1962 IRE International Convention Record

PART 8

Sessions Sponsored by

IRE Professional Groups on

Communications Systems

Vehicular Communications

at

the IRE International Convention, New York, N.Y.

March 26-29, 1962

The Institute of Radio Engineers

1962 IRE INTERNATIONAL CONVENTION RECORD

An annual publication devoted to papers presented at the IRE International Convention held in March of each year in New York City. Formerly published under the titles CONVENTION RECORD OF THE I.R.E. (1953 & 1954), IRE CONVENTION RECORD (1955 & 1956), and IRE NATIONAL CONVENTION RECORD (1957, 1958, & 1959).

Additional copies of the 1962 IRE INTERNATIONAL CONVENTION RECORD may be purchased from the Institute of Radio Engineers, 1 East 79 Street, New York 21, N.Y., at the prices listed below.

Part	Sessions	Subject and Sponsoring IRE Professional Group	Prices for Members of Sponsoring Professional Group (PG), IRE Members (M), Libraries and Sub. Agencies (L), and Nonmembers (NM)			
			PG	M	L	NM
1	8, 16, 23	Antennas & Propagation	$ .70	$ 1.05	$ 2.80	$ 3.50
2	10, 18, 26, 41, 48	Automatic Control Circuit Theory	1.00	1.50	4.00	5.00
3	1, 9, 17, 25, 28, 33	Electron Devices Microwave Theory & Techniques	1.00	1.50	4.00	5.00
4	4, 12, 20, 34, 49	Electronic Computers Information Theory	1.00	1.50	4.00	5.00
5	5, 13, 15, 22, 29, 47, 54	Aerospace & Navigational Electronics Military Electronics Radio Frequency Interference Space Electronics & Telemetry	1.20	1.80	4.80	6.00
6	3, 11, 31, 35, 42, 45, 50, 52	Component Parts Industrial Electronics Product Engineering & Production Reliability & Quality Control Ultrasonics Engineering	1.40	2.10	5.60	7.00
7	30, 37, 43, 51	Audio Broadcasting Broadcast & Television Receivers	.80	1.20	3.20	4.00
8	7, 24, 38, 46, 53	Communications Systems Vehicular Communications	1.00	1.50	4.00	5.00
9	2, 19, 27, 32, 39, 40, 44	Bio-Medical Electronics Human Factors in Electronics Instrumentation Nuclear Science	1.20	1.80	4.80	6.00
10	6, 14, 21, 36	Education Engineering Management Engineering Writing & Speech	.80	1.20	3.20	4.00
		Complete Set (10 Parts)	$10.10	$15.15	$40.40	$50.50

Responsibility for the contents of papers published in the IRE INTERNATIONAL CONVENTION RECORD rests solely upon the authors and not upon the IRE or its members.

1962 IRE INTERNATIONAL CONVENTION RECORD

PART 8 - COMMUNICATIONS SYSTEMS; VEHICULAR COMMUNICATIONS

TABLE OF CONTENTS

Special Communications Techniques
(Session 53: sponsored by PGCS)

SIMULTANEOUS TRANSMISSION AND RECEPTION WITH A COMMON ANTENNA

J.V. Tilston
Sinclair Radio Laboratories Ltd.
Downsview, Ontario, Canada.

The problems involved in operating two or more equipments on one antenna are discussed and some methods are given for their solution. These involve the use of various types of linear passive filters.

The particular type of filter to be used depends on the number of transmitters and receivers to be coupled, on the frequency separation between them, and whether the frequencies are to be fixed or variable.

Practical information is given on how the various types of filters, may be constructed and used in the different frequency bands from 2 Mc. to 10,000 Mc.

Introduction

Whenever transmitters and receivers are operated simultaneously, it is necessary to provide a certain amount of isolation between them.[1,2,3] The most obvious method of providing this isolation is to use a separate antenna for each equipment.

For antennas which are spaced far apart, more than enough isolation can be obtained in this way, so that no interference problem arises.

As the number of systems in a given area steadily increases, the distance between the various antenna sites decreases, thus reducing the isolation between adjacent stations. In addition, the frequency spectrum becomes more crowded. Both these factors contribute to a rise in interference. This interference may again be restored to a non-objectionable level by the insertion of appropriate filters into the antenna feed lines. However, it is often more economical to connect two or more equipments to a common antenna by means of a diplexer or multicoupler. This arrangement can often be made to provide the required isolation and to eliminate the need for multiple antennas and cable runs.

In addition to the problem of interference which is encountered in multiple-antenna installations, there is the problem of pattern distortion. This is particularly serious in many shipborne applications and it was primarily from pattern considerations that some of the earlier multicouplers were developed[12]. A single favourable antenna location could then be made to serve several transmitters and receivers operating on different frequencies. Similarly, at repeater sites, a diplexer makes possible the use of a single antenna located at the top of a tower for both reception and transmission.

When a common antenna is to be used it is first necessary to establish the amount of isolation needed, over the band of frequencies to be used. This may be done by examining the properties of the transmitters and receivers.[1]

Once the isolation-frequency relationship has been established, the next consideration is the manner in which the isolation is to be achieved. Here several possibilities exist, and of these, only those containing solely passive linear networks will be examined.

Determination of the Isolation-Frequency Relationship when a Transmitter on Frequency f^1 and a Receiver on Frequency f^2, are to be connected to a Common Antenna.

Let f_1 = frequency to which the transmitter is tuned.
Let f_2 = frequency to which the receiver is tuned.
$V_R (f_2)$ = amplitude of the interfering voltage which the receiver can withstand without disensitizing, at its resonant frequency f_2.
$V_T (f_1)$ = the amplitude of the output voltage from the transmitter at its resonant frequency f_1.
$L_T(f_1, f - f_1)$ = the number of db by which the transmitter output, at frequency f, is below the transmitter carrier at frequency f_1
$L_R(f_2, f - f_2)$ = the number of db above the reference voltage on frequency f_2, which an interfering signal, on frequency f, may attain without densensitizing the receiver.
$N (f)$ = the required isolation, in db, between transmitter and receiver, at the frequency f.

It can be shown that:

$$N(f) = 20 \log_{10} \frac{V_T(f_1)}{V_R(f_2)} - L_T(f_1, f-f_1) - L_R(f_2, f-f_2)$$

Curves of L_T and L_R can, in many instances, be obtained from the manufacturer. In most cases, the curves supplied are for a typical transmitter or receiver, and some safety factor must be applied to allow for variations between sets and for such things as tube aging. Additional isolation must also be provided at those points where intermodulation products are a problem.

A typical transmitter spectrum is shown in figure 1.. This would give L_T.

Figure 2 shows a typical receiver desensitizing voltage curve and so gives L_R. Values of $V_T(f_1)$, and $V_R(f_2)$, are, of course, supplied with the equipment data.

In general it is found that the required isolation X frequency curve has two frequencies, namely the transmitter frequency and the receiver frequency, where the isolation must be much greater than at all other frequencies. This is because the main interference problems arise from the transmitter carrier, and from noise, or spurious radiation from the transmitter, on the receiver frequency.[1,14]

Figure 3, illustrates how the necessary attenuation at these two frequencies varies as the frequency separation between transmitter and receiver is varied. While these curves are applicable to many V.H.F. installations the same type of curves may be obtained for other frequency ranges. It should be noticed that for fairly wide frequency separations the main cause of trouble is the transmitter carrier, so that more attenuation is required at this frequency than at the receiver frequency. As the frequency separation becomes narrower, a point is reached where the same amount of attenuation is required at the receiver frequency as at the transmitter frequency.

At even closer spacings the receiver frequency requires more attenuation than does the transmitter frequency. As an example of this latter case, figure 4 illustrates the required isolation between a V.H.F. receiver on frequency f_o, and a transmitter tuned 180 Kc. away from f_o.

It should be stressed, however that in any given installation, spurious responses could alter this picture somewhat. In addition, from the example in figure 4, it would seem that no isolation is required at frequency $f_o + 0.090$, or mid-way between the transmitter and receiver frequencies. This is because intermodulation has not been taken into account.

A more accurate estimate can be made if the intermodulation patterns of the transmitter and receiver are measured. If this is done, interference may be seen to result on the receiver frequency due to a product formed by the transmitter carrier and transmitter noise near $f_o + 0.090$. Additional isolation will probably also be needed to take care of intermodulation products which could be formed in such non-linear devices as imperfect antenna joints. In practice, for a 0.4 percent frequency separation in the 150 Mc. band, a minimum isolation of about 30 db should be furnished everywhere in the band.

Throughout the present discussion it has been assumed that the transmitter and receiver are properly aligned. It is sometimes stipulated in specifications that no densensitizing is to take place under any conditions of transmitter tuning. Since varying the tuning controls on most transmitters has a marked effect on the noise output, much more isolation must be provided in this case. In fact in the example given the isolation at the receiver frequency may well need to be over 100 db instead of 75 db.

Methods of Obtaining the Necessary Isolation.

Once the amount of isolation over the band has been determined, the next step is to choose the best method for achieving it. Some of these methods are:

Separate Antennas.

One way to provide the necessary isolation is to use separate antennas for each frequency. This method often proves undesirable from the standpoint of both cost and size.

Common Antenna with Diplexer

In the diplexer method of operation, the transmission lines from the transmitter and receiver pass through their respective filtering sections and are then connected to a common antenna terminal as in figure 5.

The number of filters in each arm depends on the filter type used and on the frequency separation.

The system must be designed so that the transmit frequency is passed with little attenuation from the Tx terminal to the antenna terminal, while at the same time being sufficiently isolated from the receive terminal. The receive frequency must pass freely from the antenna terminal to the receive terminal.

In order to accomplish this end a large variety of filter types may be employed. These include low-pass, high-pass, band-pass, band-reject, double-tuned band-reject and hybrid ring types.

Filter Types used in Diplexers

High-Pass and Low-Pass Filters

A diplexer may be made with high-pass filters on one side of the antenna terminal and low-pass filters on the other.[4,5] In this system there are two broad useable frequency bands separated by a relatively narrow crossover region. The use of complimentary filters leads to a good impedance match.

Diplexers of this type can be readily designed in the various frequency bands from 100 Mc. to 5000 Mc. by the use of fairly simple image-parameter methods. An advantage of this type of system is that within wide bands of operation, either the transmitter frequency or the receiver frequency can be changed without any change being made in the diplexer. In general the crossover region is not as narrow, for the same resonant insertion loss, as can be achieved with band-reject filters, nor is the system as flexible as some of those comprising band-pass filters.

Band-Pass Filters

A common way to construct a diplexer is to make the filters of the band-pass type.[6,7,8] One advantage of using the band-pass filter is the ease with which the system can be designed,[9] due to the fact that there exist universal formulae which are applicable regardless of the physical configuration used. This point is brought out in figure 6. Here, three different types of band-pass filters are analized. It should be noticed that the formulae dealing with insertion loss on resonance, and with the frequency response of the filter, are in terms of the unloaded Q or Q_o, and the loaded Q or Q_L, and these formulae are the same for all band-pass filters. These formulae are however approximations, but it is found that they can often be applied, with fairly good accuracy, even over bandwidths as wide as 2:1.

Once the desired values of Q_o and Q_L have been established, it can be decided what physical form the filter will take. Usually, lumped elements are used in the MF and HF ranges, while coaxial elements are used at V.H.F., U.H.F. and L band. Occasionally coaxial band-pass filters are used at much higher frequencies, although more often waveguide sections are used in the frequency bands where the waveguide sizes prove to be convenient.

Upon examing band-pass responses, it becomes evident that they fundamentally pass one frequency, while rejecting all others, except for any secondary pass bands. In addition as the frequencies become further and further removed from the resonant frequency of the filter, they become attenuated more and more.

This latter feature might, in some cases, be desirable, but it should be remembered that in the example cited earlier it was not necessary.

Moreover, it is a feature, that is paid for with a higher insertion loss than could be obtained with other filter types.

It is true that by the use of modern network theory a more nearly optimum stop-band response may be obtained but this often entails the use of several band-pass filters in cascade and may still not be as economical as using fewer filters of a different type.

Band-Reject Filters

As can be seen from figure 7, band-reject filters also possess universal insertion loss, and response formulae, so that ease of design is a feature here too. Again the attenuation loss on resonance, and the frequency response may be obtained by properly choosing Q_o and Q_L.

The band-reject filter attenuates one frequency by a given amount, and attenuates all others by a lesser amount. The further removed a frequency is from resonance, the less it is attenuated. The response of the band-reject filter closely resembles that which is needed to give the type of frequency-isolation curve of the previous example.

It has an advantage over the band-pass filter in many applications, since for the same size, and the same insertion loss for the frequency to be passed, it offers much more attenuation to a single frequency to be stopped, than does a band-pass filter. The single band-reject filter has the disadvantage, when used in diplexers, that if the insertion loss of the frequency to be passed is to be kept under a given limit, then the loaded Q of the filter must be changed when the frequency separation is changed.

Two Section Band-Reject Filters

One of the most useful, and easily constructed filters for diplexer use is the two-section Band-Reject filter, whose behaviour is described in figure 8. In this case also, it has been found that universal formulae apply, although a quantity σ has been defined instead of the loaded Q. The behaviour, although a little more complicated than either of the two former examples, is still reasonably easy to predict.

To pass one frequency and reject another, with this type of filter, it is merely necessary to tune one section for a null at the frequency to be stopped, and to tune the other section so that the frequency to be passed is attenuated as little as possible. In practice there is usually some interaction between these two sections and so the tuning procedure may have to be repeated two or three times. The behaviour of this type of section also has the type of response which was found to be desirable previously. At MF and HF frequencies the filter may be constructed of lumped constant elements.

At V.H.F. and U.H.F. the resonant sections are readily formed from coaxial, open-circuited, quarter-wave sections which are placed in parallel with the main transmission line. At still higher frequencies wave-guide sections may be used, the resonant frequency being changed by means of posts of variable length which may be inserted into the guide walls.

Comparison of Band-Pass, Band-Reject and Two-Section Band-Reject Filters.

In figure 9, the performance of the band-pass, band-reject and two-section band-reject filters are compared. Each filter gives an insertion loss of 1 db at frequency f_o. The unloaded Q or Q_o of the band-pass filter was made equal to that of the band-reject filter, and the volume of the two-section band-reject filter was made equal to the volume of the band-pass filter. These curves illustrate many of the points mentioned earlier.

An isolation-frequency curve for a diplexer containing two, two-section, band-reject filters in the receive arm, and one in the transmit arm is shown in figure 10. This type of response has proven quite satisfactory for frequency separations of about three per-cent or more.

Hybrid Ring Filters

When the frequency-separation between transmitter and receiver becomes in the order of one per-cent or less, the type of frequency response given in figure 10 turns out to be inadequate. At first glance, this seems unusual since it approximates the shape of curve calculated in figure 4 which was for a frequency separation of only 0.1 per cent.

The trouble is that in practice there is usually more noise in the system than was anticipated and also a certain amount of rectification occurring in joints and connectors, and this noise becomes more and more objectionable as the frequency separation decreases. In fact as was mentioned earlier for a 0.4 per cent separation it is desirable to drop the whole isolation curve, such as that of figure 4 or figure 10, by about thirty decibels so that at least this amount of isolation is present at any frequency within about plus or minus ten per cent of those actually being used to transmit or receive. This does not mean that band-pass filters need be used in the diplexer since it is not desirable to have their relatively high resonant insertion loss if it can be avoided.

What is needed is a minimum of perhaps thirty or forty db everywhere, plus the calculated isolation at the transmit and receive frequencies. Thus it would seem desirable to somehow combine the good features of both the band-pass and band-reject types.

One type of isolating network, not yet considered here is the hybrid ring (or magic T in waveguide).. This circuit will theoretically isolate a transmitter from a receiver by an infinite amount provided a three db insertion loss can be tolerated. A more easily attainable isolation figure is forty db since the loads are seldom perfectly matched and the ring is not perfectly balanced.

The arrangement of figure 11 is one way in which a diplexer arm may be constructed using a hybrid ring and a band-reject filter. The lines here represent coaxial cables. If the parallel stub and band-reject filter were removed from the circuit then very little transmitted signal would reach the antenna, since the two possible paths to the antenna terminal differ in length by a half-wave.

When the band-reject filter is installed as shown, and tuned to the transmitter frequency, a short-circuit is effectively placed across the ring at a point a quarter-wavelength removed from the transmitter junction. Thus the transmitter signal travels clockwise to the antenna terminal with little attenuation.

At other frequencies it is desirable for signals from the transmit terminal, travelling clockwise, to arrive at the antenna terminal with the same amplitude, but opposite phase, to those travelling anti-clockwise. This is the reason for the parallel stub which is really a compensating network to balance the ring at the receiver frequency.

A similar ring is used in the receive arm of the diplexer.

Figure 12 illustrates the desirable type of response which is obtained from a diplexer which used hybrid ring elements.

Multicouplers

When it is desired to combine more than two equipments on one common antenna a multicoupler is used. As with diplexers, a large variety of filters may be employed. One of the earliest and most successful designs for V.H.F. and U.H.F. use consists of a number of parallel band-pass channels.[11,12,13] A modified version of such a multicoupler is shown in figure 13[10]. Equipments T_1, T_2,T_3,T_4 are connected through their respective band-pass filters F_1,F_1 etc. to a common antenna junction. Line B is a matching section which compensates for the varying input impedance of the filters over the band of frequencies used. Early parallel type multicouplers used only one type of resonator filter such as F_1,F_2,F_3,F_4. Thus the number of channels that could be accommodated in this scheme depended to a large extent on how many resonators could surround the antenna junction with quarter wave coupling lines L_1,L_2,L_3,L_4. As the desired frequency separation became smaller, the isolation required between channels became larger, so that larger and larger band-pass filters were needed.

However, if the problem of switching the various frequencies into the proper channels is divorced from that of supplying enough isolation between channels, then it is possible to use smaller resonators, designated in figure 13 by P_1,P_2,P_3,P_4, to switch the frequencies, and larges ones, F_1,F_2,F_3,F_4 to supply most of the isolation needed. Even so, the system is limited by the number of switching filters P_1,P_2,P_3,P_4 that can be fitted.

Besides the disadvantage of the limited number of channels, the system described above has the disadvantage that it must be designed for a specified number of channels which cannot be changed without a major change in the matching. Thus channels cannot be added or subtracted at will.

Figure 14 shows a system in which channels can be added or subtracted. Equipments T_1 to T_{40} may be transmitters or receivers. The lower part of this figure shows the "through" transmission line terminated in a parallel multicoupler. In many cases this lower section would be replaced by a dummy load or by a short circuit. Thus only the action of the upper four channels will be discussed here, and, these can be understood by considering the first channel which is shown surrounded by dashed line.

If equipment T_{10} is tuned to frequency f_{10}, then there is a path of very low attenuation at f_{10} through filter F_{10}. On the other hand filter W_{10} stops frequency f_{10}, but lets all other frequencies pass with little attenuation. At the same time filter F_{10}, stops all other frequencies but f_{10}. Nine channel tuneable multicouplers based on this scheme have been built in the 225-400 Mc. range and twelve channel units have been designed in the 150 Mc. communications band. These systems have proven to be much more flexible than previous systems and seem to be limited mainly by the losses and power handling capacity of the "through" line as the number of channels is increased.

Conclusions

An outline has been given of various ways in which two or more equipments may be connected to a common antenna.

The particular type used in any given

instance will depend on several factors such as frequency separation, power level, size limitations and allowable insertion loss. In many cases it is evident that band-pass filters do not supply the best answer to the diplexing problem. On the other hand for fixed-tuned two-channel systems the two-section band-reject and the hybrid-ring diplexers, offer a reasonable solution by providing the required isolation with low insertion loss.

In multicouplers, however, band-pass filters, sometimes in conjunction with band-reject filters, can be used to advantage in systems which are not only tuneable, but are also flexible from the standpoint of the number of channels needed.

Bibliography

1. N.H. Shepherd "Bi-Dimensional Interference Analysis". Third Conference on Radio Interference Reduction, Armour Research Foundation of Illinois Institute of Technology. February 26, 1957.

2. J.F. Byrne "The Selectivity and Inter modulation Problem in U.H.F. and Communication Equipment". I.R.E. Transactions on Airborne Electronics. September 1952.

3. S.F. Meyer "Multi Transmitter-Receiver Installation Interference - Problems and Cures". Communications Horizons, December 1961.

4. M.E. Breese "Diplexing Filters", I.R.E. Convention Record, Vol. 2, Part 8, Communications and Microwave, 1954.

5. "Final Report on Band Separation Filters", Sperry Gyroscope Co., Sperry Report No. 5224-1266-8.

6. T.H. Moriarty "Resonant Ring Diplexers in Forward Scatter Systems" Electronics, Vol. 32, No. 27, July 3, 1959.

7. H.J. Carlin "U.H.F. Multiplexer Uses Selective Couplers", Electronics, Vol. 28, No. 11, November 1955.

8. J.F. Cline and B.M. Schiffman "Tuneable Passive Multicouplers Employing Minimum-Loss Filters" I.R.E. Transactions on Microwave Theory and Techniques. Vol. M.T.T-7, No. 1, January 1959.

9. P.I. Richards "Universal Optimum - response curves for Arbitrarily Coupled Resonators". Proc. of the I.R.E., Vol. 34, 1946.

