be sufficient for all practical purposes.

Figure 5 shows the error probability P(E)as a function of the signal-to-noise ratio η for the cases M = 2 and M = 4; for sake of comparison, the corresponding curves for the standard DCPSK are also shown (dashed lines). We can observe that the Double Differentially Coherent demodulation scheme exhibits a penalty of about 4 dB in the binary case, and 4.7 dB for the quaternary case, with respect to the standard DCPSK in the same conditions (distortionless channel with additive Gaussian noise).

5. COMPARISON WITH THE STANDARD DCPSK IN THE PRESENCE OF DOPPLER SHIFT.

The results obtained in the previous section for the Double Differential demodulator are clearly independent with respect to a carrier frequency shift induced by the Doppler effect. On the contrary, as it was shown in sect. 2, in a standard DCPSK demodulator a Doppler shift of Δf hertz induces a rotation of $2\pi\Delta fT$ radians on the signal plane, degrading the overall system performance.

If we consider an Air-to-Air or Air-to-Ground digital transmission, the Doppler shift depends on the carrier frequency f_o and on the relative transmitter-receiver radial velocity v according to the equation:

$$\Delta f = f_o \cdot v/c \tag{25}$$

c being the light velocity. Remembering that the symbol duration T can be related to the information rate R as:

$$T = (\log_2 M)/R \tag{26}$$

M being the number of permitted phase increments, we obtain the expression of the phase shift δ due to the Doppler effect:

$$\delta = 2\pi \cdot f_o \frac{v}{c} \cdot \frac{\log_2 M}{R} \tag{27}$$

Since the error probability increases as δ increases, there exists a critical value $\delta^\dagger$ of δ for which the performance of a standard DCPSK system and of the Double Differential scheme are equal. For $\delta > \delta^\dagger$ the Double Differential scheme outperforms the standard DCPSK, and such a condition occurs when:

$$\frac{f_o}{R} > \frac{\delta^\dagger}{2\pi} \cdot \frac{c}{v \cdot \log_2 M} \tag{28}$$

Clearly, $\delta^\dagger$ depends both on the error probability chosen for the comparison, and on the value of M. Figure 6 shows, on a plane $\{f_o, R\}$, for $P(E)=10^{-4}$, M=2 and several values of the relative speed v, the boundary between the (upper) region in which the DDPSK outperforms the standard DCPSK, and the (lower) region in which the standard DCPSK exhibits better performance. Fig. 7 is the same as fig. 6, except for M=4.

6. SOME AUXILIARY REMARKS.

If we consider the expression (8) which defines the algorithm used by the Double Differential demodulator, we can see that the addition of a constant phase increment of β radians each symbol does'nt change the second difference ψ. Consequently, the demodulator can accept equally well any modification of the phase coding which adds a systematic phase increment each symbol. In particular, if we choose $\beta = \pi/M$, we obtain the well known CCITT signaling scheme "B", which, in the case of $M = 4$ is also known as "Pseudooctonary" phase modulation scheme.

Such a choice seems to be particularly interesting in the case of burst-mode operation, because of its well known properties of fast acquisition of symbol synchronizazion. The symbol timing procedure can be exactly the same as for the conventional pseudooctonary receiver.

As far as the implementation is concerned, we can estimate the hardware complexity of the DDPSK demodulator is only slightly increased with respect to the conventional DCPSK receiver, and the same technology can be employed equally well.

7. CONCLUSIONS.

The Double Differential phase demodulator examined in the paper has the property of being insensitive (almost in the ideal case) with respect to a Doppler shift; the theoretical analysis in the case of distortionless channel with additive Gaussian noise shows that in the absence of Doppler shift, it suffer a penalty of about 4 dB with respect to a conventional DCPSK receiver. Moreover, a comparative analysis with a DCPSK receiver in the presence of a Doppler shift shows that DDPSK exhibits better performance for combinations of carrier frequency and information rate which are of practical interest in the field of the Air-to-Air and Air-to-Ground digital transmission.

A further insight is needed to investigate the effects of various impairing factors originated in real channels, such as intersymbol and interchannel interference, as well as the sensitivity of the system to a symbol timing jitter.