10. W.V. Tilston "New Methods of Coupling Filters to allow a number of Transmitters and/or Receivers to be Connected to the same Antenna." Sinclair Radio Laboratories Report, February 1, 1957.

11. R.E. Canfield "Final Development Report on Mutual Environmental Tests of U.H.F. Equipment Employing Selective Filters." Phase 1, Item 6-16, April 1956. Collins Radio Co., Cedar Rapids, Michigan.

12. A.G.D. Watson, J.H. Jones, D.L. Owen, "Common Aerial Working for V.H.F. Communications". J.I.E.E. Part III A, Vol. 94, No. 14, March-April 1947.

13. "Nontactical Radio Equipment System Study", Contract AF30(602)-1251, Item 5, Final Engineering Report, July 1, 1955", Motorola Inc., National Defense Division, Chicago 14, Illinois

14. N.H. Shepherd and A.C. Giesselman "Interference - A Look at the Ounce of Prevention", I.R.E. Transactions on Vehicular Communications, Vol. 13, September 1959.

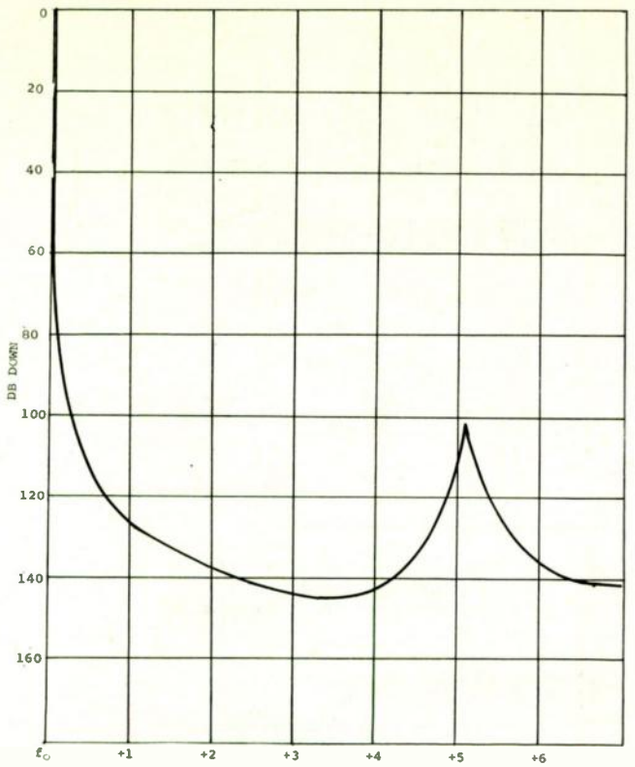


Fig. 1.

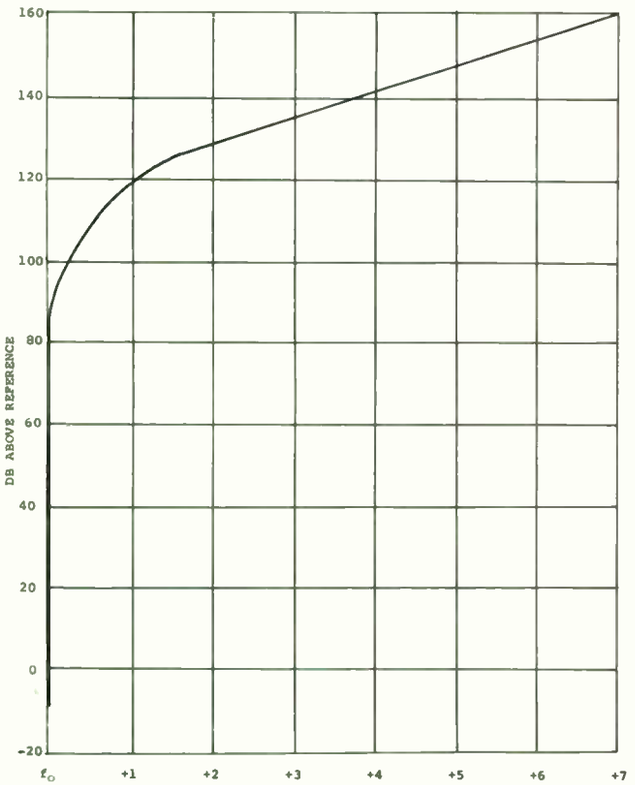


Fig. 2.

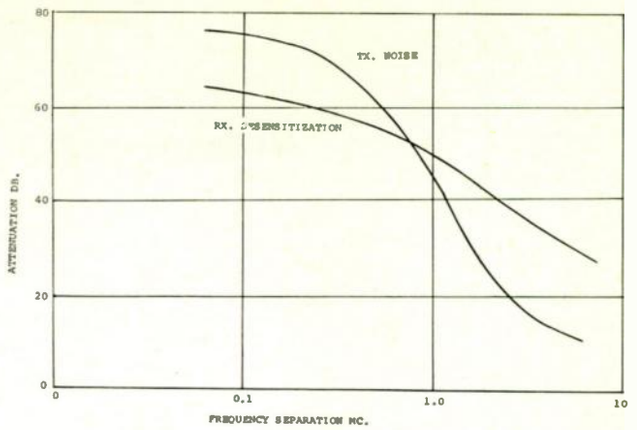


Fig. 3.

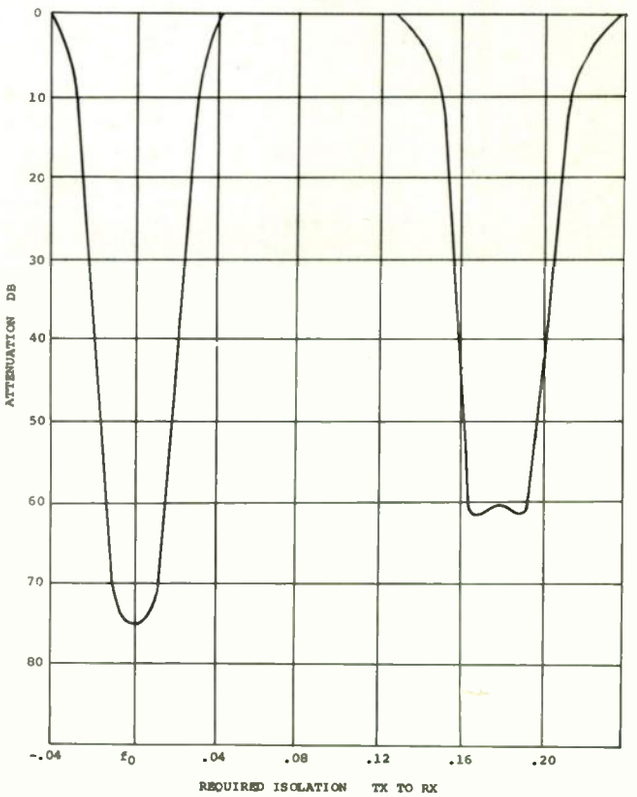


Fig. 4.

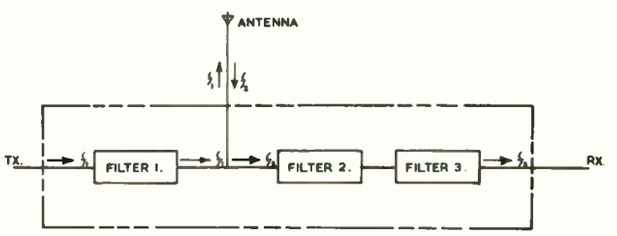
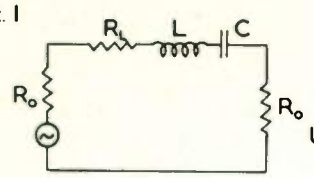


Fig. 5.

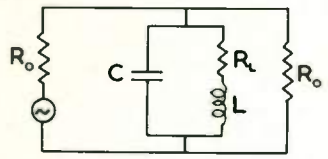
Ex. 1



LOADED $Q = Q_L = \frac{\omega_0 L}{R_o}$

UNLOADED $Q = Q_o = \frac{\omega_0 L}{R_L}$

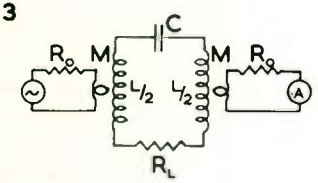
Ex. 2



$Q_L = \frac{R_o}{2\omega_0 L}$

$Q_o = \frac{\omega_0 L}{R_L}$

Ex. 3



$Q_L = \frac{R_o \omega_0 L}{2\omega_0^2 M^2}$

$Q_o = \frac{\omega_0 L}{R_L}$

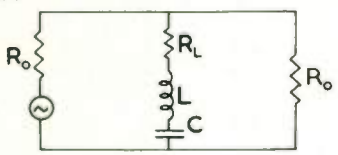
IN GENERAL.

RESONANT INSERTION LOSS $= 20 \log_{10}\left(1 + \frac{Q_L}{Q_o}\right)$

$$\frac{\text{POWER INTO LOAD AT FREQ. } f_0}{\text{POWER INTO LOAD AT FREQ. } f} = \frac{P_0}{P_L} = 1 + Q_L^2\left(\frac{f}{f_0} - \frac{f_0}{f}\right)^2$$

Fig. 6.

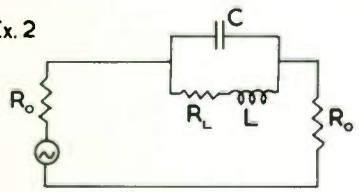
Ex. 1



$Q_L = \frac{2\omega_0 L}{R_o}$

$Q_o = \frac{\omega_0 L}{R_L}$

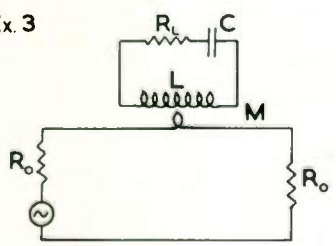
Ex. 2



$Q_L = \frac{2R_o}{\omega_0 L}$

$Q_o = \frac{\omega_0 L}{R_L}$

Ex. 3



$Q_L = \frac{2R_o \omega_0 L}{\omega_0^2 M^2}$

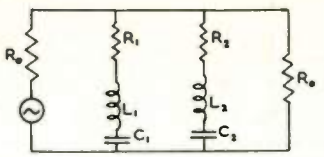
$Q_o = \frac{\omega_0 L}{R_L}$

INSERTION LOSS AT RESONANCE $= 20 \log_{10}\left(1 + \frac{Q_0}{Q_L}\right)$

$$\frac{P_0}{P_L} = 1 + \frac{1}{Q_L^2\left(\frac{f}{f_0} - \frac{f_0}{f}\right)^2}$$

Fig. 7.

EX. 1



L_1, C_1 TUNED TO f_1.

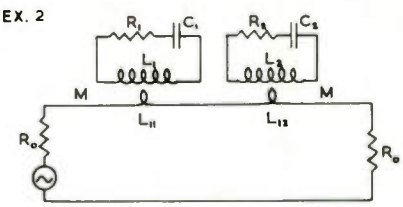
L_2, C_2 TUNED TO f_2.

$Q_o = \frac{\omega_0 L_1}{R_1} = \frac{\omega_0 L_2}{R_2}$

$$\sigma = \frac{\text{VOLTAGE ACROSS LOAD AT } f_0 \text{ WITHOUT FILTERS}}{\text{VOLTAGE ACROSS LOAD AT } f_0 \text{ WITH FILTERS}}$$

$$= 1 + \frac{R_o}{R_L} \frac{1}{1 + 4Q_o^2\left(\frac{f_0 - f_1}{f_0}\right)^2} \qquad R_L = \frac{R_1 + R_2}{2}.$$

EX. 2



L_1, C_1 TUNED TO f_1.

L_2, C_2 TUNED TO f_2.

$Q_o = \frac{\omega_0 L_1}{R_1} = \frac{\omega_0 L_2}{R_2}$

$$\sigma = 1 + \frac{\omega_0^2 M^2}{2R_o}\left[\frac{1}{R_1} + \frac{1}{R_2}\right] \frac{1}{1 + 4Q_o^2\left(\frac{f_0 - f_1}{f_0}\right)^2}$$

$$\frac{\text{VOLTAGE ACROSS LOAD AT } f_0}{\text{VOLTAGE ACROSS LOAD AT } f} = \frac{V_0}{V_L} =$$

$$= 1 + (\sigma - 1)\frac{\left[1 + 4Q_o^2\left(\frac{\Delta\omega}{\omega_0}\right)^2\right]\left[1 + j2Q_o\frac{\Delta\omega}{\omega_0}\right]}{\left[1 + j2Q_o\frac{\Delta\omega + \Delta\omega_1}{\omega_0}\right]\left[1 + j2Q_o\frac{\Delta\omega - \Delta\omega_1}{\omega_0}\right]}.$$

$\Delta\omega_1 = \omega_0 - \omega_1$. $\quad \Delta\omega = \omega - \omega_0$.

Fig. 8.

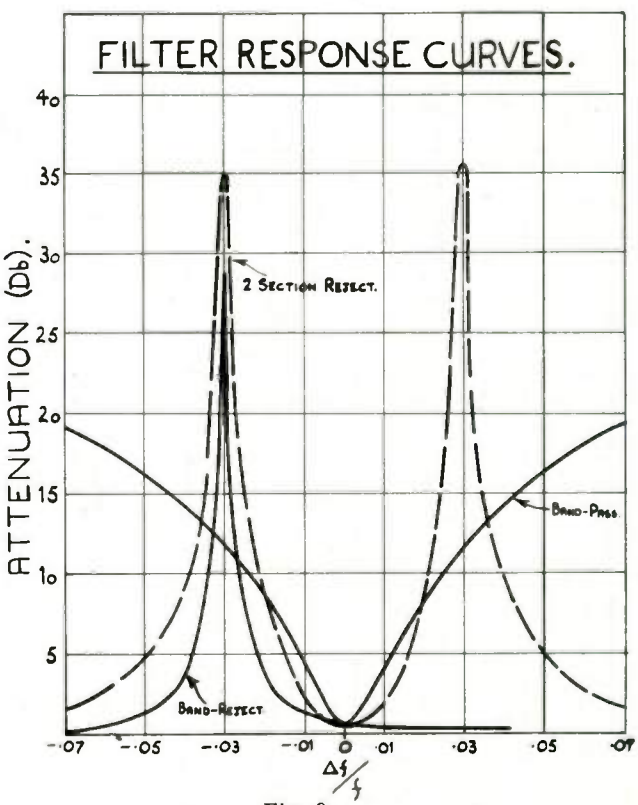


Fig. 9.

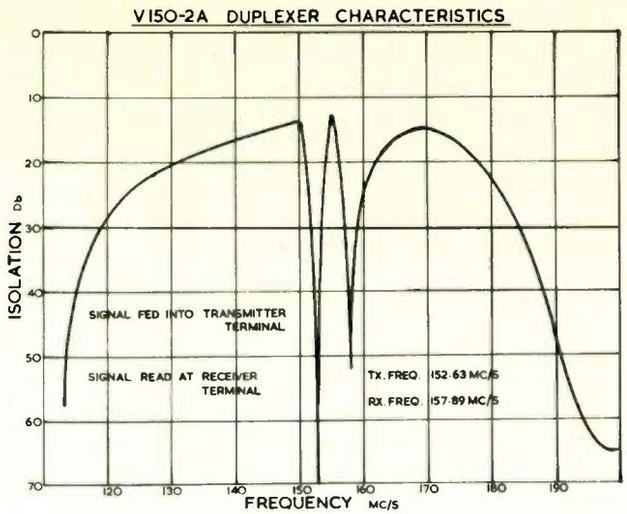


Fig. 10.

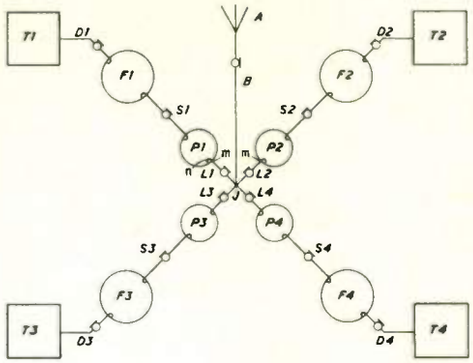


Fig. 13.

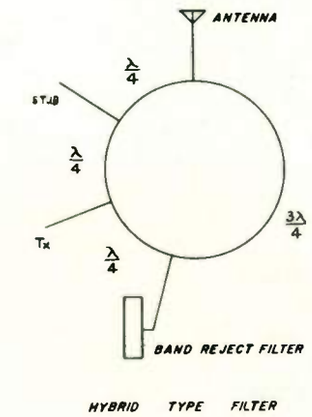


Fig. 11.

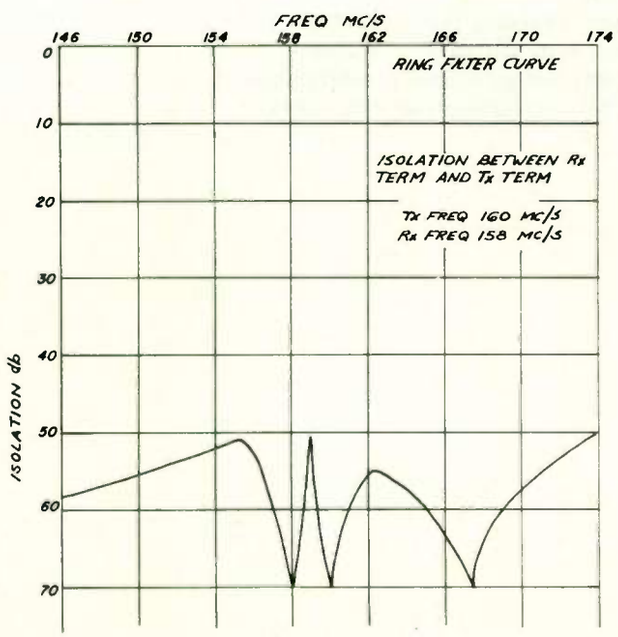


Fig. 12.

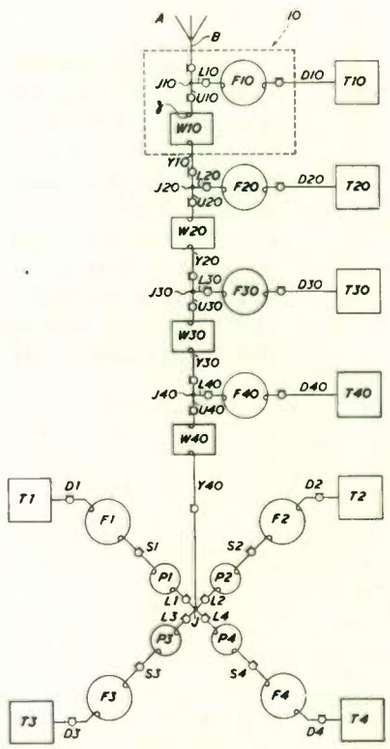


Fig. 14.

AN IMPROVED FM OSCILLATOR-MODULATOR

A. R. Cumming
General Electric Company
Lynchburg, Va.

Abstract

This paper presents the design considerations of an oscillator modulator particularly suitable for use in radio transmitters intended for the Personal Communications. The design under discussion derives a frequency modulated signal by varying the capacitance of a non-linear diode in the crystal oscillator circuit with the input audio signal. The main design aspects are the stability of the crystal oscillator and obtaining the maximum deviation without exceeding the maximum allowable distortion of the output waveform. A theoretical and practical investigation of the optimum operating point of the crystal oscillator to meet the requirement is presented.

AN EXPERIMENTAL FAST ACTING AGC CIRCUIT

A. L. Hopper
Bell Telephone Laboratories, Incorporated
Murray Hill, New Jersey

Summary

Conventional backward-acting AGC circuits acting fast enough to counteract rapid fading will, if stable, suppress neighboring voice frequencies. However, forward-acting AGC is inherently stable, so it can be made very fast by inserting a sharp cut-off filter in the control path. This permits the suppression of 200 cps. fading with negligible effect on the 300-3,000 cps. voice band. Analysis and tests show that fading modulation products are suppressed by about 20 db.

Introduction

Mobile telephone systems of the future will probably require hundreds of channels operating in the UHF band. Bandwidth will have to be conserved, and if this requires the use of AM, rapid fading will pose an AGC problem.[1,2,3] For example, at 850 mc, vehicle speeds of 60 miles per hour produce fading rates of more than 100 cycles per second. This rate cannot be suppressed by conventional (i.e., backward-acting) AGC without significantly cutting into the voice band, because the filter in the feedback path must have a gradual cutoff to avoid regeneration.[4]

The conventional AGC problem is illustrated in Fig. 1. The incoming r-f wave is modulated at a voice frequency of 300 cps. and also a fading frequency of 100 cps. The AGC must suppress the modulation at the fading, but not the signal frequency. Some typical suppression plots are shown in Fig. 1. In (a) it will be seen that only 10 db suppression is provided at 100 cps. In (b) this suppression is increased to about 22.5 db, but the voice band is seriously affected. It is apparent that other techniques are needed to suppress rapid fading. We will therefore consider the use of forward-acting AGC.