8. BIBLIOGRAPHY.

1 - CASTELLANI, V.; PENT, M..1977, "A Computer simulation oriented approach to the error rate evaluation of DCPSK non-linear receivers", Alta Frequenza, vol. XLVI, n.8, p.334

2 - CASTELLANI, V.; PENT, M.; PATTINI, F.; PORZIO GIUSTO, P..1977, "Design guidelines for a satellite regenerative repeater", Genova, Symposium on advanced Satellite Communication Systems using the 20-30 GHz bands.

3 - KATO, M.; INOSE, H..1970, "Differentially coherent detection of phase modulated waves with error correcting capability", Electronics and Communication in Japan, vol. 53-A, n. 12, p. 54.

4 - LOMBARD, D.; IMBEAUX, J..1975, "Multidifferential PSK demodulation for TDMA transmission", London, Int. Conf. on Satellite Communication Systems Technology.

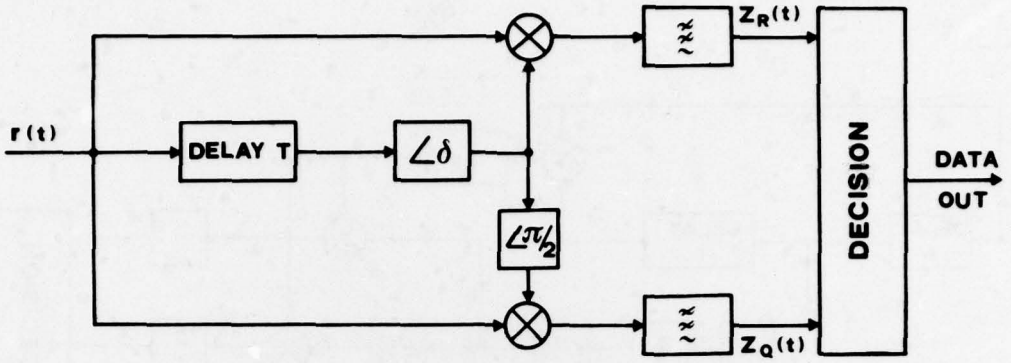


Fig. 1 - Block diagram of a standard 4ϕ-DCPSK receiver

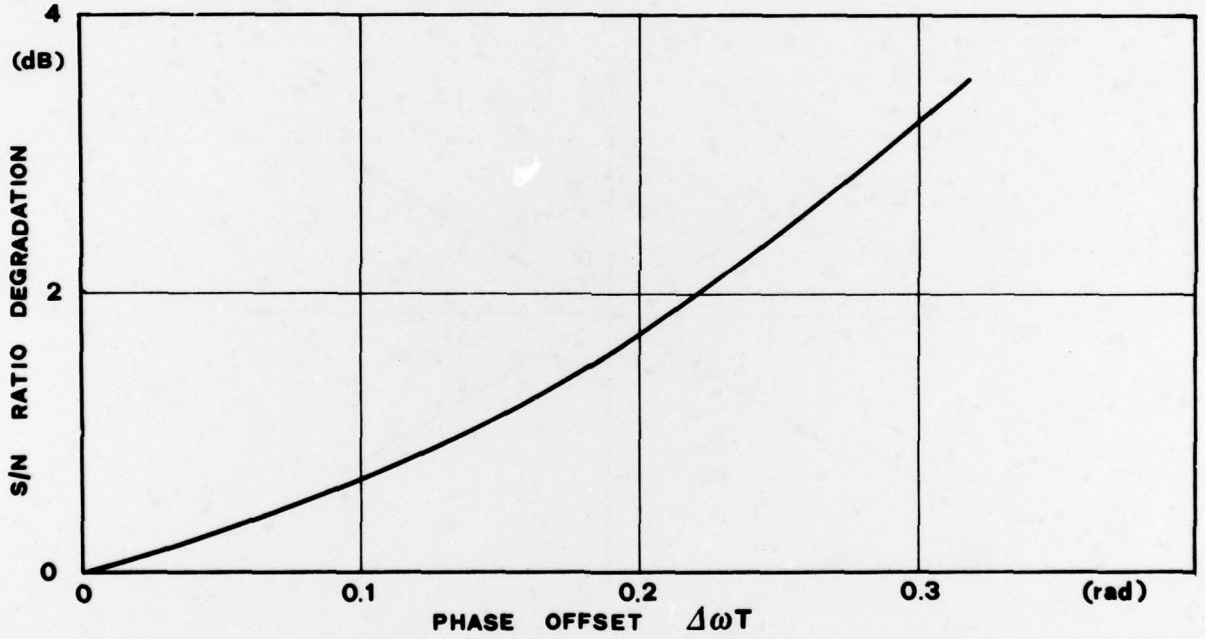


Fig. 2 - Performance degradation of a 4ϕ-DCPSK demodulator due to a normalized frequency shift $\Delta\omega T$