Operation of Forward-Acting AGC (FAAGC)

The received i-f wave modulated at 100 and 300 cps. is fed to an envelope detector whose output is shown in Fig. 2(a). The signal modulation is removed by the low-pass filter (LPF) which delays the control wave as shown at (b). This delay must be matched in the signal path by a band-pass filter (BPF). Resulting signal (c) is divided by control (b) to produce wave (d) containing the signal but not the fading modulation.

This circuit will suppress 100 cps. fading by at least 30 db with negligible effect at 300 cps. Moreover, it will suppress fading and signal intermodulation products, which fall in the voice band of frequencies, by about 20 db.

Because of its limited range of control, forward-acting AGC should be used in combination with conventional AGC, which should precede it in the circuit. Thus conventional AGC will accommodate the large slow variations in input level, leaving only the rapid fading to be suppressed by FAAGC.

Low-Pass Filter Characteristics

The LPF should have a sharp cutoff between 200 and 300 cps., but the phase shift in the passband should be as small and as linear as possible since it must be matched by the BPF. If we were to use a conventional LPF, we would find that the slope of the phase characteristic corresponded to a delay of some three milliseconds which would probably be impractical to match. However, by permitting some transmission in the region above cutoff, the phase shift can be reduced by a factor of three and a very good delay match achieved up to at least 100 cps. The measured transmission and phase characteristics of such a special LPF are shown in Fig. 3. The phase is quite linear up to 200 cps. where the transmission is down only 3.5 db. At 300 cps. it is down about 14 db, a cutoff rate of about 18 db per octave. In the region above cutoff, the transmission is down 17.5 db.

The special LPF was designed as a five-section maximally flat filter approximated by two semi-infinite slopes[5] having corner frequencies at 200 and 300 cps. as shown in Fig. 4(a). For this filter the angles of the poles and zeros are equal and at 0°, 36° and 72°, respectively. To permit realizing an

unbalanced bridge structure, the largest angle was reduced to 70° resulting in the configuration shown in Fig. 4(b).

The filter was synthesized in the form of three constant resistance bridged-T networks connected in cascade. This synthesis results in an "all-pass" loss of 4.5 db (in addition to that shown in Fig. 3).

Matching Delay With Band-Pass Filter

Although FAAGC may be used to advantage with any type of AM, and possibly FM too, we will assume in the following analysis that SSB modulation is used to conserve bandwidth. To simplify the analysis, it is assumed that only two modulations are present at a time, a sine wave signal modulation with index m_s and a modulation due to fading with index m_f. It is further assumed that m_s and m_f are small compared to unity.

As shown in Fig. 2(a) and (b), the output of the envelope detector is delayed in passing through the LPF. The amount of this delay, assuming sine wave modulation since m_f is small, is the phase delay of the filter φ/f. For 200 cps. this is the slope of the dashed line in Fig. 3.

In the signal path the BPF should provide the same delay. In this case, however, we assume an upper SSB signal with a single sine wave modulation. The envelope delay of such a signal is:

$$\frac{\varphi_{sideband} - \varphi_{carrier}}{f_{sideband} - f_{carrier}} = \frac{\Delta\varphi}{\Delta f}$$

With the foregoing assumptions, we may normalize the carrier frequency phase in the band-pass (signal) circuit to the zero frequency phase in the low-pass (control) circuit. This is shown in Fig. 5. After this normalization, the envelope delay $\frac{\Delta\varphi}{\Delta f} = \frac{\varphi_{BPF}}{f}$ may be compared directly to phase delay $\frac{\varphi_{LPF}}{f}$. The phase and delay errors are then $(\varphi_{LPF} - \varphi_{BPF})$ and $\frac{\varphi_{LPF} - \varphi_{BPF}}{f}$, respectively.

In addition to matching the LPF delay, the BPF should provide "flat" transmission in the voice band, otherwise sideband amplitude will vary with frequency. To obtain sufficient delay (approximately one millisecond) at i-f, a pair of Collins Radio 455 kc mechanical filters were connected in cascade.[6]

Envelope Detector and Variolosser

The envelope detector uses a diode operating at high voltage levels to provide linear detection. The result is a nearly ideal transfer characteristic over a range of output between 0.1 and 10.0 volts. The variolosser uses a forward-biased diode to provide a small signal conductance which varies directly with control current. Both signal and control currents are fed from constant current sources, the former through blocking capacitors as shown in Fig. 6. These capacitors provide high and low impedance for control and signal currents, respectively, to keep the control from appearing in the output.

When the control current is large compared to the signal (and also to the saturation current of the diode), the small signal resistance of the variolosser diode is:

$$\frac{de}{di} \doteq \frac{kT}{qi}_{control}$$

where k is Boltzmann's constant, T the absolute temperature and q the charge on an electron.

The variolosser voltage,

$$E_{out} = i_{signal} \cdot \frac{de}{di}$$

$$= \frac{kT}{q} \cdot \frac{i_{signal}}{i_{control}}$$

This shows that the variolosser output is proportional to the ratio of signal to control currents, which is the relationship assumed in the analysis.

The detector-variolosser combination has a control range in excess of 24 db. This is believed to be adequate since only a small percentage of rapid fades are expected to exceed this depth.[2]

To measure suppression due to FAAGC, the loop is alternately opened and closed as shown in Fig. 6. Since the control current has a dc component, the proper bias current must be maintained in the variolosser diode in the open loop condition. The LPF and the following low-frequency amplifier must also pass dc. The amplifiers shown in Fig. 6 provide proper levels and impedances in the experimental circuit. Further development should reduce the circuit complexity.

Simulation of Fade - SSB Signal

A very simple model of fading assumes a receiver moving directly toward a fixed transmitter and directly away from a fixed reflector as shown in Fig. 7. If the transmitted carrier is at frequency f_o and the receiver velocity is v, the fade frequency

$$F = 2\frac{v}{c} f_o = \text{Doppler frequency}$$

where

c = velocity of light

For f_o = 850 mc and v = 60 mph., F = 152 cps.

When v = o (no fading), the received (upper sideband) signal is:

$$E = \cos 2\pi f_o t + m_s \cos [2\pi(f_o+S)t+\theta_s] \quad (1)$$

where

m_s = index of signal modulation

and S = signal frequency

With fading, the received signal E is made up of four voltage vectors as shown in Fig. 7. These are:

Direct carrier =

$$E_D = \cos[2\pi(f_o+F/2)t + \theta_f]$$

Direct sideband =

$$E_S = m_s \cos[2\pi(f_o+F/2+S)t + \theta_f + \theta_s]$$

Reflected carrier =

$$E_R = m_f \cos[2\pi(f_o-F/2)t - \theta_f]$$

Reflected sideband =

$$E_{RS} = m_f m_s \cos[2\pi(f_o-F/2+S)t - \theta_f + \theta_s] \quad (2)$$

where the relative amplitude of the reflected wave, m_f, may be considered as a fading modulation index.

We note that the direct and reflected waves have been shifted up and down, respectively, in frequency, by one-half the fading frequency, F.

For the frequencies and path length differences being considered, we may assume that the sidebands fade always in phase with the carrier, and (2) is based on this assumption. As this equation shows, there are four different frequency components in the received signal. But at a particular instant of time, the four corresponding voltage vectors E_D, E_S, E_R, and E_{RS} will appear as shown in Fig. 7. The relative phases of these vectors at this instant may then be found by comparing the arguments of the cosines in (2). From this we find that

$$\angle E_S - E_D = \angle E_{RS} - E_R = \angle \sigma$$

which shows that the carrier-to-sideband phase difference is always the same for both direct and reflected waves. This is shown in Fig. 7. It is important that the experimental fade simulator maintain this correct phase relationship.

The maximum fading frequency, F, which must be accommodated is being investigated by other members of the Laboratories. For the purposes of this paper, we may assume that F_{max} = 150 cps.

Suppression by FAAGC - Analysis

For small values of m_f and m_s, a Taylor series expansion gives an approximation to the detected envelope at frequencies F, S, 2F, (S+F) and (S-F). This is the envelope of the variolosser output with the FAAGC loop open, since the variolosser diode resistance is constant.

Frequency	Magnitude
F (fading)	m_f
S (signal)	m_s
2F	$\frac{m_f^2}{4}$
S + F	$\frac{m_f m_s}{2}$
S - F	$\frac{m_f m_s}{2}$

(3)

With the FAAGC loop closed, the envelope given by (3) is modified by the amplitude and phase characteristics of the LPF and applied to the control lead to vary the variolosser diode resistance. Envelope (3) is thus divided by the control envelope. An approximate analysis shows that with the FAAGC loop closed the envelope of the variolosser output is:

Frequency	Magnitude
F	$m_f\left((1-\gamma_f\cos\beta_f)^2 + (\gamma_f\sin\beta_f)^2\right)^{1/2}$
S	$m_s\left((1-\gamma_s\cos\beta_s)^2 + (\gamma_s\sin\beta_s)^2\right)^{1/2}$
2F	$\frac{m_f^2}{4}\left\{(1-\gamma_{2f}\cos\beta_{2f}-2\gamma_f^2\cos2\beta_f+2\gamma_f\cos\beta_f)^2 + (\gamma_{2f}\sin\beta_{2f}+2\gamma_f^2\sin2\beta_f-2\gamma_f\sin\beta_f)^2\right\}^{1/2}$
S + F	$\frac{m_f m_s}{2}\left\{\left(1-\gamma_{s+f}\cos\beta_{s+f}+2\gamma_f\gamma_s\cos(\beta_f+\beta_s)-\gamma_s\cos\beta_s-\gamma_f\cos\beta_f\right)^2 + \left(\gamma_{s+f}\sin\beta_{s+f}-2\gamma_f\gamma_s\sin(\beta_f+\beta_s)+\gamma_s\sin\beta_s+\gamma_f\sin\beta_f\right)^2\right\}^{1/2}$ (4)
S - F	$\frac{m_f m_s}{2}\left\{\left(1-\gamma_{s-f}\cos\beta_{s-f}+2\gamma_f\gamma_s\cos(\beta_f-\beta_s)-\gamma_s\cos\beta_s-\gamma_f\cos\beta_f\right)^2 + \left(\gamma_{s-f}\sin\beta_{s-f}-2\gamma_f\gamma_s\sin(\beta_f-\beta_s)-\gamma_s\sin\beta_s+\gamma_f\sin\beta_f\right)^2\right\}^{1/2}$

where

γ = unity based transmission factor of the LPF at the subscript frequency, e.g., in Fig. 3, if F = 200 cps., the transmission is down 3.5 db, therefore $\gamma_f = 0.668$.

β = phase error in radians at the subscript frequency. This is $(\varphi_{LPF} - \varphi_{BPF})/57.3$ in Fig. 5. If F = 200 cps., the phase error is 18.8° and $\beta_f = 0.328$.

The suppression of each frequency component by FAAGC is given by the ratio of the corresponding magnitudes shown in (3) and (4) with the loop open and closed. The calculation of these suppressions was made using a high-speed digital computer.

Measuring Suppression by FAAGC

An approximate simulation of the type of fading being considered is shown in Fig. 8. The direct carrier is simulated by a fixed oscillator operating at 453.8 kc. The reflected carrier, attenuated by m_f, the fading modulation index, is simulated by an adjustable oscillator differing in frequency by F, the fading frequency; therefore, its frequency is (453.8 kc-F). The direct and reflected sidebands are simulated by the modulation of signal frequency, S, with the fixed and adjustable oscillators, respectively; the latter output being attenuated by m_f to simulate the proper magnitude of the reflected sideband. The SSB filter suppresses the unwanted (lower) sidebands produced by the modulators. Modulation indices m_f, and m_s may be adjusted as shown.

As shown in Fig. 8, the output of the fade simulator is fed to the experimental FAAGC circuit. Its output is in turn connected through an envelope detector to an audio frequency wave analyzer which measures the magnitude of frequency components F, S, 2F, etc. With frequencies F and S and indices m_f and m_s adjusted to the desired values, the suppression due to FAAGC is measured by observing on the wave analyzer the reduction in output when the FAAGC loop is closed.

The wave analyzer must be able to discriminate between closely spaced frequencies, and this selectivity demands a high degree of frequency stability in the fixed and adjustable oscillators. However, the fade simulator shown in Fig.

8 is quite satisfactory where frequency S is to be varied. Where S is held constant for a series of measurements, the phase shift type of SSB generation is preferable.[7,8] The use of a rotating phase shifter to simulate the fading frequency has also been considered; however, the required speeds are unusually high, for example, if F = 150 cps. the phase shifter must rotate at 9,000 rpm.

It has been found that the magnitude of the (S+F) and (S-F) components of the envelope are sensitive to slight amplitude and phase variations in the composite fading signal, i.e., the vectors shown in Fig. 7 must have the correct magnitudes, and angles "σ" must be equal. If the magnitude at E_{RS} is in error by 1 db, (S-F) is changed by 2 db. Because of this sensitivity, precise design and adjustment of the measuring circuit is required for valid suppression measurements.

Suppression of Fading Frequencies

With the loop open, the magnitudes of F and S relative to unity (the magnitude of the direct carrier) are given by (3). With $m_f = m_s = 0.2$, these magnitudes are equal and are plotted as 0 db in Fig. 9. With the loop closed, these magnitudes are reduced to those shown in (4), and the dashed lines in Fig. 9. Note that the suppression peak at 100 cps. corresponds, as would be expected, to the frequency of zero phase error shown in Fig. 5.

The calculated suppressions for F = 100 and S = 300 cps. are 39 db and 1.5 db, respectively. The measured suppressions are seen to be in good agreement except at the lower frequencies.

The suppression of 2F is somewhat poorer than that of F especially at the lower frequencies. However, the second harmonic of the fading frequency is usually considerably below the fundamental without suppression, for example, if $m_f = 0.2$, (3) shows that 2F/F is -26 db. If F = 100 cps., the suppression of 2F is approximately 7 db. In this case, suppressed 2F is 33 db below unsuppressed F.

Suppression of Modulation Products

The suppression of the modulation products of frequencies (S±F) is important, since most of these products fall in the voice band between 300 and 3,000 cps. and, therefore, cannot be removed by subsequent filtering.

As in the case of F and S, the relative magnitudes of these components with the loop open is given in (3). Compared to the signal, S, (S±F) will be down by $m_f/2$, i.e., 20 db if $m_f = 0.2$. This loop open (no suppression) level is normalized to 0 db in Fig. 10.

With the loop closed, the magnitudes of (S±F) are reduced to those shown in (4). These magnitudes are plotted by dashed lines in Fig. 10 for S = 300, 1,000 and 3,000 cps. When F = 100 cps., Fig. 10 shows the following suppressions for frequencies within the voice band:

S	Freq.	Suppression Calculated	Suppression Measured
300 cps.	S + F	-13.5 db	-13.0 db
1,000 cps.	S - F	-17 db	-16 db
1,000 cps.	S + F	-22 db	-20 db
3,000 cps.	S - F	-12 db	- 2 db

It will be noted from (4) and Fig. 10 that the suppressions of (S+F) and (S-F) are not equal. Furthermore, the suppressions of these modulation products do not increase markedly with increasing frequency S. These rather unexpected effects appear to be the result of transmission (γ) in the voice band combined with frequency dependent phase errors (β) (see Figs. 3 and 5). It is true that a "conventional" sharp cutoff LPF would improve the (S±F) suppression, especially in the upper part of the voice band. However, as mentioned previously, it would appear to be impractical to match the delay of such a filter.

By inspection of (3) and (4), it will be noted that the suppressions shown in Figs. 9 and 10 are independent of m_f and m_s, based on the approximation that these indices are small.

Discussion of Results

Suppressions by conventional AGC and FAAGC may be compared by examination of Figs. 1, 9 and 10. For example, with F = 100 cps. and S = 300 cps., Figs. 1(a) and 9 show the following suppressions:

Frequency	Conventional AGC	FAAGC
100 cps.	-10 db	-39 db
300 cps.	- 3 db	- 1.5 db

The advantage of FAAGC is even more apparent in the suppression of modulation products, since Fig. 1(a) shows practically no suppression of these products by conventional AGC.

It is possible that some combination of backward-acting AGC and filtering might yield some suppression of these components; however, the inability to match delays appears to be a basic difficulty.

As mentioned previously, maximum suppression by FAAGC requires matching delays at the fade frequencies and "flat" transmission in the signal path. In practice it may be difficult to obtain these characteristics with the required accuracy. To determine this accuracy, the analysis given herein is being extended to include signal path transmission irregularities.

Conclusions

A forward-acting AGC circuit for combating the effects of rapid fading in UHF mobile telephone systems has been subjected to preliminary analysis and test. Considerable improvement in the suppression of fading and fading modulation products has been shown. Results of suppression measurements are in fair agreement with theory; however, more exact analyses and test methods are indicated. Also, the advantages have yet to be confirmed by subjective testing.

Forward-acting AGC, especially after further development, should be an important contribution to the solution of the rapid fading problem.

Acknowledgments

This forward-acting AGC study was first proposed and the analysis started by H. G. Unger. The author wishes to thank A. J. Rack for the design of the special low-pass filter, J. F. Ossanna and B. P. Bogert for their help in analysis and J. L. Wenger for the experimental work.

References

1. W. D. Lewis, "Coordinated Broadband Mobile Telephone System," I.R.E. Trans. on Vehicular Communications, Vol. VC-9, No. 1, pp. 43-48, May, 1960.

2. Nathan W. Feldman, "An Analysis of Radio Flutter in Future Communications," I.R.E. Trans. on Vehicular Communications, Vol. PGVC-13, pp. 27-36, September, 1959.

3. W. L. Firestone and H. Magnuski, "Application of Single Sideband for Mobile Communication," I.R.E. Trans. on Vehicular Communications, Vol. PGVC-11, pp. 48-54, July, 1958.

4. B. M. Oliver, "Automatic Volume Control as a Feedback Problem," Proc. of the I.R.E., Vol. 36, pp. 466 - 473, April, 1948.

5. Richard F. Baum, "A Contribution to the Approximation Problem," Proc. of the I.R.E., Vol. 36, pp. 863-869, July, 1948.

6. Don L. Lundgren, "Electromechanical Filters for Single-Sideband Applications," Proc. of the I.R.E., Vol. 44, pp. 1744-1749, December, 1956.

7. Donald E. Norgaard, "The Phase Shift Method of Single-Sideband Generation," Proc. of the I.R.E., Vol. 44, pp. 1718-1735, December, 1956.

8. J. D. Griffiths, "Voice Modulated SSB & DSB Peak to Average Envelope Power Ratios," 1961 I.R.E. International Convention Record, Part 8, "Communication Systems - Techniques," pp. 85-101.

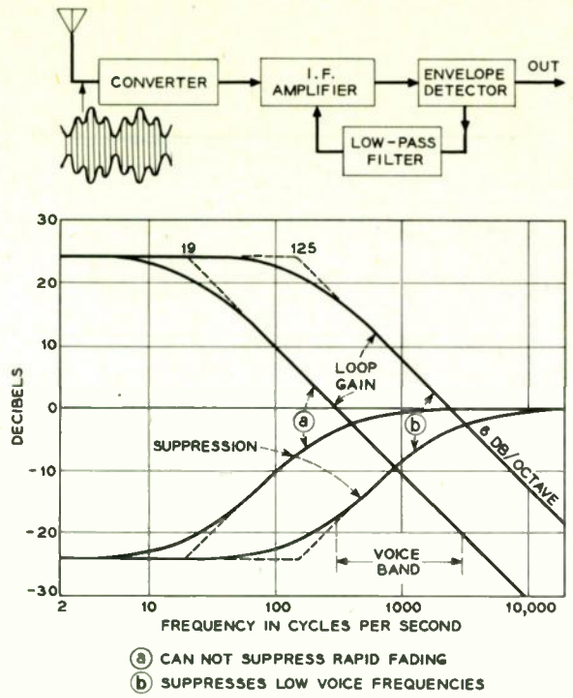


Fig. 1.

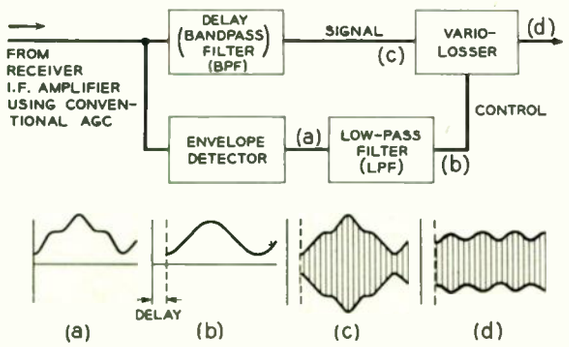


Fig. 2.

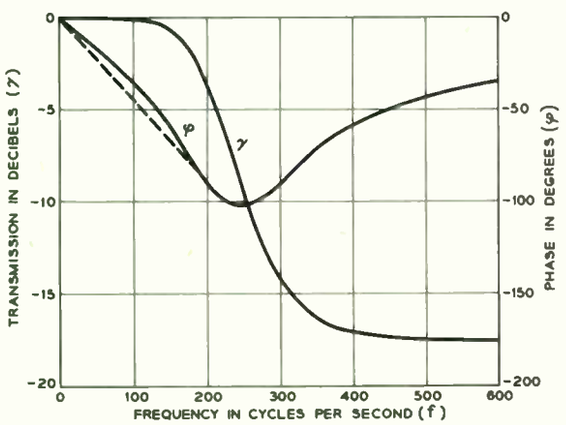


Fig. 3.