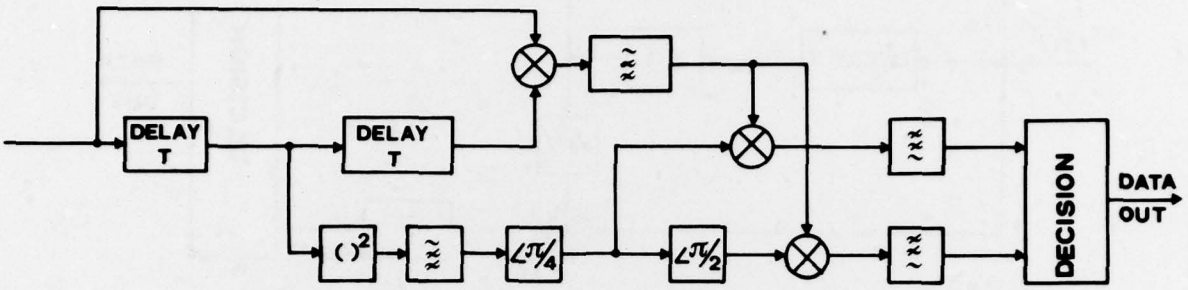


Fig. 3 - Block diagram of the Double Differential demodulator

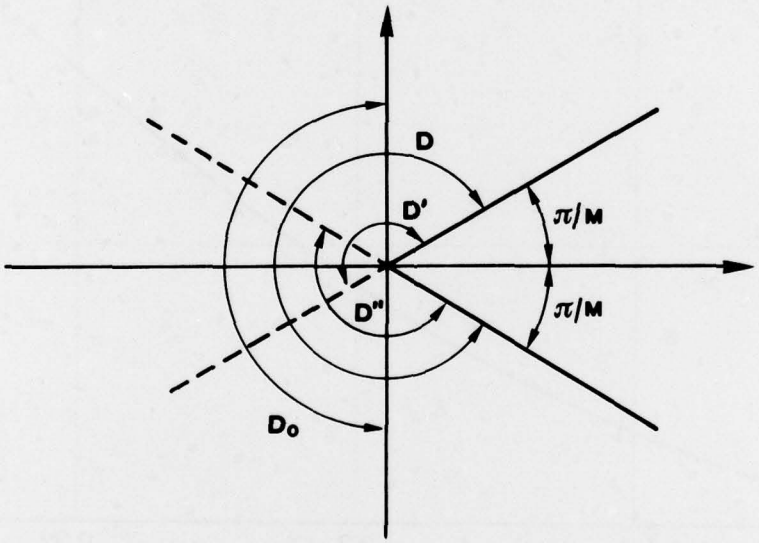