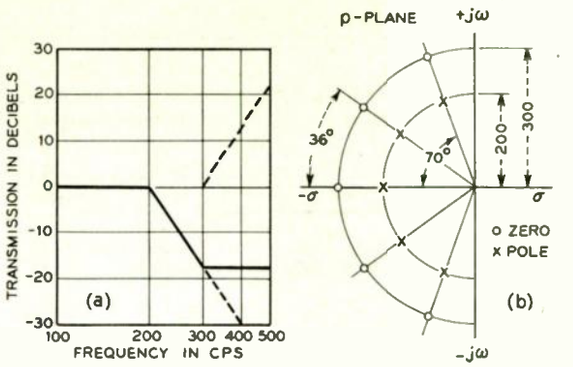


Fig. 4.

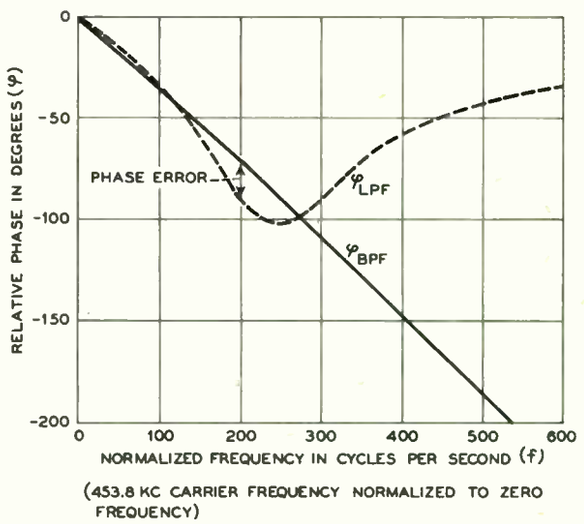


Fig. 5.

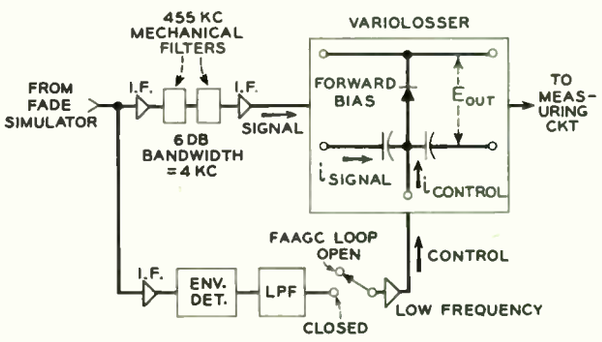
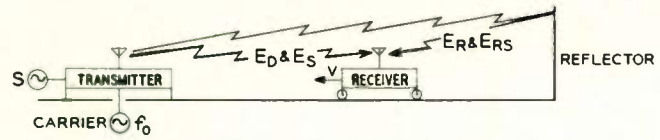
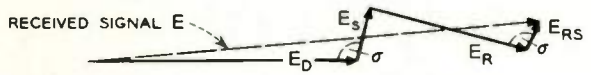


Fig. 6.



FADING FREQUENCY = $F = 2\frac{V}{C} f_0$ FADING MODULATION INDEX = m_f

SIGNAL FREQUENCY = S SIGNAL MODULATION INDEX = m_s



VOLTAGE VECTOR	RELATIVE MAGNITUDE	FREQUENCY V = 0 (NO FADE)	FREQUENCY WITH FADE
E_D = DIRECT CARRIER	1	f_0	$f_0 + \frac{1}{2}F$
E_S = DIRECT SIDEBAND	m_s	$f_0 + S$	$f_0 + \frac{1}{2}F + S$
E_R = REFLECTED CARRIER	m_f		$f_0 - \frac{1}{2}F$
E_{RS} = REFLECTED SIDEBAND	$m_f m_s$		$f_0 - \frac{1}{2}F + S$

Fig. 7.

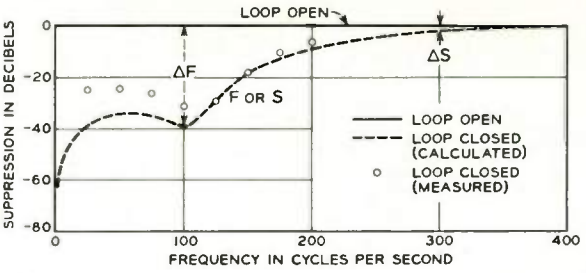


Fig. 9.

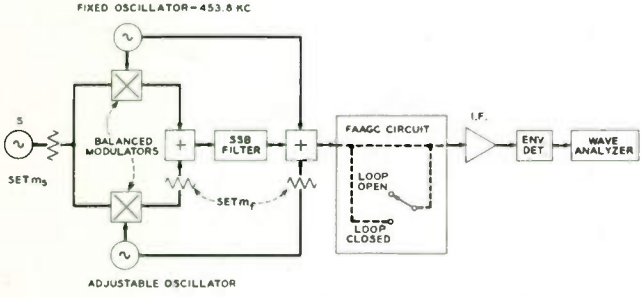


Fig. 8.

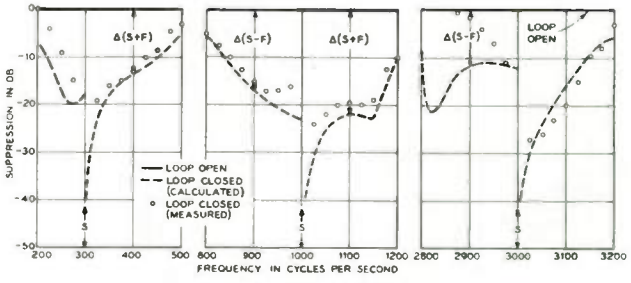


Fig. 10.

A MINIATURE TUNED REED SELECTOR OF HIGH SENSITIVITY AND STABILITY

L. G. Bostwick
Bell Telephone Laboratories, Incorporated
New York, New York

Summary

This paper describes a selective contacting device that is responsive only to sustained frequencies in a discrete narrow band and is insensitive to speech and noise interference. It is of small size suitable for use in a pocket carried radio receiver and is sufficiently stable to permit 33 discrete resonant frequencies, spaced 15 cycles apart in less than an octave between 517.5 and 997.5 cycles per second. It has a threshold sensitivity of about 35 microwatts and other operating characteristics that are essential in large capacity systems.

Introduction

Tuned reed selectors used as selective receivers in multifrequency systems involving large numbers of individual selections, such as Personal Radio Signaling[1] must operate within close and specifiable limits in order to avoid false signaling and to assure satisfactory performance under devious environmental and circuit conditions. In particular three operating characteristics, or their equivalent, must be controlled; namely, the resonant frequency, the sensitivity (current or power needed at the most sensitive frequency), and the bandwidth (the frequency band in which contacting occurs with an input power twice that needed at the most sensitive frequency).

The permissible variation in these characteristics is much smaller than would seem necessary from first considerations. Resonant frequency changes that seem negligible compared to the frequency spacing between adjacent selectors often become important when other system requirements are considered simultaneously. For example, the frequency range over which contacting will occur depends upon the electrical input level and the selector bandwidth. Consequently, feasible limits for both of these latter quantities must be considered and in determining allowable frequency deviations from nominal, the lowest probable input level and the narrowest bandwidth must be taken into account. On the other hand excessively high input levels cannot be allowed even in those unusual instances where conserving power is unimportant because this necessitates wider channel separations in order to avoid transient operation of adjacent selectors particularly those having high sensitivities. Furthermore high input levels result in longer decay times which often cannot be tolerated. When these and other related factors are considered and the widest manufacturing tolerances are sought, it is found that the above three selector characteristics are closely interrelated and one cannot be relaxed without making one or both of the others more stringent.

The tuned reed selectors described in this paper have factory adjustment provisions and sufficient structural stability to control in a practical manner the resonant frequencies, the sensitivities, and the bandwidths within adequate and compatible limits. As a result it is feasible to use 33 discrete resonant frequencies, 15 cycles apart, in less than an octave between 517.5 and 997.5 cycles. An available electrical power of 35 microwatts at each individual resonant frequency will just operate the contact and a power level of 100 microwatts will close the contact to a low resistance over 20 per cent or more of the reed period. These and other capabilities to be described distinguish these selectors from many others that are not adequate for reliable operation in large systems.

General Description

Fig. 1 is a photograph showing one complete reed selector and another with the outside shell removed. Fig. 2 is a partially exploded view showing the subassemblies and indicating how the parts are fitted together. The shell is formed from permalloy sheet and serves as an effective shield from extraneous fields and as a high permeability flux path for the internal magnetic circuit. All parts are electrically insulated from the shell. The over-all dimensions excluding the terminals are 3/8" x 15/32" x 1-7/16" and the complete selector weighs about 8 grams.

As shown in these photographs, a tuning fork formed from two reeds brazed to a base block serves as the resonant element. This balanced type of structure does not require a massive support as would a single cantilever reed in order to isolate it from extraneous influences, an important matter for a miniature device. This fork is freely supported within the shell by a compliant frame that further isolates any small residual vibration of the fork base from the rest of the selector and yet is sufficiently stable to permit the vibrating contact on the end of the tuning fork tine to be precisely positioned with respect to the stationary contact. This latter contact is carried by a loop of wire spot welded to a rotatable stud that fits into a tapered hole in an insulating bushing in the frame between the tines. A magnetic polepiece is positioned between the open ends of the tines forming two equal gaps. Polarizing magnetic flux is set up in these gaps by a small permanent magnet attached to the opposite end of the polepiece. The energizing coil surrounds the center portion of the polepiece.

The tuning fork is made of a nickel-iron-molybdenum alloy[2] (vibralloy) having controlled elastic and magnetic properties. Annealed permalloy with low coercive force and high permeability is used for the polepiece and shell to reduce magnetic flux changes. The materials and

shapes of other parts are chosen to minimize dimensional changes with time and environmental conditions.

Frequency Selection and Fine Tuning

The range of resonant frequencies is obtained with tuning forks that have the same over-all length but varying free tine lengths. The small dimensions of these forks require the brazing fillets and the free reed lengths and thicknesses to be precisely controlled. By special attention to rolling of the reed stock, precise jigging of the reeds and base block, and brazing with minimum fillet dimensions, it is feasible to produce forks in which the individual tine frequencies are sufficiently close to chosen nominal frequencies spaced 15 cycles apart so that they may then be accurately tuned to these desired frequencies.

Precise or fine tuning is accomplished with spring sliders that may be moved along the tines. This requires a slider that will stay in place under shock and vibration, will provide an adequate tuning range, and will allow the necessary fineness of frequency adjustment. This is achieved by means of small spring clips that snap on and ride along the edges of the tines. These sliders are shaped so that pressure at the center releases the force with which the slider seizes the reed and permits it to be moved. Each slider has a mass of about 1 milligram and provides a tuning range of about 10 cycles on forks near 500 cycles and about 25 cycles on forks near 1000 cycles. The sliders may be moved in increments less than a thousandth of an inch, permitting the resonant frequencies to be readily set to a desired value within ±0.05 cycle. The seizure forces are large so that shock and vibration acceleration in excess of 1500 G are required to move the sliders.

Contact Facility and Sensitivity Adjustment

The sensitivity is adjusted in manufacture by changing the contact gap separation. A fine rhodium wire having a resonance frequency above the frequency range of the tuning forks is supported by a loop of larger wire that may be rotated on a tapered stud through the frame. The fine wire is pretensioned with a prescribed force against the loop wire to form a lift-off type of contact that is accurately positioned and will follow large tine excursions without objectionable interference with the tine motion. This construction[3] results in a contact that makes to a low resistance with the vibrating contact on the reed for intervals of time that may be 25 per cent or more of the reed period depending on the applied power. The operating sensitivity of the selector is precisely set by rotating the loop on the stud axis and thereby causing the end of the contact wire to move toward or away from the reed contact. The point of contact is close to the axis of rotation so that a fine control of the contact gap may be achieved.

Bandwidth Control

The bandwidth or sharpness of the resonance curve is determined primarily by three dissipative factors, namely: internal frictional losses in the reed material, viscous losses in the air surrounding the reeds, and eddy current losses in electrically conducting parts. The last factor has been chosen as the adjustment or control means for bandwidth. A copper washer is placed around the polepiece and where flux changes due to motion of the reeds induce eddy currents in the copper. By selecting the proper washer thickness and diameter and setting the magnet strength to yield the proper flux density, eddy currents are developed when the tines vibrate that absorb energy and reflect into the system as an effective mechanical resistance that broadens the resonance curve by the desired amount.

Vibrating System Parameters

Tabulated below are some measured and derived data that show the magnitudes of the more important vibrating system constants of two selector samples with resonant frequencies nearly an octave apart. These are typical values that will be of interest to those concerned with the vibrational mechanics, electromechanical coupling, and other analytical design factors.

TABLE 1

	Nominal Freq. 517.5	Nominal Freq. 997.5
Reed Dimensions	Length 1.4 cm Thickness 0.015 cm Width 0.254 cm	1.01 cm 0.015 cm 0.254 cm
Effective Reed Stiffness	1.45×10^5 dynes/cm	3.88×10^5 dynes/cm
Resonant Freq. as Brazed	560 cycles	1068 cycles
Resonant Freq. with Contact	530 cycles	1011 cycles
Resonant Freq. with Slider as tuned	517.5 cycles	997.5 cycles
Effective Reed Mass as Brazed	.0118 grams	.0087 grams
Effective Reed Mass with Contact	.0130 grams	.0096 grams
Effective Reed Mass with Slider as tuned	.0138 grams	.0099 grams
Electrical Impedance at Resonant Frequency	478 +j231	448 +j430
Electrical Blocked Impedance at same freq.	220 +j277	235 +j485
Electrical Motional Impedance at same freq.	258 -j46	213 -j55
Current to just close Contact	.275 milliamps	.275 milliamps
Bandwidth	1.1 cycles	1.3 cycles
Effective Mechanical Resistance of Fork at Resonance	.19 mech. ohms	.16 mech. ohms
Electromechanical Coupling Factor	$2.24 \times 10^5 \lfloor 5^0$ dynes/abamp	$1.88 \times 10^5 \lfloor 7.2^0$ dynes/abamp
Effective Magnetic Gap Stiffness (each gap - from freq. shift measurements)	$-.02 \times 10^5$ dynes/cm	$-.02 \times 10^5$ dynes/cm
Corres. Gap Flux Density	200 gauss	200 gauss
Max. Tine Flux Density (assuming Fringe Flux equal to Gap Flux)	4000 gauss	4000 gauss

Performance Objectives

Consideration of the over-all system operating requirements for Personal Radio Signaling pertaining to such factors as the needed number of individual selections, practical radio receiver power levels, calling rates, and environmental conditions, led to the following objectives for the performance of the reed selectors.

(1) Nominal Frequency Range - 517.5 to 997.5 cycles

(2) Nominal Frequency Separation - 15 cycles

(3) Frequency Deviation Limits - ±0.3 cycle

Including adjustment tolerances, aging, shock, magnetic changes and all other instabilities except those due to temperature changes.

(4) Temperature-frequency deviation limits - ±0.2 cycle

Over temperature range of 35°F to 110°F (2°C to 43°C).

(5) Nominal bandwidth - 1.0 cycle

(6) Bandwidth deviation limits - 0.8 - 1.4 cycles

Resulting from temperature changes and all other causes.

(7) Nominal current to just operate contact - 0.25 milliamps

For a nominal 500 ohm coil impedance at resonance.

(8) Just operate current deviation limits - ±3.0 db

Resulting from temperature changes and all other causes.

These objectives are mutually consistent in that the limits given in each case are as large as can be tolerated without reducing the limits on some other factor. There are other important design considerations that must not be neglected such as weight, size and shape, contact life, shock tolerances, corrosion resistance, magnetic interaction and so forth, and with respect to which the selectors must, of course, be adequate. However, the above-tabulated characteristics are the most significant from an operating standpoint and are sufficient under marginal conditions to assure positive operation and avoid false signaling.

Typical Measured Data

Presented below are measured data showing that the above-described reed selector meets these objectives. By means of the spring sliders the two tine frequencies are made alike within a small fraction of a cycle and of values to give a combined fork frequency well within requirements. Attention is given in the assembly and adjustment procedure to magnetically and mechanically stabilize the whole structure. The magnet is stabilized well below its maximum remanence, the whole final assembly is subjected to a moderately high temperature to relieve residual stresses, and the tines are vibrated at a suitable level to bring them into a normalized magnetic state prior to final adjustment. The resulting selectors have resonant frequencies that will remain within ±0.3 cycle from their nominal frequencies at normal room temperatures and under reasonable conditions of mechanical shock and electrical overload. Negligible changes occur under shocks up to 1500 G (2 milliseconds duration) or with input levels 20 db above the just operate values.

Frequency stability with temperature is achieved by making the forks of a nickel-iron-molybdenum alloy of such a composition that magnetic permeability changes are small and the temperature coefficient of Young's modulus is low and of a magnitude to compensate for dimensional changes with temperature. Operate current stability is realized by additional attention to the design geometry and materials so that changes in temperature cause variations in contact separation that are a small fraction of a mil-inch.

Fig. 3 and Fig. 4 are graphs of measured data showing variations with temperature in the resonance frequency, just operate current, and bandwidth of two typical samples, one at each end of the nominal frequency range. The range covered by these graphs is much wider than that required for most applications. In the more common temperature range of 35°F to 110°F (2°C to 43°C), the deviations are well within the limits tabulated above.

Fig. 5 and Fig. 6 are electrical impedance diagrams of the same two selector samples with resistance and reactance as coordinates and frequency as the variable parameter. This form of plot emphasizes the interesting values near resonance and may be used for analytical purposes[4]. From these graphs, it can be determined that the conversion of electrical to effective mechanical power is about 46 per cent and the available electric power necessary to just operate the contact is about 33 microwatts.

Nominal Operating Levels and Times

The electrical power source supplying selectors in a system must have an available power capacity sufficient to cause dependable contacting under the worst temperature and adjustment conditions. These worst conditions obtain when the frequency deviation from nominal and the just operate current are at their maximum values. Considering the limits permitted in these selectors and making allowance for contact quality

and life with some statistical advantage taken of the small chance of all limiting conditions occurring simultaneously, it was determined that the minimum electrical input power should be 6 db above that needed to barely close the contact of a nominal selector. At this level, the time required to close the contact after energizing the coil is equal to the time needed for the reed amplitude to decay below contacting amplitude after the coil current is stopped. For nominal selector constants, this time is approximately 225 milliseconds. Input levels higher than 6 db above just operate will result in faster operating times and slower decay times but the sum of the operate and decay times will increase less than 20 per cent up to input levels 12 db above the nominal just operate value.

Contact Capacity and Life

The contact has greater capability than would at first seem likely. Such a light contact is most frequently used in circuits to change the potential on a tube or transistor and thereby trigger some desired signaling or switching function without the contact current exceeding a few milliamperes. The contact closure is intermittent at a rate corresponding to the frequency of the selector and the duration of the individual closures is a small fraction of a millisecond depending upon the frequency and input level. These short closures, however, occurring at a rate of several hundred times per second may control current pulses that have an integrated or averaged power that is a substantial fraction of a watt.

The maximum power that can be controlled depends mostly upon the reactive elements in the contact circuit and the life needed from the selector. As an example of what may be expected, Fig. 7 shows changes that occurred in the resonance frequency and the sensitivity of a typical selector when operated continuously (except for a few minutes about every 100 hours during check test) over a period of 1500 hours. The electrical input was 9 db above the just operate value and the contact closed a 12 volt battery through a 240 ohm resistor giving a closure current of 50 milliamperes. Throughout the test period the resonance frequency changed only slightly and the just operate current increased about 20 per cent. This later change was due to erosion of the contact wire which increased the contact gap. Erosion was minimized by connecting the fine contact wire to the negative side of the battery. At the end of the test the diameter of the contact wire was approximately half its original value.

Applications

The manner in which these selectors are used in the circuits of the BELLBOY Personal Radio Signaling system will be described in a paper to be published on the pocket radio receiver. In this system three tuned reed selectors are operated simultaneously in the receiver and these trigger a transistor oscillator that gives an audible signal. The power controlled by the contacts in this case is small.

The substantial power capacity of the contacts can be used to operate relays and other devices directly. Pulses of current from a battery at the selector frequency can be supplied to a smoothing or integrating capacitor and the relatively constant voltage across the capacitor used to operate a sensitive dc relay. The battery may be at the location of the reed selector or may be supplied by superposition over the same circuit used to transmit the selector frequency.

The contact may also be used as a synchronous rectifying means to generate dc from the same ac source that operates the selector as shown in Figure 8. When the source frequency corresponds with that of the reed selector, the contact of the selector closes in synchronism once each cycle to send unidirectional pulses to the capacitor and relay in parallel. The capacitor smoothes the pulses and gives a nearly constant current in the relay winding. For maximum sensitivity it is desirable that the contact closures occur near the peaks of the supply voltage wave and this is accomplished by connecting a large reactance (either inductive or capacitative) in series with the selector winding. This reactance also serves to attenuate the supply voltage applied to the selector winding to avoid overdriving the reeds because a supply voltage large enough to operate a relay is ordinarily many times that needed to operate the reed selector. Combination circuits using reed selectors and mercury wetted contact relays provide a simple means of selectively controlling substantial powers to perform a multiplicity of functions over a single pair of wires.