Fig. 4 - Phase diagram for M-ary systems

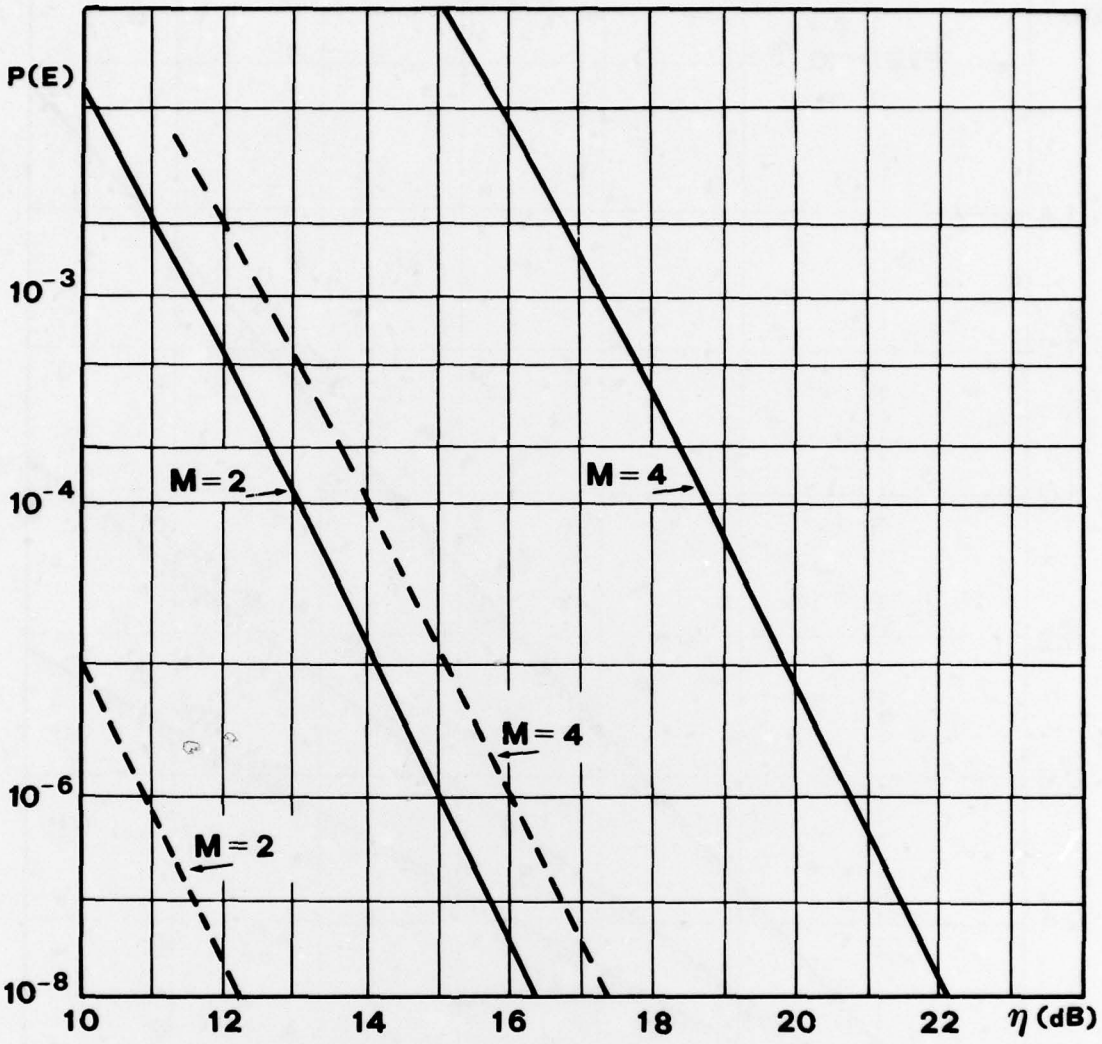


Fig. 5 - Error probability as a function of η
solid lines = DDPSK
dashed lines = DCPSK

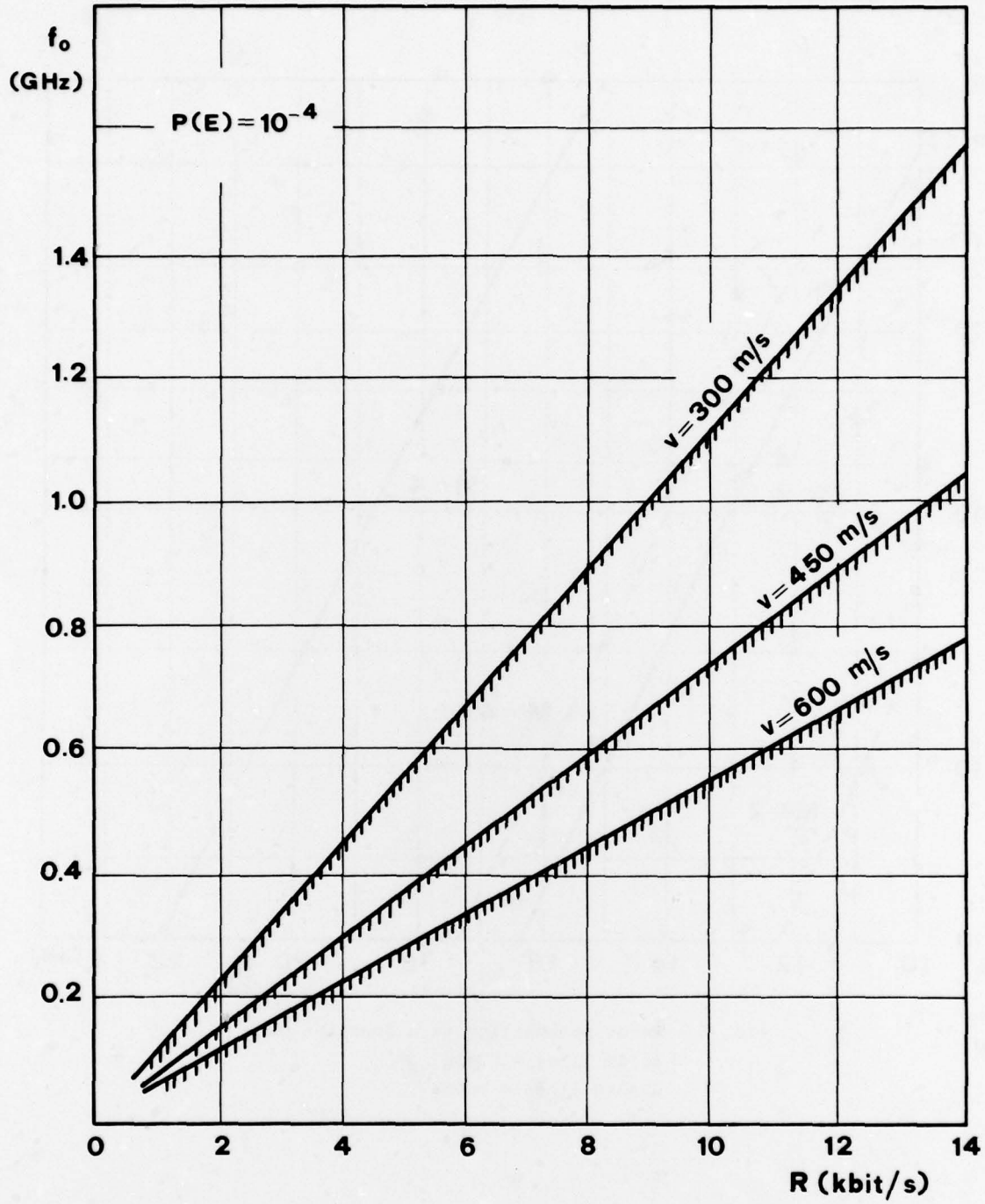


Fig. 6 - Relative performance between DDPSK and DCPSK (M=2)
in the upper regions DDPSK outperforms DCPSK