When operated just below the contacting level these selectors have a Q (resonant frequency to bandwidth ratio) in the range of 500 to 1000 and therefore may be used effectively in a selective bridge or filter circuit as described in a previous paper[5]. The use of suce a selective circuit in the feedback loop of a single transistor oscillator results in an attractively simple source of frequency having a precision corresponding to that of the selector.

Acknowledgments

Original suggestions regarding the construction of this selector and skilled model work were contributed by the late R. L. Guncelle. Essential refinements in the design and in the fabrication techniques were made by K. F. Bradford and D. H. Wenny. E. J. Kasello carried out most of the adjusting and testing. The successful outcome of this development is due in no small measure to their efforts.

References

1. D. Mitchell and K. G. Van Wynen - "150 MC Personal Radio Signaling System" - Bell System Technical Journal - September 1961

2. M E. Fine and W. C. Ellis - "Thermal Variation of Young's Modulus in Some Fe-Ni-Mo Alloys" - Journal of Metals - September 1951 - Also U. S. Patent 2,561,732

3. Alternatives disclosed in U. S. Patent 2,877,319

4. F. V. Hunt - "ELECTROACOUSTICS" Harvard University Press - John Wiley and Son, New York - 1954

5. A. C. Keller and L. G. Bostwick - "Vibrating Reed Selectors for Mobile Radio Systems" - also H. M. Pruden and D. F. Hoth - "Vibrating Reed Signaling for Mobile Radio" - Both papers in Transactions of the AIEE Vol. 68, 1949 - Further Information in U. S. Patents 2,583,542 and 2,630,482

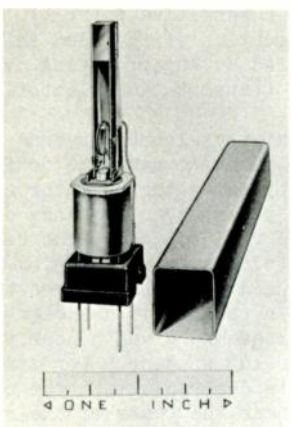


Fig. 1.

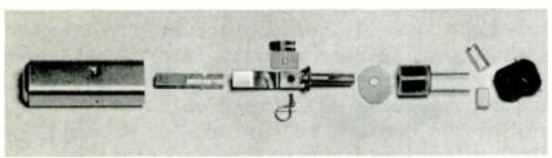


Fig. 2.

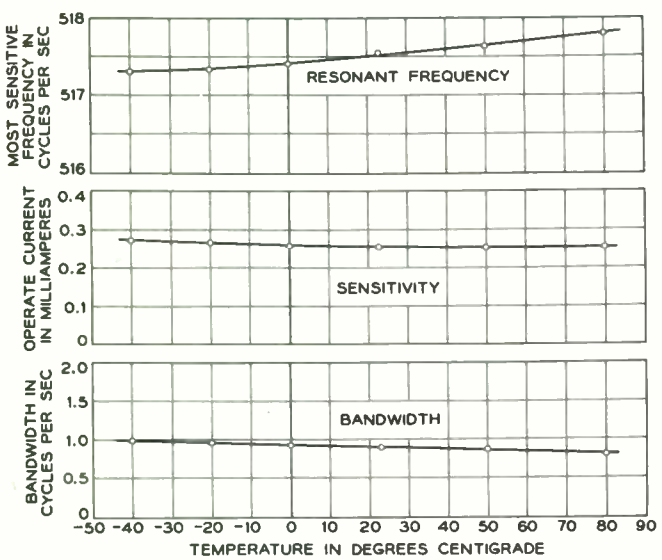


Fig. 3.

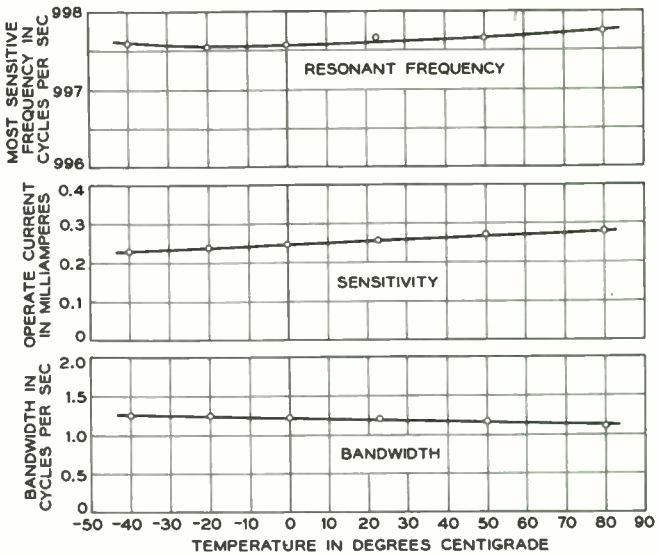


Fig. 4.

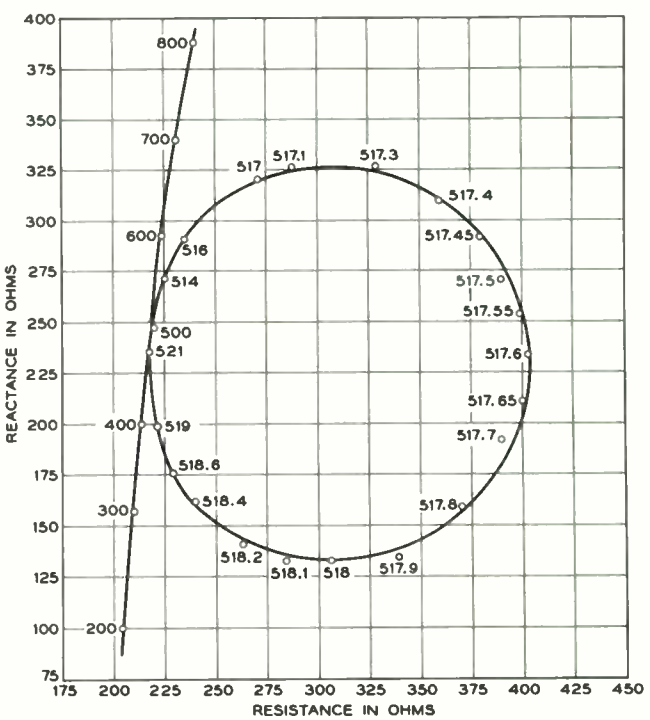


Fig. 5.

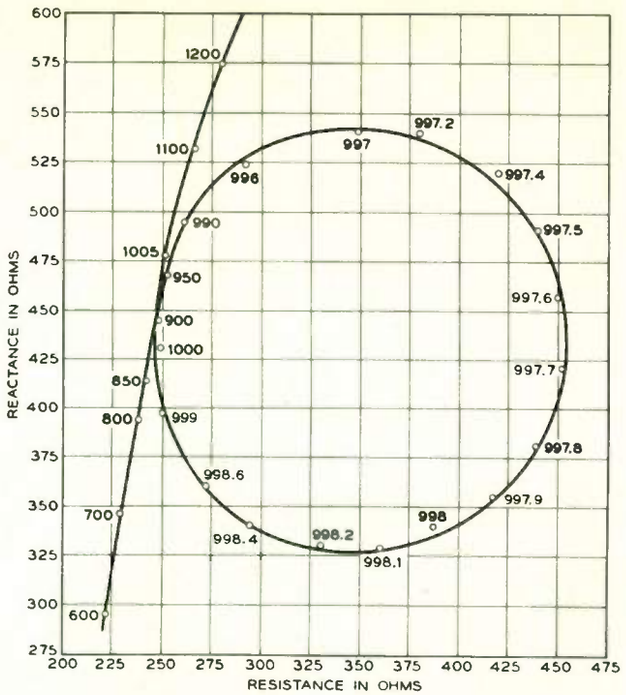


Fig. 6.

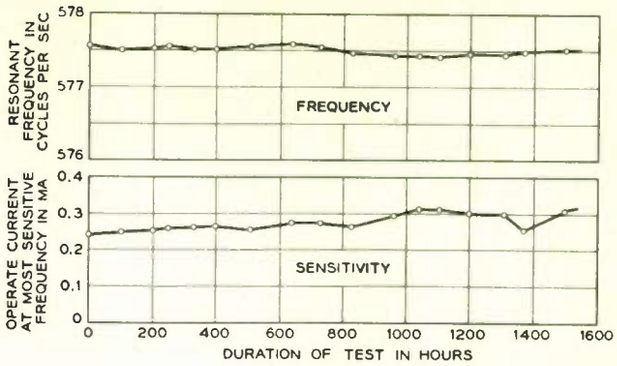


Fig. 7.

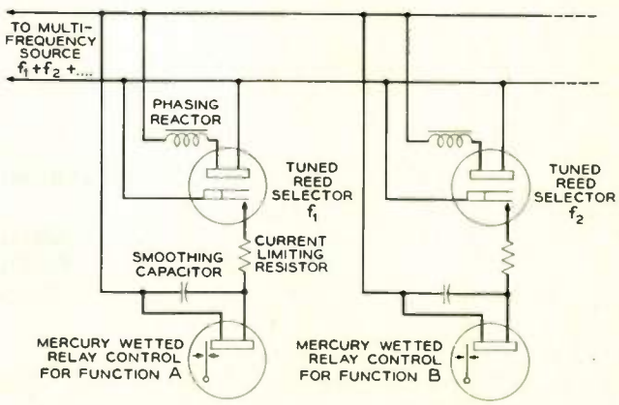


Fig. 8.

TWO-WAY RADIO FOR MOBILE TELEPHONE SERVICE

Martin Cooper
Motorola, Inc.
Chicago, Ill.

Abstract

The increasing use of mobile radio systems in telephone service applications has imposed unique requirements on equipment manufacturers. This paper discusses these requirements with emphasis on the operational procedures used in existing telephone systems and on the inherent equipment problems. Design features important to this application are presented along with specific examples in practical installations.

SPACE RADIO TELESCOPES

A. E. Lilley
Harvard University Observatory
Cambridge, Mass.

The study of the universe with radio telescopes is an excellent example of one way communications. In fact, the entire history of astronomical research has been communications in a one way sense. The only way we can learn anything at all about the universe is through the study of electromagnetic radiation emitted by objects in space.

The introduction of radio techniques into the field of astronomical research has revolutionized the study of astronomy. The contributions of radio research in astronomy have ranged from the study of meteors to the moon, the planets, the sun, the galactic system, and to objects at very great distances in the universe. Any one of these topics would exhibit its own particular new body of information, by virtue of radio data which have been gathered over the past decade. In many cases our study of the solar system and the universe with radio waves has led to totally unexpected results.

There are many cases to choose from, and I have chosen the planet Venus as an example. The radio wave study of the planet Venus with terrestrial radio telescopes has been keenly pursued on an international basis since 1955. The results of the radio studies of the planet Venus have led to something of a dilemma regarding its physical environment. The radio data have posed serious questions which must be answered before we may look forward to a detailed exploration of the planet with space vehicles.

Before the emergence of the new radio studies of Venus, our concept of the planet was very similar to the earth. As a matter of fact, astronomers regarded Venus as something of a "twin" of the earth. Venus moves about the sun in an almost perfectly circular orbit (eccentricity is 0.0068 and is the smallest of the planetary system) and its distance from the sun is about 72% of the earth's distance. At close approach to the earth, it subtends an angular diameter slightly greater than a minute of arc, and its sidereal period of revolution is 224.7 days. It moves into close proximity with the earth once each synodic period which occurs on an interval of about 584 days. Its size and mass are close to that of the earth, having an equatorial diameter of approximately 7,800 miles and a mass of 0.815 that of the earth. Based on optical studies, the orientation of its polar axis and its period of rotation were not clearly established. More recently, the Doppler broadening of radar echoes observed at the Jet Propulsion Laboratory suggest that Venus may have a very slow rotation rate. It may be so slow that it keeps the same surface directed toward the sun at all times, much as the moon does with respect to the earth.

Venus is continuously covered with clouds, and there have been no direct photographs of its surface. In addition, optical and infrared determinations of the temperature of the planet have actually referred to the temperature near the cloud tops on Venus, or that altitude where the infrared opacity becomes larger than unity. Carbon dioxide has been clearly identified in its spectrum, and there is a possibility that water vapor has also been detected.

The infrared determinations of temperature led to values near 285°K, not an unexpected value for the higher altitudes on Venus.

The above information led to a number of "models" of the Venus surface environment. These range from a hot steamy tropical environment to a planet completely covered with an ocean. Although the models were substantially different, all were based on a thermal environment not unlike that of the earth.

The above picture of the planet Venus was completely disrupted when the first radio measurements in the microwave region were carried out at the Naval Research Laboratory by C. H. Mayer and his associates. Using the NRL 50-foot diameter radio telescope, the temperature of Venus was measured at wavelengths of 3 and 10 cm by microwave radiometers, and the astronomical community was greatly surprised to learn that the radio temperature was 600°K. Since the temperature is the same at 3 and 10 cm, evidence existed for a true thermal origin of the radiation. It seemed plausible that the 3 and 10 cm radiation would pass freely through the cloud cover at Venus and that the thermal source was in fact the "surface" of the planet Venus. But a temperature of 600°K is hotter than a broiling oven, and this demanded a most efficient greenhouse mechanism

to permit the surface to reach such a high temperature.

There is considerable disagreement in the theoretical arguments which have arisen concerning the Venus greenhouse mechanism, and it is not entirely clear that we can account for the 600°K. However, our knowledge of the physics of the atmosphere of Venus is most fragmentary. Particularly, our knowledge of the water vapor content, which is vital to the greenhouse mechanism, has practically no observational support.

Venus has now been observed from the infrared to the L-band region of the microwave spectrum. Observational data exists at the following wavelengths: 0.8 microns, 8.13 microns, 4 millimeters, 3, 10, 12, and 21 centimeters. The spectrum of Venus is presented in terms of the temperature measured at each of these wavelengths and is shown in Fig. 1. It is this spectrum which has completely altered our concept of the environment of Venus.

A number of models have been constructed to explain the spectrum shown in Fig. 1. Starting with a temperature at the surface of Venus near 600°K, it is possible to explain the shape of the spectrum shown in Fig. 1 in several ways. If the surface pressures are sufficiently high, carbon dioxide will provide a slight nonresonant absorption which is capable of attenuating the thermal radiation originating at the surface. With suitable assumptions, this effect can account for the shape of the spectrum. A second mechanism, again starting with a 600°K surface, is the attenuation of the microwave radiation by very small water droplets. For droplets of suitable size, the loss in energy due to scattering is small compared to the absorption, and the absorption leads to a reduction in the apparent brightness temperature of the planet at the shorter wavelengths. Both these models have their difficulties; the carbon dioxide nonresonant absorption process requires high surface pressures, and the water droplet model requires a substantial quantity of H_2O in the atmosphere of Venus. Based on present data, we cannot be confident of either of these models.

There is a third model which falls into this category of a 600° surface, which suggests that the surface of the planet Venus is covered by vast storms, and that the source of the heat is frictional. All of these models have their peculiar difficulties, but all have a common feature of a "surface" of Venus near 600°K.

A completely different model, the ionospheric model, can account for the spectrum shown in Fig. 1 through the postulation of a very rich ionosphere enveloping the planet. In this model, unlike the first three discussed, the surface of Venus can be more like that of the earth and is "comfortable" from the point of view of temperature. However, to account for the spectrum shown in Fig. 1, it is necessary to postulate an electron temperature of 600°K, and $\int n_e^2 dl \sim 4 \times 10^{25}$. With these assumptions, the ionosphere is opaque at the longer radio wavelengths and does not become thin until wavelengths near 2 cm. This model suggests that the electron density would be in excess of 10^9 per cm^3. An immediate difficulty of this model is the question of the maintenance of such a rich ionosphere. In addition, it is difficult to see how such a medium would permit the return of radar echoes which resulted from the JPL and Lincoln Laboratory measurements. Nevertheless, the ionospheric model does produce a theoretical spectrum which agrees quite closely with the observed data.

All of the models outlined above convey radically different visualizations of Venus as a planet. Ultimately, we will undoubtedly learn what the true conditions are. But part of our hope for realizing the true conditions on the planet Venus rests on spacecraft reconnaissance. We look forward within the decade to landing instrumentation on the surface of the planet. To some extent, we need to know what the true conditions on Venus are before we can adequately design the instrumentation which will be parachuted to the surface. The unusual importance of the interpretation placed upon the radio spectrum of Venus can perhaps be best emphasized by examining the requirements imposed on scientific apparatus which will ultimately be deposited on the surface. Suppose that the greenhouse model is correct, and that the surface temperature is actually hotter than a broiling oven. The scientific instrumentation must be designed in such a manner that "normal" operation will be possible in an environment that is exceedingly hot. While this would likely be possible, a development program costing time and money would be required.

On the other hand, let us suppose that the ionospheric explanation of the spectrum is correct. The instrumentation may be successfully parachuted to the surface of Venus and upon landing operate in a "comfortable" terrestrial-like thermal environment. However, the objective in programs of this type is to conduct scientific measurements and relay the information to man on the planet earth. This requires telemetering from the surface of the planet through the ionosphere. At this point the ionospheric model imposes a restraint, because no telemetry frequencies presently operating, nor none presently contemplated, will be able to penetrate an ionosphere as rich as that postulated for Venus. Telemetry wavelengths as short as one centimeter would be necessary.

Thus it is immediately clear that we must be able to resolve the true state of physical conditions

of Venus before we can intelligently plan for a direct scientific exploration on the surface of the planet. We may confidently expect that radio techniques will contribute toward the solution of this fascinating problem.

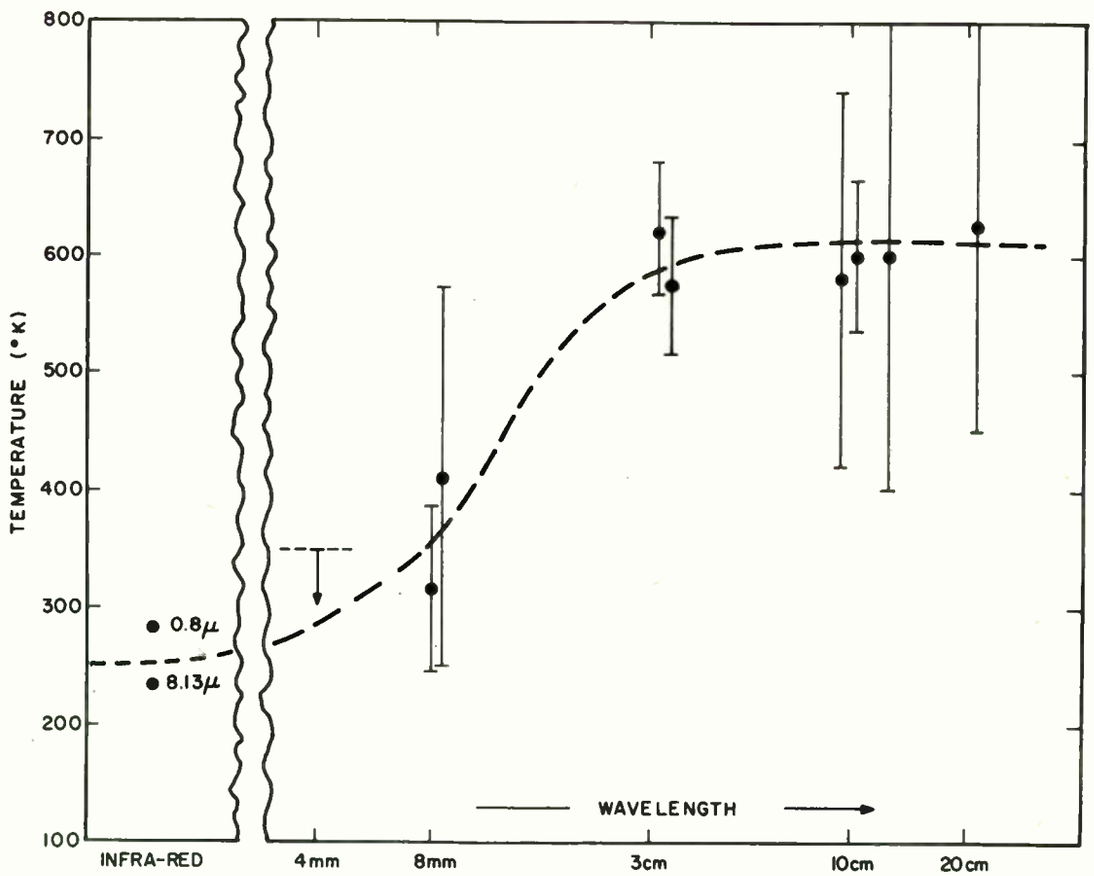


Fig. 1. Infra-red and radio spectrum of Venus. The approximate spectrum of the planet Venus is shown ranging from the infra-red to a long wavelength limit near 21 cm. Two measurements in the infra-red at 0.8 and 8.13 microns are shown followed by a lower limit observation of less than 350° K at 4 mm. The remainder of the radio observations range from 8 mm to 21 cm and define the variation of temperature with wavelength that constitutes the ground-based radio telescopic determination of the Venus spectrum. The dashed line is a free curve through the observational points. The flatness of the spectrum from 3 to 21 cm is regarded as observational proof of a thermal radiation mechanism.

INTEREST IN OUTER SPACE

L. V. Berkner, President
Graduate Research Center of the Southwest
Dallas, Tex.

Dr. Lilley, I am delighted to hear you admit that to ascertain what is on Venus, we have to send someone up there to have a look!

First of all I am not going to talk this evening about space communications. We have dealt with that subject for three successive conventions of IRE. I would be the first to admit that we have a few technical problems left in obtaining practical and reliable space communications, but I believe that if we face facts flatly, we ought to ask our political friends in Washington to move a little faster and we would have space communications in a hurry.

So I will speak of basic problems in exploration of deep space. After centuries of wondering what is really in the heavens, man has suddenly - in just the past three or four years - been able to extend his influence beyond the Earth to the whole Solar and Planetary System. Already we have begun by launched probes which are circling the Sun, or which have reached the moon.

You will note that I limit the sphere of man's influence to the Planetary System encircling our Sun. Within the foreseeable limits of physics, it is quite impossible to imagine any design for space craft that could travel significantly outside of our Solar System to the nearest stars as has occasionally been suggested. Based on the most sophisticated physical computations, the incredible energies required make interstellar travel quite impossible at any foreseeable time.

So beyond our Planetary System, we will still have to be content with the observational methods that have been discussed by Professor Lilley, and will be discussed by Dr. Bernie Oliver. But the developments of the last decade bring human exploration of the Moon, the Planets and their satellites - in fact the whole of our Solar System - generally within the reach of our practical technology in the immediate future. To explore the Solar System we simply require a space craft with its associated means of launching in the required orbit or space track having sufficient power and sufficient total energy. Such powers and energies are within our technical reach.

Essential, of course, to the success of such interplanetary space craft is a whole variety of electronic requirements. First we might mention navigation; second, control; third, approach control if the space craft is to approach the planet or some object in space, or to orbit it; fourth, landing control if it is to land; fifth, earth-planetary communications and telemetry through Dr. Lilley's very thick ionosphere, and finally collection, reduction and transmission of experimental data. For planetary exploration means scientific observation and exploration of the highest scientific character to ascertain what is there, how it came to be there and what are its changes with time.

As manned exploration proceeds (and you will notice I say "manned" because I believe unquestionably that to do a proper job of exploration, we must have men in interplanetary space vehicles in order to exercize and render judgment on the spot for the whole range of space activities ranging from military to scientific) - as each manned operation proceeds, many obvious electronic requirements must be realized to serve a whole variety of needs.

What are the real challenges of space to electronics? I would summarize them as follows: The first challenge is the new order of distance. Communication distance is no longer measured in thousands of miles, but rather in millions or tens of millions, or hundreds of millions of miles.

Second, in view of the length of time required by these expeditions, new orders of component and system reliability are imperative. The fellow who is going to have to travel days, weeks, or months at a time in a space craft is going to be quite interested in both component and system reliability.

Third, miniaturization and weight reduction of a new order will be required. It is now estimated that it will cost about $25,000 per pound to put a pound on the moon and bring it back again. This probably will come down by a factor of 10 or 100 in the next ten or fifteen years as more sophisticated and complex space trains are used that permit coupling or decoupling of exploration devices.

Fourth, we are reminded of the circuit performance and circuit sophistication that is required

to perform the multitude of tasks. Fifth, I would specify the electric power-supply of adequate capacity and low weight-power ratio and, finally, a high order of systems concepts with the proper relative allocation of weight, power and space among the methods of experiment, the methods of data reduction, the methods of encoding, and the methods of transmission all of which are interrelated and involved in getting the necessary data back to earth.

Consider distance: To overcome distance one suspects today that coherent radiation of very short wave lengths, let us say, of 1800 to 30,000 angstroms holds major real promise. Using antennas which are sized in the decimeter scale, the beam-width can be a few hundred meters across at the earth, for distances measured in units of earth-lunar distances.

So to conserve power we should think about beam-widths from a transmitter on the moon transmitting to the earth that are just a few hundred meters across, or beam-widths of just planetary size, over inter-planetary distances using antennas whose radii are measured only in the scale of meters. Thus very low power is needed for very high speed transmission over multi-million mile circuits. But at the Earth we require receiver sites with perfectly clear weather, or relay-satellites above the atmosphere.

We should be able to use the coherent radiation of the whole variety of molecular resonances (which we are now learning to exploit) with magnetic fields employing small magnets at the cryogenic temperatures - and generally, cryogenics is fairly easy to exploit, except on Venus or Mars, since the temperature of space looking away from the Sun is less than $4^{o}K$.

But, of course, very great stability and precision in directing the beam of coherent radiation is essential. Likewise, the perfection in reliability is going to be most critical. This is not only going to require ever improved concepts in component assembly, but also a highly educated circuit redundancy in the event of component failure since even the best components will fail and systems failure is not tolerable.

This sophistication of circuitry within the limits of weight, that are required, of course, implies solid-state circuits of very high efficiency and of relatively negligible weight. At $25,000 a pound you can afford some pretty good engineering. Even on manned expeditions, we must remember that the electronic circuits represent the equivalent competence of many men in the performances of the repetitive processes they can produce.

Electric power-supplies represent perhaps the weakest link in our whole space technology at the moment, and certainly will require much effort. Efficient converters of chemical to electrical energy, better solar cells, or perhaps nuclear power supplies, if they can be reduced in weight, are going to be required.

Above all, originality in system-organization and systems-design of our electronics systems should yield great rewards. This means, of course, the optimization of the total electronics system in the design and relative contribution of each element, from the experiment through data reduction, encoding, and the telemetry of the data.

The exploration of space is certainly man's greatest challenge to date. Such exploration is made possible only through advanced electronic systems which are the very nerve system and the brain of the complicated high speed inter-planetary space crafts that are now being planned.

But in return for great effort, I believe the potential rewards of space exploration are very great. Success in space gives us a measure of technological dexterity which is necessary for any nation which would stay in the Sun. But beyond this I would predict that through the new opportunities for science in space, perhaps merely from astronomy conducted from some orbiting satellite or from the moon, or perhaps the discoveries on the planets which provide new scientific vistas, may come indeed another revolution in science.

INTERSTELLAR COMMUNICATION

B. M. Oliver
Hewlett-Packard Company
Palo Alto, Calif.

Speculation as to the existance of extraterrestrial life is as old as the knowledge that other possibly habitable worlds exist. The early Greeks imagined that there might be beings on the moon. So did the astronomer Kepler and the science fiction writer Jules Verne. At the turn of the century the airless moon was no longer seriously considered a possible abode for life and attention turned to the other planets of our solar system, particularly Mars and Venus. As knowledge about the surface conditions of the other planets has increased, the possibility that intelligent life exists there has grown more remote. Today no one seriously expects to find intelligent life elsewhere in the solar system, though some feel that ammonia-based life forms may exist on Jupiter, and that there is probably primitive vegetation on Mars.

Oddly enough, as further knowledge about our solar system has shown it to be barren of intelligent life except (debatably) for Earth itself, further knowledge in other fields has greatly increased our estimates of the density of intelligent life elsewhere in the universe, on planetary systems around other stars. There are of course a tremendous number of stars—about one hundred thousand million in our own galaxy, the Milky Way. And there are on the order of one hundred thousand million galaxies. The feeling that in all this vast cosmos life must exist somewhere besides on earth is very appealing and plausible, but not scientific. The real questions are: How many stars have planetary systems rich in heavy elements? , and: Given a suitable planet, how likely is life to begin and to evolve?

Only a few decades ago, planetary systems were thought to be very rare, the result of near collisions between two stars. On this basis, only about one star in a billion would have planets. Today planetary systems are thought to be produced as a necessary step in the evolution of a typical star.[1,2,3] In order to condense to a small sphere, the whirling gas mass must lose most of its initial angular momentum. It can only do this by casting off from its equatorial periphery satellite gas masses and imparting angular momentum to these by magnetohydrodynamic interaction. As each satellite mass is thus accelerated into orbit the central mass contracts further and the process repeats, until the central mass becomes a star. The satellite masses then condense and cool to become planets and moons. In the solar system over 99% of the total mass is in the sun, while over 98 of the angular momentum is in the planets. Since the galaxy as a whole is rotating, typical gas masses must go through this process to condense into stars, and we conclude that planetary systems must be the rule rather than the exception. Several nearby stars show periodicities in their apparent motion which could be explained by the revolution of a planet of roughly Jupiter's mass.

The older stars, called Population II stars, formed out of almost pure hydrogen gas. It is unlikely that they produced stable planets, or if so, that life could begin on them. But these stars convert their hydrogen to heavy elements in the course of their life and when, in death, they explode as supernovae, they cast these heavy elements into space. There these elements mix with more hydrogen to form the gas masses out of which the second generation stars, so-called Population I, like our sun condense.[4] Planets of Population I stars are possible abodes for life.

A few decades ago the origin of life was a complete mystery and was thought to be highly unlikely unless precisely the right conditions were present. Today the right conditions are believed to be present in at least one planet of a typical Population I star.[5,6] The primitive atmosphere of a planet like the earth is believed to consist of ammonia, methane, water vapor and other molecular gases. Upon irradiation with ultra-violet light such a gas mixture forms amino-acids, the building blocks of proteins. On the primitive earth or any similar planet, these amino-acids would rain down into the then fresh ocean to form a literal consommé covering the planet. Given a few phosphate radicals it is not difficult to visualize the formation of large quantities of the molecules out of which deoxyribonucleic acid polymerizes. Those who have studied the origin of life now feel that, given a planet only approximately like the primitive earth, life is certain to start.

On a fair fraction of these it is likely to develop into complex and possibly intelligent forms.

Thus, during this century, while the probability of intelligent extraterrestrial life has virtually vanished for the rest of our solar system, it has grown enormously for other stellar systems. We now contemplate a universe teeming with life, but this life is so remote from us as to almost preclude our ever establishing physical contact. nearest star, α - Centauri, is four light years away. By contrast our solar system is only a milli-light year in overall diameter. While man may roam the solar system with instruments (or even in person) during the next decades, to explore thousands of light years into space looking for other life still seems impossibly difficult. The only means of contact which appears feasible today is some form of electromagnetic communication.

The problem of interstellar communication can be divided into three major aspects. There is the technical problem. Can we signal over these great distances and, if so, how can we best do it? There is the acquisition problem. How do we attract the attention of another race, or they ours? And finally there is the communication problem itself. How do we exchange meaningful information with a totally alien civilization?

The technical problem obviously depends on distance, and this depends upon the density of communicative races in our stellar neighborhood. A sincere effort to estimate this density was made last November at a conference at the National Radio Astronomy Observatory at Green Bank, West Virginia. On the basis of quantitative estimates of the numerous independent factors involved, the conclusion was reached that if we can signal out to ten light years our chances of contact are extremely slim, but that if we can signal out to 1000 light years (and do so for an extremely long time) our chances of contact are good. Stars likely to have habitable planets exist about 10 light years from us. Within a radius of 1000 light years there are tens of thousands of candidates.

The ultimate range of an interstellar communication system is determined by three factors: the energy radiated per symbol, the directive gain of the antennas, and the noise. Let us look first at the noise. Because a thermal noise source has the spectral power density of kT watts per cps (at frequencies, ν , such that $h\nu << kT$) it is convenient and customary to consider any noise source having a spectral power density, $\psi(\nu)$, as equivalent to a thermal source having the temperature $\tau(\nu) = \psi(\nu) / k$. Figure 1 shows the effective noise temperature for three sources of noise germane to the present problem.

Cosmic noise is received from all parts of the sky but is most intense toward the center of the galaxy. [7] It is the noise which Jansky [8] first heard and which started the science of radio astronomy. It falls off rapidly with increasing frequency as indicated by the line in Figure 1.[9]

Even an ideal amplifier, one without any thermal noise, is not noise free. Quantum effects produce a noise referred to its input of $h\nu$ watts per cps, giving an effective temperature of $h\nu / k$ as shown. [10] (Photomultipliers show a similar fluctuation due to randomness of photon count.) Modern masers approach this ideal quantum limited performance.

The total noise from these two sources reaches a minimum around 8 kmc of about 0.5° Kelvin. This would be a good frequency for interstellar communication if our receiver itself were in space. With an earth based receiver the atmosphere radiates noise into the antenna due to absorption (and re-radiation) of energy by oxygen and water vapor.[11] Taking this into account the quietest frequency range is from about 2 to 10 kmc with an effective noise temperature of 5° to 10° Kelvin. Above 50 kmc the strong oxygen, water vapor, and carbon dioxide resonances render the atmosphere opaque. When it clears up again in the infra red, $h\nu/k$ has reached several thousand degrees. Thus we have two likely regions for communication, one in the infra red and optical part of the spectrum, and a much quieter one in the microwave region.

The advent of optical masers, or lasers, has focused interest on the possibility of using optical frequencies for interstellar communication.[12] Very large directive gains are possible with optical "antennas" of modest size. However, beam widths much less than a second of arc are impractical because of atmospheric turbulence and aiming problems. Such beams do not provide enough directive gain over present large microwave antennas to compensate for the increased noise at optical frequencies (≈20,000° K). Thus the microwave region still seems best at the present state of the art.

Figure 2 shows a performance comparison of two optical systems and one microwave system. The ordinate is the number of photons, $\overline{\eta}$, received per pulse for the optical systems, and the ratio of received energy per pulse to kT for the microwave system. Each of these quantities gives the signal to noise ratio for the system involved, so the ordinates are comparable. The present laser with 10 joules per pulse at a wavelength of 0.7 micron, and equipped with 10" telescope "antennas" at both transmitter and receiver, would give the performance shown by the left line. An ideal quantum limited receiver is assumed. We see that the signal to noise ratio falls to unity

at a distance of 0.04 light year. Raising the radiated energy to 10,000 joules per pulse gives the performance shown by the middle dotted line: unity SNR at a little over 1 light year. The line on the right assumes a microwave system radiating 10^4 joules per pulse at 21 cm, with 600 foot transmitting and receiving antennas, and a receiver noise temperature of 10°Kelvin. We see that unity SNR is reached at 100 light years. At ten light years the SNR is 20 db. As a matter of fact 10^6 joules per pulse (one megawatt for one second) is certainly possible today, and this would give a 20 db SNR at 100 light years, 0 db at 1000 light years. The same results would be obtained with 230 foot antennas at X-band.

We may conclude that the technical problem is barely within the present art, assuming equally dedicated cooperation at the other end!

What about the acquisition problem? This is made difficult by the dimensionality. Whether we transmit or listen, or both, we must not only search in space, star by star, but also in time, over an extended period, and possibly in frequency as well. If we could eliminate one dimension of search, we could simplify the problem enormously.

If we (or the other race) had some means of generating extremely short pulses of extremely high power, say 10^{16} watts for 10^{-10} seconds, we (or they) could blanket the best part of the microwave region with energy. A very broad receiver sensitive from about 2 to 10 kmc, would then be indicated and would not require searching in frequency. Barring this, the high pulse energy can only be achieved by using the highest available power for long pulses and this requires a narrow band receiver.

Frequency search can also be eliminated, or greatly reduced, by transmitting on or near some natural frequency such as a spectral line. This is an inherent property of lasers and one of the strong arguments for their use. Project Ozma began with the ingenious suggestion of Cocconi and Morrison of Cornell [13] that we listen on the hydrogen line frequency of 1420 mc (21 cm). They argued that since radio astronomers listen on this line a great part of the time, radio astronomers anywhere in the universe would be doing the same, and any intelligent race trying to communicate would choose this frequency or one nearby.

A good acquisition signal should differ markedly from all natural signals. It need not be a series of prime numbers —mere pulses will do— but it should not be so efficiently encoded in the Shannon sense as to resemble thermal noise. The acquisition signal can convey instructions for locating and receiving another signal more efficiently encoded, but the acquisition signal itself should be attention getting. The acquisition problem needs further study and should be reviewed in the light of each new technological development.

How should we attempt to convey meaning? Since the nearest likely stars are 10 light years away, no answer can be received for at least 20 years. Certainly we should not waste time sending simple pulses till we get a reply. The other race would probably consider us so stupid they might not bother to answer. We should, I believe, send a repeated series of messages which constitute a course of instruction—facts about our civilization, our language, our science, and ourselves. The other race can then respond with a much more sophisticated initial message, perhaps one which would indicate some more efficient means of communication.

In constructing these first acquisition messages we are justified, I believe, in making one very important assumption about any intelligent race: that they have eyes. Sight is such an important sense that it is hard to imagine technological development in a race devoid of it. Further, it has developed in a great variety of earth's life forms. Flies have eyes, so do scallops, and mice and men. Since pictorial matter provides a good, if not the best, means of conveying meaning in the absence of a common language, the initial signals should probably be pictorial. Other races would probably decide on pictorial messages for the same reasons.

With this in mind, let us assume we have made contact! After years of futile listening we receive a peculiar series of pulses and spaces from ϵ-Eridani. The message is repeated every 19 hours and 53 minutes, apparently the length of their day. The pulses occur at separations which are integral multiples of a minimum separation. Writing ones for the pulses and filling in the blanks with the appropriate numbers of zeros we get the binary series shown in Figure 3. It consists of 1271 ones and zeros. 1271 is the product of two primes 31 and 41. This strongly suggests that we arrange the message in a 31 x 41 array. When we do so, leaving blanks for the zeros, and putting down a dot for each pulse we get the non-random pattern of Figure 4.

Apparently we are in touch with a race of erect bipeds who reproduce sexually. There is even a suggestion that they might be mammals. The crude circle and column of dots at the left suggests their sun and planetary system. The figure is pointing to the fourth planet, evidently their home. The planets are numbered down the left hand edge in a binary code which increases in place value from left to right and starts with a decimal (or rather a binary) point to mark the beginning. The wavy line commencing at the third planet indicates that it is covered with water and

the fish-like form shows there is marine life there. The bipeds know this, so they must have space travel. The diagrams at the top will be recognized as hydrogen, carbon, and oxygen atoms, so their life is based on a carbohydrate chemistry. The binary number six above the raised arm of the right figure suggests six fingers and implies a base twelve number system. Finally, the dimension line at the lower right suggests that the figure is eleven somethings tall. Since the wavelength of 21 cm on which we received the message is the only length we both know, we conclude the beings are 231 cm, or seven feet, in height.

This message certainly does not exceed Shannon's information limit. Better messages from this standpoint can easily be constructed. In fact the present message is an extension of an earlier one constructed by Frank Drake, of the National Radio Astronomy Observatory,[14] which used only 551 zeros and ones to convey a somewhat similar amount of information. But it is surprising how much we could learn about another race from only 266 pulses.

The likelihood of other intelligent life and the virtual impossibility of physical contact with it combine to make the study of interstellar electromagnetic communication a fascinating and possibly fruitful subject. To signal over the required distance seems technically possible. The communication of meaning appears easy. If we solve the acquisition problem (and the economic problem of mounting the sustained effort required), electronics may some day answer the age-old question: Is man alone in the universe?

References:

1. O. Struve, "Stellar Evolution," Princeton, Chap. 2, 1950.
2. H. Alfven, "Origin of the Solar System," Oxford, 1954.
3. G. P. Kuiper, Astrophysics, Chap. 8 (J. A. Hynek, ed.), McGraw-Hill, 1951.
4. J. H. Oort, Stellar Populations (D. J. K. O'Connell, ed.), North Holland Publishing Co., Amsterdam, 1958, p. 415.
5. C. Sagan, Radiation, 15, 174 (1961).
6. A. I. Oparin et al., eds., "The Origin of Life on the Earth," Pergamon, 1959.
7. H. C. Ko, Proc. IRE, 46, p. 208, Jan. 1958.
8. Karl G. Jansky, Proc. IRE, 21, p. 1387, October, 1933.
9. J. H. Piddington, Monthly Notices, Royal Ast. Soc. 3 (1951).
10. M. W. P. Strandberg, Phys. Rev. 106, p. 453, Mar. 1961.
11. D. C. Hogg, J. Appl. Phys. 30, p. 1417, Sept., 1959.
12. R. N. Schwartz and C. H. Townes, Nature 190, 205 (1961).
13. G. Cocconi and P. Morrison, Nature 184, 844 (1959).
14. Private Communication to Members of The Order of the Dolphin.

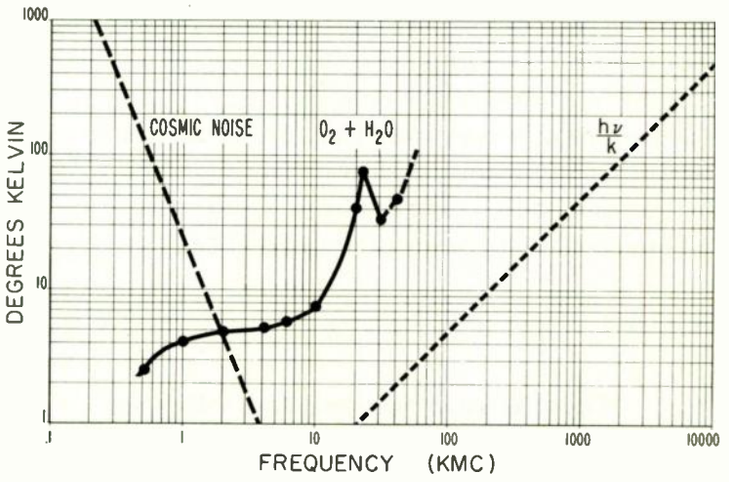


Fig. 1. Equivalent noise temperatures from various sources.