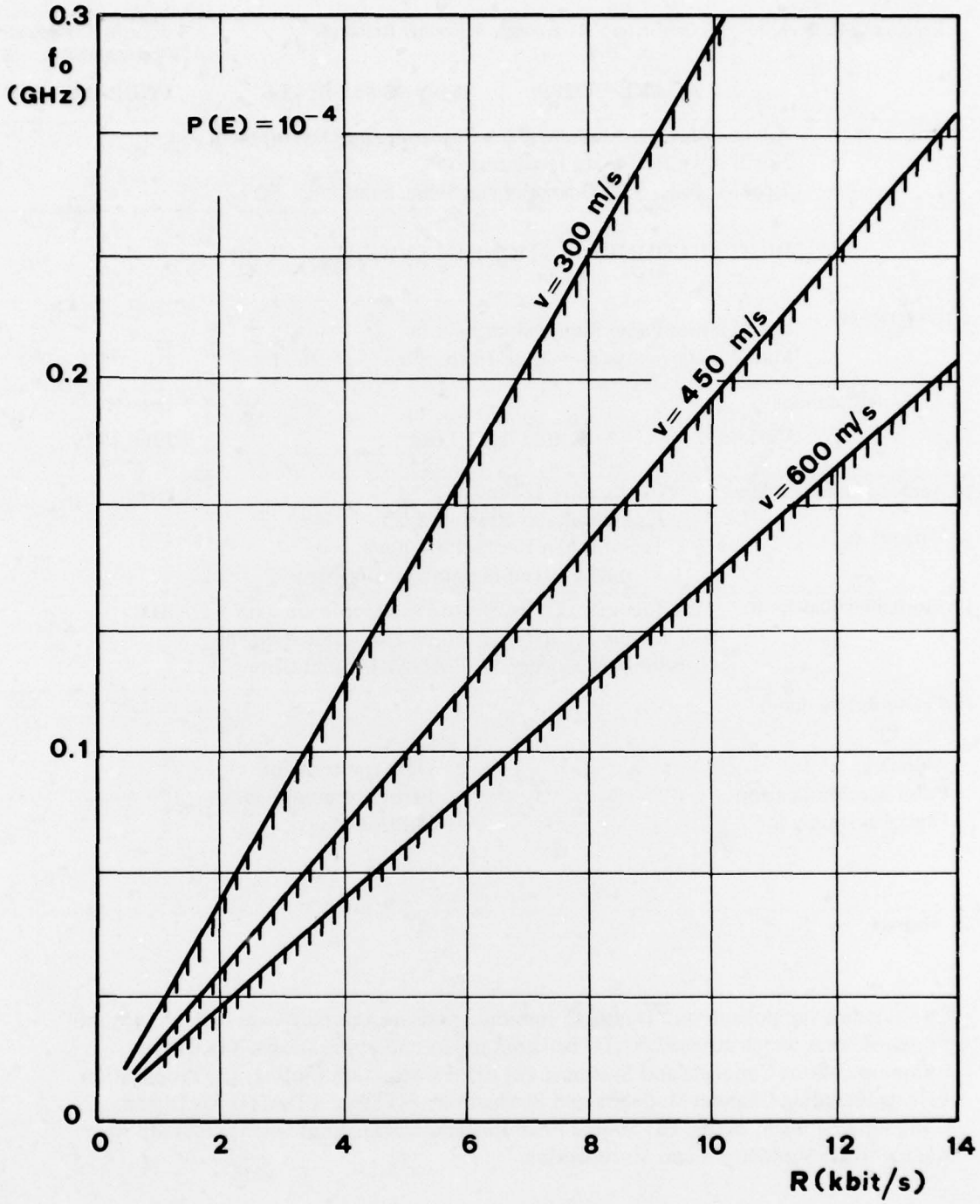


Fig. 7 - Relative performance between DDPSK and DCPSK (M=4)
In the upper regions DDPSK outperforms DCPSK

REPORT DOCUMENTATION PAGE

1. Recipient's Reference	2. Originator's Reference	3. Further Reference	4. Security Classification of Document
	AGARD-CP-239	ISBN 92-835-0242-6	UNCLASSIFIED

5. Originator Advisory Group for Aerospace Research and Development
North Atlantic Treaty Organization
7 rue Ancelle, 92200 Neuilly sur Seine, France

6. Title DIGITAL COMMUNICATIONS IN AVIONICS

7. Presented at the Avionics Panel Symposium held in Munich, Germany, 5–9 June 1978.

8. Author(s)/Editor(s) Various — Edited by H.Lueg*

9. Date June 1979

10. Author's/Editor's Address Various

*Institut für Technische Elektronik der Reine-Westf. Technischen Hochschule Aachen
51 Aachen, Templergraben, Germany

11. Pages 476

12. Distribution Statement This document is distributed in accordance with AGARD policies and regulations, which are outlined on the Outside Back Covers of all AGARD publications.