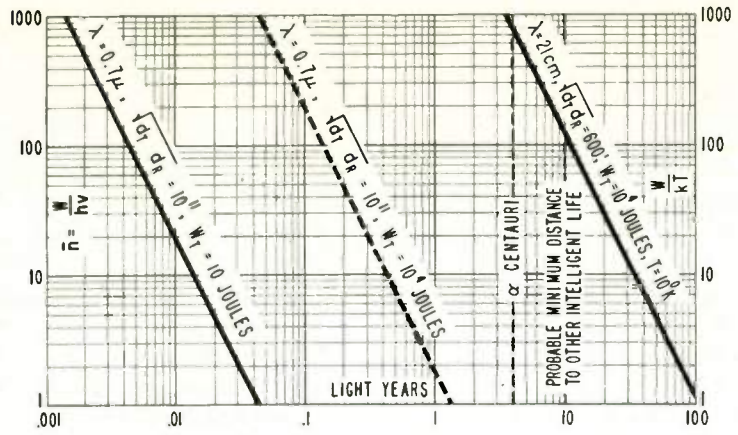


Fig. 2. Comparative DX performance of laser and microwave systems.

266 PULSES - 1271 BITS - WHAT DO THEY TELL US?

Fig. 3. An imitation message from outer space.

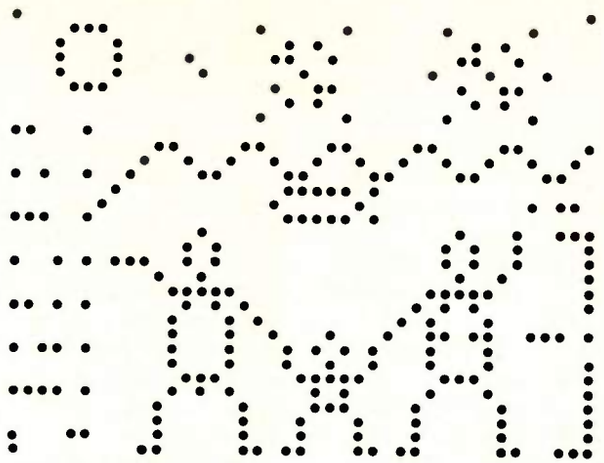


Fig. 4. The same message rearranged. (We may have it reversed left to right, but we will never know unless we compare physical parity experiments with the other race or star field visible to both of us.)

COMMUNICATION DEVELOPMENTS IN LIFE SCIENCES

Lee B. Lusted
Oregon Primate Research Center
Beaverton, Ore.

I would like to talk about some communication projects for the biomedical sciences which I hope will keep us all healthy while we are eagerly awaiting that next message from outer space.

I have two subjects that I would like to discuss and, to begin with, I won't go into much detail, but I will try to outline the subjects.

The first thing I would like to talk about is the development of a family of computers for use in the biomedical sciences. One of the things that is interesting about the biomedical sciences is that this field is data rich and theory poor. This means that it is easy to accumulate a large amount of data to work with, and frequently there is inadequate theory for handling this data.

What is urgently needed in the biomedical sciences is a family of computers modeled along the lines of the laboratory oscilloscope. I think that this is a new idea for the life sciences.

These computers, I would like to suggest, should be of several sizes: From very small machines for individual laboratory use to extremely large-scale computers of great speed and capability, with the emphasis on data manipulation and flow rather than on computation and precision. The machines in the biomedical laboratory must control as well as process data under a variety of conditions where there is a great deal of data, and by this I mean that the computer is actually a part of the experiment, so that the experimenter can use the computer itself to help develop the experiment as it goes along.

This means that the computer needs a great variety of peripheral equipment; and in particular devices are needed that may not be very readily available now, particularly for smaller computers. For instance, it appears that a cathode ray tube is a very useful device to put on equipment to be used for processing biological data. The biologist is used to looking at this kind of display for his data, and he can use the scope as a quantitative as well as a qualitative element.

We need optical input and output devices; we need analogue signal convertors; we need magnetic tape in very many different kinds of forms, and we need a full range of manual controls. These things help the investigator manipulate the equipment during the experiment.

I believe that only with such a complete range of computing and data-processing machines can the biomedical scientist who is increasingly aware of the need to quantify, structure and reduce his data—and he is oftentimes very ill at ease with mathematical procedures—I believe only if we provide him with various types of computing equipment adapted to his needs—will he be able to make much progress in the quantification in the life sciences.

Now I would like to discuss the second subject which concerns a computer based health information processing system. The reason that I think it is very important to talk about a health information processing system at the present time is because there is a shortage of physicians developing in this country, and it interests me that when the physician shortage and what we should do about it is discussed, I have not seen any comments that we should provide the physician and the hospitals with an information processing system which would help them carry on patient care more effectively. I would like to outline some of the medical background data, about why this problem is important and later we can talk about some things that could go into the development of this kind of a system.

There are at least three areas where data processing systems can help with patient care. One of these areas is medical records. Physicians need readily accessible records. The second area is medical literature in readily retrievable form. Lastly, a hospital information processing system is needed so that the laboratory, the nurse's station, the interns, etc., and get the information they need as readily as possible.

Let us look at these areas briefly. A study of what interns do during their working hours in a teaching hospital here in the East, showed that they spend from one to four hours a day writing in patient's charts, whereas they spend about ten minutes a day with each patient. There ought to be some way to cut down on the amount of charting that the interns have to do. A study of what nurses do during their working hours showed that they spend a sixth of their time in copying medication

orders. Some kind of data processing system might help with this problem.

What about the problem of medical records? A large city hospital found, in studying its records, that it had records of a million people in the file, and that the records of about 100,000 new patients a year were being added to this file. To store this patient data in a computer memory, a random access storage of 140 million alphanumeric characters would be required. Also, it has been found that, because the same patients may be seen in a number of different hospitals in the city, there is considerable medical record overlap. This overlap has been estimated at about 35 per cent among various hospitals. This means that if a medical information processing system could be developed so that the hospitals could communicate with each other on patient records, for instance through a central filing system, then this system should help hospitals to operate much more efficiently.

What about the problem of medical literature? The National Library of Medicine is installing a computer-based system for medical literature searching. This system, called MEDLARS, should go into use in fiscal 1964. It is planned to punch something like 250,000 characters a day at first, and by 1968 it should be punching something like 400,000 characters a day.

This system specifically will be to help prepare and to print the Index Medicus. Index Medicus lists 125,000 papers a year which have been culled from 1600 medical journals. But the point is that this is now only about 15 or 20 per cent of the medical articles that are now being published.

I am proposing that a health information processing system is very much needed to supply information on medical literature and patients' records. This system should be developed first for one hospital, then expanded to several hospitals in a city and finally to hospitals in a region. It is timely to do this now because of the pressure for increasing the amount of medical care in the United States.

In closing, I want to mention a quotation which I saw recently in an electronics journal which said: "Electronics has been through five major areas: radio, radar, television, industrial and missiles. We are now in the beginning of a newer age, the medical age."

Note: For a more complete explanation of various aspects of health information processing systems and for some proposed solutions, see "The Application of Computers in Diagnosis," by Lee B. Lusted in Circulation Research, September, 1962.

THE IMPACT OF NEW TECHNIQUES ON TELE-COMMUNICATIONS IN THE NEXT TEN YEARS

Richard R. Hough
American Telephone and Telegraph Company
New York, N. Y.

I am glad we were brought somewhat back down to earth again, because I am going to try to stay on the ground even though today it seems as if the telecommunications business is somewhat in orbit.

My subject is: "The impact of new techniques on telecommunications in the next ten years." It is pretty difficult to try to look ahead that far because your vision gets somewhat hazy. But I think that progress being a continuous function—and this is particularly true in tele-communications—it is very helpful to look a little bit behind and obtain some landmarks, so that we can project into the future. The way we are moving today is an indication—a very strong indication—where we are going. Although you can point to individual technical break-throughs that have been of tremendous importance to tele-communications, their application has been more evolutionary rather than revolutionary. This has been, of course, in accord with the need and the economics of the situation.

I think there are really three factors to be considered when we talk about telecommunications progress. These are: quality, quanity and utility, or what you might call convenience.

Speaking about quality first, some 30 years ago, a telephone conversation was equivalent to talking to someone about 35 feet away, and the naturalness in addition to the volume suffered in the same proportion. Today, it is essentially face to face.

This improvement, of course, has come about through refinement of the vacuum tube, the invention of negative feed-back and more recently the transistor as well as improvements in the tele-phone instrument itself.

When you talk about quantity, I think the progress there is rather obvious. In 1920, there were about 13 million telephones in this country. Today they are in the neighborhood of 75 million. In 1920, there were about 17 billion telephone conversations, and last year there were approximately 97 billion. So the history has been one of very substantial increase in quantity. In addition to this, you must add other services such as video transmission, private lines, and TWX.

From the standpoint of utility, in addition to interconnection of all telephones in this country, some 98 per cent of all the world's telephones are now accessible from telephones in this country. There has been a tremendous increase in the variety of services: The use of the switched telephone network for data communications, the availability of broad chunks of band width for communication purposes, and new services which provide wide area, flat rate telephone service and data service for customers who have large volumes to scattered points.

Very important, when you speak of utility, is the matter of cost. Services do not have utility until the cost is within reason for the service performed. As we look to the future, I think it is safe to say that future progress is going to be along the same lines of quality, quantity and utility, and I think it is significant that, as all facets of business and government grow and expand, the need for more and better communications seems to go up at a substantially faster rate.

Not only do better communications stimulate more use, but also the need is increasing rapidly, and this stimulates the rapid development of communication facilities.

I think the most significant development in recent years, and one that is going to have, in my estimation, a most far reaching impact on telephone communications for the future, is the rapid development in the field of data handling and processing. Tremendous increases in the amount of data and information that can be digested and utilized—and it seems the more we are able to handle the more we have to handle—enables us to do, in all facets of business and government, many things that we either could not do previously or at least could not do with the assurance of thoroughness and accuracy and to the depth that we can do today.

As we increase the use of sophisticated data processing, information handling equipment, then the need for greater and greater volumes of communications grows by leaps and bounds. The decentralization of our industrial activity that has been going on since World War II increases the need for communications. As our industrial world makes more use of data processing equipment,

general purpose computers, while their management operations are decentralized, their processing of data tends to be centralized. Because, if you can do this data processing in larger chunks—well to put it simply, you get more "bits per buck," and this is what is happening. It is having a major impact on the requirements for tele-communications and is going to be, in the years ahead, a most important factor with regard to the development of communications services and instrumentalities.

How is this affecting the three factors that I spoke about which have been the measures of advances in tele-communications in the past? First of all, from the quality standpoint, when we say a telephone conversation today is essentially equivalent to face-to-face, you might wonder what is there to obtain in the future from the standpoint of improved quality?

Well, it is a little bit frustrating to telephone engineers because we no more than begin to do a fairly decent job from the standpoint of voice communications, and we find this great demand for data communications which imposes a much more severe requirement on the quality of transmission.

Noise and hits on the line which went by unnoticed in voice communication, represent substantial errors in data communication. And the faster you transmit the data, the more serious this is.

Also, as we are required to transmit larger volumes of data we need broad continuous band widths. We also have switching problems; some data systems require very short messages and the switching interval becomes a significant factor, as opposed to telephone switching.

Looking at overseas service, some of the same requirements that we find in domestic communications are becoming more and more important, both volume-wise and quality-wise, for overseas. Hence, the advent of the undersea telephone cable which provides the quality of a communications that we are able to obtain on land lines. As we look to greater volumes to more locations around the world, satellites will come into the picture.

With regard to quantity, it is obvious that, as we start to handle large volumes of data, then we must have more circuits and all this means really broader bands, more frequency space. While, fortunately it looks as if the voice band will be the mainstay of data communications in the future, just as it is for voice communications, there will be increasing needs—and there are increasing needs—for broader and broader band widths over and above voice bands.

As these demands for larger volumes of communication services increase, it is going to become more and more important to pool facilities. Our switched telephone network is something that we have available to us every day. This is a great pool of facilities which we all share and, as we require greater and greater volumes, it is most important that developments go forward on this common sharing of circuits and switching equipment.

More than in any other part of the business, we do have the tools to meet this demand for greater volume. It seems that the ability to provide larger volumes of communications over long distances more cheaply has kept somewhat ahead of the demand and, as you might expect, today we have facilities either available or under development that will meet the needs of the future as we see them.

A new coaxial cable is going across the country now which will carry approximately 9,000 voice circuits. The latest microwave system used in transcontinental service will carry 11,000. Using a 2-inch wave guide in the 50 kilomegacycle range with time division channelizing techniques, somewhere in the neighborhood of 200,000 telephone conversations could be carried in a single wave guide, or, if you like, 200 TV channels.

With the advent of the optical Maser, great new possibilities are opened up for large band widths and great volumes—circuits in the millions of equivalent voice band widths.

It seems that we have the tools to provide the long distance distribution of the great volumes of data communications that will be required. And as we look ahead, I am sure we will find that there will be more chatter between machines over our circuits than between people.

From the standpoint of utility, as this volume increases, there is a great need for a wide range of services to be available and, just as difficult as the technical problems of providing the instrumentalities, are the problems of shaping the rate structures so that the customer can obtain the service he needs in the way he wants.

We have made a lot of progress in this recently. Some of the new services mentioned are the results of trying to tailor the services to meet the wide variety of needs of many customers. As we look further ahead, as the public travels more and more, on the move a greater percentage of the time, there is going to be an increasing need for mobile communications and eventually personal telephones that you can carry around with you.

The same things, the technology and the techniques, that are making possible the handling of tremendous volumes of data are very promising for solving many of the problems of the tele-communications industry. For handling these volumes of communications, the solid state devices and techniques, digital techniques, enable us to do more in less space with less power with longer

life and less cost, and these are the things that are required in order to make large volume communications practical and economical.

Consider station equipment, for example. There are a lot of things that are technically possible to do, but have not in the past been economically feasible in the great volume of station equipment. With the availability of transistors and with their reliability increasing in recent years, and prices going down, it becomes feasible to put kinds of circuitry in the telephone instrument that you would never dream of before.

One of our most serious problems, and one that we really don't have any answer to yet, is the matter of distribution. You can talk of the use of the optical Maser to provide great volumes of circuits for interconnecting central offices and between towns. But when you come to get relatively broad bands to the individual home, then we have some real problems and we look forward to the day when some of the new techniques not only solid state devices such as transistors and diodes, but some of the new solid state circuits that can be manufactured as a unit, providing the means to make the distribution of broader bands more economical. Along with this will come the possibility of increased volumes of data and television telephones in the home.

We look for these same techniques to make economical the provision of mobile telephone service for a wide range of customers at rates that can be afforded and in the volume needed.

The only conclusion to be reached is that the future of tele-communications is one of the continuing and, in fact, accelerating improvements in quality, quantity and utility.

PLANNING SYSTEMS FOR MILITARY COMMUNICATION

E. G. Fubini
Office of the Secretary of Defense
Washington, D. C.

First of all, let me express my pleasure to be at a meeting discussing advanced communications, where communication satellites have not been mentioned. I will break the rule only to say that I am completely in agreement with the fact that we had better stop talking about them and go ahead with the work.

Also, I want to assure Lee Lusted that the problem he has in the medical profession is nothing compared to the ones we have. Now let me come back a little bit more seriously to my problem.

As you know, our defense is being planned so that we are not going to have only one red button; we are going to have a large number of buttons which go from the deep red to the very pale pink. This imposes a new dimension to command, and I assure you imposes a number of requirements which are, in many ways, staggering.

Our present planning assumes a flexibility and an action of command, the like of which has never been requested of our defense establishment. First of all, our communications must survive; it is part of our policy that we must be able to absorb the first blow. The Secretary of Defense has stated recently that we can absorb the first blow, and if the enemy hits us we can retaliate and destroy him because we are that much stronger. But in order to do this, command must survive, and our communications must survive.

How do we go about it? We must, first of all, be able to make communications heard. Second, we must provide diversification and third we must provide mobility; we are using all these schemes. We have established airborne command posts which are on the alert continuously so as to provide not only the survivability of an airborne platform, but also the longer range communication which will be supplied by the higher altitude.

Also, we have provided relay airplanes to transmit communications all over the country, and these will help provide the survivability we are looking for. We are also going underground and the NBC television programs showed the Cheyenne Mountain example for the Norad Command Post.

We are also providing ship-based command posts with capability to survive and to provide command and communications. Consideration has also been given to trains for the same purpose.

Problems are involved not only in establishing these communications but also in determining what kind of information goes with this communication (this, by the way, is much more difficult than establishing the communications).

The chatter between machines which Dick was talking about is here. The trouble is this: The machines are designed a little like humans; one machine does not listen to what the other says.

The design of the command post is, in many ways, the same problem that Lee Lusted was talking about regarding his doctors, but much bigger. Our problem is that we must have a machine that takes hundreds of thousands of bits in and lets very few bits come out.

The process of doing this is very complicated. Lee Lusted was talking about a million people in a file with 8,000 new patients and 140 million characters. I assure you this is really very little when seen from our point of view. We go by data bases, and one data base may have 60 million characters. It is just par for the course, and we have at least—I don't know—40 or 50 of these data bases. And the problem is not only to put information in: this is not too difficult; it is to get data out, and that is very much more complicated.

Now, lest we believe that automation is the solution to all problems, I would like to establish that machines can and will and should, but they don't yet, analyze data and display important points and permit the man who has the information to lead, to command.

As far as the problem of communication goes—it becomes essential to decide what to send and what to do with what you send. We must avoid the tendency to send everything; you know there is such a thing as records catching dust and systems that are useless because they have too many data.

So don't forget, don't send too much information if you cannot use it. As a matter of fact, I also say don't transmit any information which you cannot use. But that perhaps is a little bit too radical.

Now to summarize very very quickly; we need flexibility because that is our national policy. We

need survivability of command because, without command, we cannot act upon a changing situation. In order to survive we must provide mobility and hardening. We must decide on information flow, we must provide quick absorption of the information displayed for decision-making, which cannot be made automatically. These are the difficult problems with which we are presented.

ERROR BURST CHAINS IN DATA TRANSMISSION

Pierre Mertz
The RAND Corporation
Lido Beach, New York

SUMMARY

Errors in data transmission do not occur at random, but come in bursts, and the bursts themselves also show "bunching" rather than complete randomness. This has been described with hyperbolic rather than classical Poisson laws. The error bursts form into occasional chains. Such chains are highly improbable in an all-Poisson distribution. Their expectancy has been determined under assumptions that within a chain the distribution is Poisson, but the long-time distribution is hyperbolic. The results show the moderate probability of the chains observed in experience. An estimate is also given of the longest expected chains in tests of given durations.

I. INTRODUCTION

It has been found in data transmission that the occurrences of errors are not completely random, but that they tend to "bunch."(1)

This "bunching" can be described mathematically, with a reasonable degree of accuracy, in terms of hyperbolic laws. The occurrences of the bursts follow these fairly faithfully, and are not described by the more conventional Poisson laws. It has further been found experimentally that, as part of this bunching, the error bursts tend to form occasional chains of fairly long duration.(2) With a completely random distribution of bursts, such chains would be of so low a probability that they would have insignificant expectancy in even long tests (say one or two successive months).

In this paper the theoretical expectancy of such long chains is examined for the case of a hyperbolic distribution. Some predicted distributions of burst chains under the various conditions and the longest expected chains in overall tests of given durations are plotted. The former are compared with some reported experiments.

Review of Statistics

Briefly reviewing the Poisson statistics, one notes that the probability $P(a,c)$ of at least c bursts in a time interval for which the long-time average is a, is(3)

$$P(a,c) = 1 - (1 + a/1! + a^2/2! \ldots + a^{c-1}/(c-1)!)\exp(-a) \qquad (1)$$

The probability $Q(a,u)$ of a burst-free interval of u time units duration is(4)

$$Q(a,u) = \exp(-au) \qquad (2)$$

It is shown in Appendix A that the longest such interval, u_o, expected in a testing time of T time units is expressed by the implicit relation

$$1/(aT) = -Ei(-au_o) \qquad (3)$$

where $Ei(x)$ is the exponential integral function tabulated by Jahnke and Emde.(5) It is also shown in Appendix A that the shortest burst-free interval, u_1, expected in the testing time T is expressed by

$$1 = -Ei(-au_1) \qquad (4)$$

In the analogous hyperbolic statistics the probability $P(a,c)$ of at least c bursts in a time interval for which the long-time average is a, is(1)

$$P(a,c) = a/(Ac + a) \qquad (5)$$

where

$$A = (\log_e T) - 1 + (1/T) \qquad (6)$$

and T has the same meaning as in Eq. (3). The probability $Q(a,u)$ of a burst-free interval of u time units is(1)

$$Q(a,u) = A/(A + au) \qquad (7)$$

In Appendix B it is shown that the longest burst-free interval expected in a testing time of T time units is expressed by the implicit relation

$$A/(aT) = N(U_o) \qquad (8)$$

where

$$N(U_o) = (\log_e \frac{1 + U_o}{1}) - \frac{1}{1 + U_o} \qquad (9)$$

and

$$U_o = u_o a/A \qquad (10)$$

The shortest burst-free interval expected in testing time T is expressed by

$$A = N(U_1) \qquad (11)$$

where

$$U_1 = u_1 a/A \qquad (12)$$

Definition of Error Burst Chain

The exact definition of an error burst chain is really a matter for considerable thought. The definition must be realistic, but it is also desirable that it lead to mathematics which can be easily handled.