13. Keywords/Descriptors

Avionics
Pulse communication
Digital systems
Data transmission
Error correction codes
Multiplexing

14. Abstract

The avionics symposium on "Digital Communications in Avionics" was divided into six principal areas which stressed NATO oriented topics and applications. (1) Digital Communications Concepts and Systems; (2) Error Correction Coding; (3) Propagation Effects including Channel Modeling and Simulation; (4) Special Devices for Digital Communications Systems; (5) Source Encoding and Data Compression; (6A) Multiple Access; (6B) Modulation and Multiplexing.

AGARD Conference Proceedings No.239 Advisory Group for Aerospace Research and Development, NATO DIGITAL COMMUNICATIONS IN AVIONICS Edited by H.Lueg Published June 1979 476 pages The avionics symposium on "Digital Communications in Avionics" was divided into six principal areas which stressed NATO oriented topics and applications. (1) Digital Communications Concepts and Systems; (2) Error Correction Coding; (3) Propagation Effects including Channel Modeling and Simulation; (4) Special Devices for Digital Communications Systems; (5) Source Encoding and Data Compression; P.T.O.	AGARD-CP-239 Avionics Pulse communication Digital systems Data transmission Error correction codes Multiplexing	AGARD Conference Proceedings No.239 Advisory Group for Aerospace Research and Development, NATO DIGITAL COMMUNICATIONS IN AVIONICS Edited by H.Lueg Published June 1979 476 pages The avionics symposium on "Digital Communications in Avionics" was divided into six principal areas which stressed NATO oriented topics and applications. (1) Digital Communications Concepts and Systems; (2) Error Correction Coding; (3) Propagation Effects including Channel Modeling and Simulation; (4) Special Devices for Digital Communications Systems; (5) Source Encoding and Data Compression; P.T.O.	AGARD-CP-239 Avionics Pulse communication Digital systems Data transmission Error correction codes Multiplexing
AGARD Conference Proceedings No.239 Advisory Group for Aerospace Research and Development, NATO DIGITAL COMMUNICATIONS IN AVIONICS Edited by H.Lueg Published June 1979 476 pages The avionics symposium on "Digital Communications in Avionics" was divided into six principal areas which stressed NATO oriented topics and applications. (1) Digital Communications Concepts and Systems; (2) Error Correction Coding; (3) Propagation Effects including Channel Modeling and Simulation; (4) Special Devices for Digital Communications Systems; (5) Source Encoding and Data Compression; P.T.O.	AGARD-CP-239 Avionics Pulse communication Digital systems Data transmission Error correction codes Multiplexing	AGARD Conference Proceedings No.239 Advisory Group for Aerospace Research and Development, NATO DIGITAL COMMUNICATIONS IN AVIONICS Edited by H.Lueg Published June 1979 476 pages The avionics symposium on "Digital Communications in Avionics" was divided into six principal areas which stressed NATO oriented topics and applications. (1) Digital Communications Concepts and Systems; (2) Error Correction Coding; (3) Propagation Effects including Channel Modeling and Simulation; (4) Special Devices for Digital Communications Systems; (5) Source Encoding and Data Compression; P.T.O.	AGARD-CP-239 Avionics Pulse communication Digital systems Data transmission Error correction codes Multiplexing

(6A) Multiple Access; (6B) Modulation and Multiplexing.

Copies of papers and discussions presented at the Avionics Panel Symposium held in Munich, Germany, 5–9 June 1978.

ISBN 92-835-0242-6

(6A) Multiple Access; (6B) Modulation and Multiplexing.

Copies of papers and discussions presented at the Avionics Panel Symposium held in Munich, Germany, 5–9 June 1978.

ISBN 92-835-0242-6

(6A) Multiple Access; (6B) Modulation and Multiplexing.

Copies of papers and discussions presented at the Avionics Panel Symposium held in Munich, Germany, 5–9 June 1978.

ISBN 92-835-0242-6

(6A) Multiple Access; (6B) Modulation and Multiplexing.

Copies of papers and discussions presented at the Avionics Panel Symposium held in Munich, Germany, 5–9 June 1978.

ISBN 92-835-0242-6

B362
4

AGARD

NATO OTAN

7 RUE ANCELLE · 92200 NEUILLY-SUR-SEINE

FRANCE

Telephone 745.08.10 · Telex 610176

DISTRIBUTION OF UNCLASSIFIED AGARD PUBLICATIONS

AGARD does NOT hold stocks of AGARD publications at the above address for general distribution. Initial distribution of AGARD publications is made to AGARD Member Nations through the following National Distribution Centres. Further copies are sometimes available from these Centres, but if not may be purchased in Microfiche or Photocopy form from the Purchase Agencies listed below.

NATIONAL DISTRIBUTION CENTRES

BELGIUM
Coordonnateur AGARD – VSL
Etat-Major de la Force Aérienne
Quartier Reine Elisabeth
Rue d'Evere, 1140 Bruxelles

CANADA
Defence Scientific Information Service
Department of National Defence
Ottawa, Ontario K1A OZ2

DENMARK
Danish Defence Research Board
Østerbrogades Kaserne
Copenhagen Ø

FRANCE
O.N.E.R.A. (Direction)
29 Avenue de la Division Leclerc
92 Châtillon sous Bagneux

GERMANY
Zentralstelle für Luft- und Raumfahrt-dokumentation und -information
c/o Fachinformationszentrum Energie, Physik, Mathematik GmbH
Kernforschungszentrum
7514 Eggenstein-Leopoldshafen 2

GREECE
Hellenic Air Force General Staff
Research and Development Directorate
Holargos, Athens, Greece

ICELAND
Director of Aviation
c/o Flugrad
Reykjavik

ITALY
Aeronautica Militare
Ufficio del Delegato Nazionale all'AGARD
3, Piazzale Adenauer
Roma/EUR

LUXEMBOURG
See Belgium

NETHERLANDS
Netherlands Delegation to AGARD
National Aerospace Laboratory, NLR
P.O. Box 126
Delft

NORWAY
Norwegian Defence Research Establishment
Main Library
P.O. Box 25
N-2007 Kjeller

PORTUGAL
Direcção do Serviço de Material
da Forca Aerea
Rua da Escola Politecnica 42
Lisboa
Attn: AGARD National Delegate

TURKEY
Department of Research and Development (ARGE)
Ministry of National Defence, Ankara

UNITED KINGDOM
Defence Research Information Centre
Station Square House
St. Mary Cray
Orpington, Kent BR5 3RE

UNITED STATES
National Aeronautics and Space Administration (NASA)
Langley Field, Virginia 23365
Attn: Report Distribution and Storage Unit

THE UNITED STATES NATIONAL DISTRIBUTION CENTRE (NASA) DOES NOT HOLD STOCKS OF AGARD PUBLICATIONS, AND APPLICATIONS FOR COPIES SHOULD BE MADE DIRECT TO THE NATIONAL TECHNICAL INFORMATION SERVICE (NTIS) AT THE ADDRESS BELOW.

PURCHASE AGENCIES

Microfiche or Photocopy
National Technical
Information Service (NTIS)
5285 Port Royal Road
Springfield
Virginia 22161, USA

Microfiche
Space Documentation Service
European Space Agency
10, rue Mario Nikis
75015 Paris, France

Microfiche
Technology Reports
Centre (DTI)
Station Square House
St. Mary Cray
Orpington, Kent BR5 3RF
England

Requests for microfiche or photocopies of AGARD documents should include the AGARD serial number, title, author or editor, and publication date. Requests to NTIS should include the NASA accession report number. Full bibliographical references and abstracts of AGARD publications are given in the following journals:

Scientific and Technical Aerospace Reports (STAR)
published by NASA Scientific and Technical
Information Facility
Post Office Box 8757
Baltimore/Washington International Airport
Maryland 21240, USA

Government Reports Announcements (GRA)
published by the National Technical
Information Services, Springfield
Virginia 22161, USA

Printed by Technical Editing and Reproduction Ltd
Harford House, 7–9 Charlotte St, London W1P 1HD

ISBN 92-835-0242-6