Although an error burst chain is somewhat analogous to a burst-free run, its definition is far more uncertain. The important difference is

that a burst-free stretch is precise. Just one error in a stretch nullifies it as burst-free. However, in an error chain, one good bit does not nullify it as such a chain. In fact, if every single bit were in error, the case would be merely that of an upset signal (i.e., marks and spaces interchanged). Thus one must decide on how many good bits are to be assumed to separate error burst chains.

For the purpose of this paper a burst chain is defined as a succession of error bursts so close together that the gap between burst initiations is not greater than some critical time interval. Where the gap is more, there two separate chains. In part of the study this critical gap is fixed (illustratively at 100 bits) independently of the length of the chain which it follows. In later parts of the study the critical gap varies with the length of the preceding chain.

General Procedure of Analysis

A rather round-about procedure is necessary. Because of the bunching one cannot assume freedom from correlation between successive bursts. Thus one must use only the appropriate formulas reviewed above.

The procedure for arriving at a distribution of chain lengths from the general hyperbolic law comprises a series of steps listed below. Some of the concepts are shown in Fig. 1.

1) Assume a given interval of time, taken to be a possible error burst chain duration, as T'. It is convenient to measure this in bit durations. Primed letters are used to distinguish the interval T' from the later over-all interval of the test duration T.

2) Assume a distribution of burst-free intervals in T', the intervals being measured in bit durations. Call the longest such interval u_o', and the shortest u_1'.

3) The maximum burst-free interval u_o' is a function of the average burst rate a', during the assumed chain duration T'.

4) Choose a criterion v', for the maximum burst-free interval in the chain, u_o', to be equal to or less than this.

5) The inequality in Step 4 fixes a', in view of Step 3.

6) Consider the assumed chain duration T' as an element of unit time in the entire test T. The average burst rate a' during T', times T', can be taken as a short-time burst rate c, per unit interval T', during parts of the entire test T.

7) Assume a long-time burst rate for the entire test as a (measured per unit time T').

8) The hyperbolic formula for P(a,c) gives the probability of the short-time burst rate c (or greater) when the long-time rate is a. This tells how much of the time T is occupied by intervals T' having a burst rate of at least c, and therefore how many such intervals there are.

9) The operation can be continued with assumed burst chains T' of different durations. The result is a distribution of burst chain durations in the test T.

II. SOLUTION OF BURST RATE IN CHAIN LENGTH

Step 5 of the general procedure is the solution of an inequality to determine the parameter a'. The components of this inequality are obtained from the equations in the statistical review.

It is desirable to explore the possibilities both of a Poisson and of a hyperbolic distribution of bursts in the interval T'. For the former one uses Eq. (3) and for the latter Eq. (8). When these are solved as equations, the solutions indicate the boundaries of the inequalities of Step 5. It is clear that in both cases what must be solved are implicit transcendental equations in the variable a'. The consequence of this is that the solutions cannot be expressed in simple analytical form. Instead, they must be obtained by such procedures as graphical solution or successive approximations. This has a profound influence on the course of the investigation from this point on. It means that it will be necessary to study specific illustrative examples, rather than generalizing through simple formulas.

The method of graphical solution can be briefly outlined. In Eq. (3) one takes the ratio R of $u_o' = v'$ to T':

$$R = u_o'/T' \qquad (13)$$

Then, with $x = a'u_o'$,

$$R/x = -Ei(-x) \qquad (14)$$

In Fig. 2, -Ei(-x) is plotted against -x. Also R/x is plotted for various values of R, as a set of diagonal fine lines. The solution for -x is obviously the crossing of these two functions for the appropriate R.

In Eq. (8) one similarly defines R, and obtains

$$R/U_o = N(U_o) \qquad (15)$$

This is in the same form as Eq. (14) and one can use the same diagonal lines for R/x and for R/U_o.

One can make some quick comments on the nature of the solutions that will be obtained. For both distributions there is no solution for R too close to 1. For the Poisson curve the limit is at R = 0.32, and for the hyperbolic curve, at R= 0.20. For both distributions there are two solutions. Here we are primarily interested in long burst chains, and consequently take the solution for the greater x or U_o, and ignore the other.

Having solved for x or U_o, and having set in advance T' and u_o', we can determine a' from Eq. (10) or just following Eq. (13). This represents the long-time average bit rate (over intervals T') that give a gap just at the boundary $u_o' = v'$. The quantity a'T' gives the expected number of bursts in the interval T'. In the procedure at Step 7 this is to be used as the short-time burst rate during particular intervals T', out of the over-all test T.

Specific Example, Chain Length Equal One Minute

A specific example may be a bit rate of 1000 per second, and for the moment, a chain length of one minute, or T' = 60,000 bits. Another necessary specification is the gap criterion for the chain, which is taken as 100 bits, so that

$$u_o' = v' = 100 \qquad (16)$$

and

$$R = 100/60{,}000 \qquad (17)$$

$$= 0.00167 \qquad (18)$$

This can be solved graphically for both the Poisson and hyperbolic distributions by enlarging the appropriate areas of Fig. 2. It is then possible to determine also u_1' for each case from Eqs. (4) and (11) respectively.

The results are listed in Table 1.

TABLE 1

Parameters	Hyperbolic	Poisson	Units of Measurement
T'	60,000	60,000	bits
A'	10.0		
$v' = u_o'$	100	100	bits
u_1'	3.3×10^{-4}	4.21	bits
a'	30.4	0.0627	bursts/bit
a'T'	1.82×10^{6}	3,770	bursts/min.

It is important to examine this table closely. First one must note that continuous processes have been used instead of the discrete processes really appropriate to digital signals, on the assumption that where one has 60,000 bits this is not of importance.

However, it is utterly unreasonable to talk of an average of 30 error bursts per bit, or of a minimum spacing between bursts of 3×10^{-4} bits. One must draw the conclusion, therefore, that 100 bit maximum gaps between error bursts in 1 minute burst chains (at a 1000 bit per second signaling rate) will not produce a hyperbolic distribution of the bursts within these chains. The Poisson assumption leads to values which, for strict theoreticians, are perhaps beginning to stretch the continuous processes used over the discrete bits, but they are generally reasonable values. An average of one burst every 16 bits, and a minimum spacing of bursts of 4 bits, sound realistically possible. Thus one is pushed to the necessity of assuming a Poisson distribution for the occurrence of the error bursts over periods of one minute or shorter, even though a hyperbolic distribution is assumed for the substantially longer periods.

This is not altogether strange. What is probably involved is the density of the bursts rather than the exact size of the time interval. When a high density of error bursts occurs, the hyperbolic distribution bunches them so closely that the formula calls for many errors to occur per bit. In practice, additional errors falling in the same bits as before are not apparent. Only new errors falling in hitherto good bits are perceived. Thus, as the density increases, the bunching is diluted and the distribution approaches the Poisson.

There is a trace of this effect in Fig. 6 of Ref. 1. There, at the highest error densities, the experimental curves veer somewhat from the hyperbolic toward parallelism with the Poisson curves. This has been noticed in other (unpublished) plots of error distributions.

Another example of shift in character of distribution is observed by Crichlow[6] in describing the distribution of atmospheric noise impulses in VLF radio. Here the statistics of the noise, over several minutes to an hour, remain essentially constant; but different from the statistics over longer periods (say up to 8 hours). Over much longer periods (beyond 8 hours) there are additional and systematic changes, but they are not at issue here.

Extension of Example

On the above basis calculations were continued for a range of chain durations (all for the 1000 bit per second rate, and 100 bit critical gap) and the results are listed in Table 2. The number of bursts per bit, a', and the shortest burst-free run, u_1', are seen to continue to be reasonably realistic over the entire range of chain durations.

Specific Examples of Over-All Tests

The solutions of Table 2 can now be applied to over-all tests, in accordance with Steps 7 to 9 of the general procedure of Section I. In this portion of the procedure it is assumed that the statistics are hyperbolic, that is, that

$$P(a,c) = a/(Ac + a) \qquad (19)$$

It is also assumed that the long-time average burst rate is 10 per minute. In Eq. (19) the a is measured per time units of T', so that a = 10 only for the case where T' = 1 minute. The A in Eq. (19) is computed from Eq. (6), with T measured in units of T'. The end product is M(T'), the number of chains of duration T' or longer, in the over-all test. It is

$$M(T') = TP(a,c) \qquad (20)$$

TABLE 2

Parameters	Burst Chain, minutes .01	.1	1	10	100	Units of Measurement
T'	6×10^2	6×10^3	6×10^4	6×10^5	6×10^6	bits
A'	5.40	7.70	10.00	12.30	14.61	
$-E1(-x')$	1.23×10^{-1}	4.30×10^{-3}	2.65×10^{-4}	1.94×10^{-5}	1.52×10^{-6}	
$x' = a'u_o'$	1.36	3.89	6.27	8.59	10.92	
a'	.0136	.0389	.0627	.0859	.1092	bursts/bit
a'T'	8.15	2.33×10^2	3.77×10^3	5.15×10^4	6.56×10^5	bursts/T'
u_1'	19.4	6.79	4.21	3.07	2.42	bits

The results are plotted as the solid lines in Fig. 3. The dotted lines are discussed in the next Section.

Where a solid line crosses the ordinate for a number of chains equal to one, the length of the longest expected chain in a test of given total duration is given. This is plotted in Fig. 4.

III. VARIATION OF CRITICAL GAP ASSUMPTIONS

The 100 bit critical gap seems unrealistically small when considering chains of 10^5 or 10^6 bits length. Exploration has therefore been made of a gap varying with the square root of chain length, as illustrated in Fig. 5.

This brings about a complication in thought. In the 100 bit gap example, all chains and gaps are definitely laid out in time, in a given test, and the total number of chains is equal to the sum of the numbers of chains of the respective lengths.

With a variable gap, when the chains of a shorter length are laid out (with a shorter critical gap), it is quite possible to find that a long chain is being subdivided into a number of pieces. Thus the "total" number of chains can include the same aggregation of bursts several times, first in a long chain, and later in shorter chains, and has therefore lost some significance. While this complicates the thought, it need not be confusing in practice if the situation is understood.

With the revised assumptions, the number of bursts per bit within T', and the shortest burst-free run, also within T', come out as plotted in Fig. 6.

For the over-all test T the assumption was made, instead of a constant long-time average burst density, of a constant bias in the hyperbolic formula (Eq. (28) of Ref. 1), namely

$$P(h,c) = h/(c + h) \qquad (21)$$

This has certain advantages, because in a continuously running test with stationary error-burst incidence, the long-time average a rises with time, but the bias h remains constant(1). The quantity h is in the same units as the average a, and was taken as 2×10^{-6} bursts per bit, or 0.12 bursts per minute. According to conditions, this gave long-time averages running from a little under 1, to a little over 2, bursts per minute.

The results, in terms of expected numbers of chains of various lengths, are plotted as solid lines in Fig. 7. The expected longest burst chains, for over-all tests of various durations, are plotted in Fig. 8. The results for the very longest burst chains, of one hour or more, are to be taken with reserve. This is because the assumption of a Poisson distribution of bursts during such a long chain is no longer really tenable.

IV. COMPARISON WITH EXPERIMENTS

There have been few experiments to measure durations of error-burst chains, and none of these were based on the specific definitions used above. The best of these tests was described by Dimock(2), and has been re-plotted in dotted lines (as maximum and minimum limits) in Figs. 3 and 7. A number of the parameters, such as the bit rate, are not exactly the same. This, however, does not prevent an order-of-magnitude comparison.

The principal characteristic to note in the comparisons in these figures is the sustained probability, at moderate levels, for the longer burst chains (say of one to ten minutes). This comes about by the bunching of bursts. With a complete Poisson distribution of bursts, this bunching is of only negligible probability, many orders of magnitude below the plots in the figure. Both sets of computations reflect this characteristic. The Dimock data are better approached in Fig. 7 than in Fig. 3. Here they are close enough for further adjustments of the hypothetical parameters to be hardly warranted until definitions of burst chains are made more specific in reporting experiments.

V. CONCLUSIONS

Conclusions which can be drawn from the foregoing investigation of the hyperbolic distribution of long error-burst chains are essentially:

1) Under certain assumptions, that are considered realistic, a fair approximation can be obtained to the observed experience. This cannot be

done with a conventional Poisson distribution of bursts.

2) One of the necessary assumptions is that the burst distribution in relatively short times (one to ten minutes) of high error density be of the Poisson type, even if the longer time distribution is hyperbolic. There has been some casual experimental evidence for such a transition. It is surmised from this situation that, where a large number of sources of impairment to give a high error density combine together, the distribution becomes Poisson. Where the sources are much sparser the distribution is hyperbolic.

3) The analysis requires the solution of implicit transcendental equations, which is done graphically. Thus, the results cannot be expressed in simple analytic form, but must be given in terms of a variety of illustrative examples. Only a small range of examples has been computed and reported here. They are, however, believed to be reasonably realistic.

4) The results will vary according to the definition adopted for a burst chain. A really complete correlation with experience is difficult to work out until the tests that are conducted use more standardized and explicit definitions of chains.

APPENDIX A

Poisson Distribution

Consider burst initiations as events. The probability of an event-free run u, when the long-time average occurrence of events is a, is[4]

$$Q(a,u) = \exp(-au) \tag{22}$$

The time τdu in a test of duration T devoted to event-free intervals between u and u + du is (with the negative sign included because of the direction of the cumulation in Eq. (22))

$$\tau du = Tq(a,u)du$$

$$= -T(d/du)Q(a,u)du \tag{23}$$

$$= aT \exp(-au)du \tag{24}$$

The number m(u)du of such intervals in that time is equal to the time τdu divided by the duration u of each interval (neglecting the du increment), namely

$$m(u)du = \tau du/u$$

$$= (aT/u) \exp(-au)\, du \tag{25}$$

The cumulated number M(u) of intervals in T of duration equal to or greater than u is

$$M(u) = aT\int_u^\infty (1/u) \exp(-au)\, du \tag{26}$$

$$M(u) = aT\int_{au}^\infty (1/au) \exp(-au)\, dau \tag{27}$$

$$M(u) = -aT\ Ei(-au) \tag{28}$$

Here Ei(x) is the exponential integral function tabulated by Jahnke and Emde.[5]

The ordinal number of an event-free run is its rank in order of duration, starting with the longest and ending with the shortest. It is M(u) of Eq. (28) at its integer values. For the longest run, u_o, in the distribution

$$M(u) = 1 \tag{29}$$

so that

$$1 = -aT\ Ei(-au_o) \tag{30}$$

For the shortest run, u_1, in the distribution

$$M(u) = aT$$

and

$$1 = -Ei(-au_1) \tag{31}$$

For convenience an abbreviated table of Ei(x) is included.

TABLE 3

x	-Ei(-x)
.0001	8.63
.001	6.33
.01	4.04
.1	1.823
1.	.219
3.	1.304×10^{-2}
10.	4.157×10^{-6}
13.	1.622×10^{-7}

APPENDIX B

Hyperbolic Distribution

An identical sequence of operations is used as in Appendix A, but the equations are taken from Ref. 1.

The probability of an event-free run is

$$Q(a,u) = h_o/(h_o + u) \tag{32}$$

where

$$h_o = A/a = \text{bias constant}$$

The time τdu devoted to event-free intervals of length between u and u+du is

$$\tau du = T\, q(a,u)$$

$$= -T\ (d/du)\ Q(a,u)\ du \tag{33}$$

$$= T\ h_o/(h_o + u)^2\ du \tag{34}$$

The number m(u)du of such intervals is

$$\tau du/u = T(h_o/u)/(h_o + u)^2\, du \qquad (35)$$

The cumulated number M(u) of such intervals is

$$M(u) = T\int_u^\infty (h_o/u)/(h_o + u)^2\, du \qquad (36)$$

$$= T\left[\frac{1}{h_o + u} - \frac{1}{h_o}\log_e \frac{h_o + u}{u}\right]_u^\infty \qquad (37)$$

$$= T\left\{-\frac{1}{h_o + u} + \frac{1}{h_o}\log_e \frac{h_o + u}{u}\right\} \qquad (38)$$

That is

$$M(u) = (T/h_o)\, N(U) \qquad (39)$$

where

$$N(U) = \log_e\left(\frac{1 + U}{1}\right) - \frac{1}{1 + U} \qquad (40)$$

and

$$U = u/h_o \qquad (41)$$

For the longest run in the distribution

$$1 = (T/h_o)\, N(U_o) \qquad (42)$$

or

$$1 = (aT/A)\, N(U_o) \qquad (43)$$

For the shortest run in the distribution

$$1 = (1/A)\, N(U_1) \qquad (44)$$

An abbreviated table of N(U) is appended.

TABLE 4

U	N(U)
.0001	8.21
.001	5.91
.01	3.63
.1	1.489
1.	.1932
10.	4.40×10^{-3}
100.	4.93×10^{-5}
1000.	4.99×10^{-7}

REFERENCES

1. Mertz, P., "Statistics of Hyperbolic Error Distributions in Data Transmission," IRE Convention Record, Part 8 - Communications Systems, March 1961, pp. 160-166.

2. Dimock, P.V., "Data Transmission on Long-Haul Telephone Facilities," Proceedings of the National Electronics Conference, vol. 16, October 1960, pp. 29-36.

3. Campbell, G.A., Collected Papers, American Telephone and Telegraph Company, 1937, p. 224; also, "Probability Curves, Showing Poisson's Exponential Summation," Bell System Technical Journal, vol. 2, 1923, pp. 95-113.

4. Davenport, W.B. and W.L. Root, An Introduction to the Theory of Random Signals and Noise, McGraw-Hill Book Company, 1958, p. 115.

5. Jahnke, E. and F. Emde, Tables of Functions, Dover Publications, 1943, pp. 1, 6-8.

6. Crichlow, W.Q., "Noise Investigarion at VLF by the National Bureau of Standards," Proc. IRE, vol. 45, June 1957, pp. 778-782.

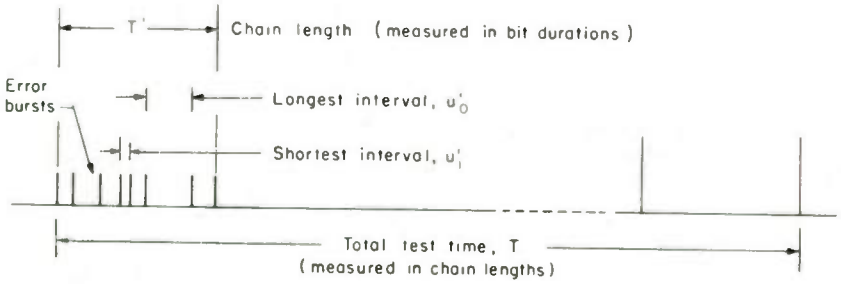


Fig 1 — Total test time divided into chain lengths (chain length is at least T' long if critical gap exceeds u'_o)

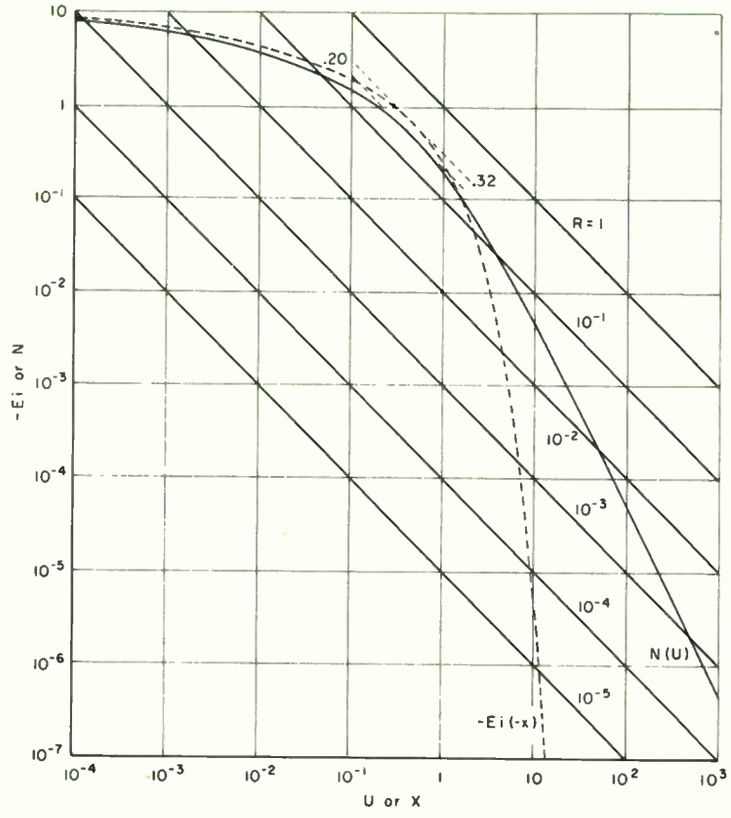


Fig. 2 — Solution of implicit equations in Ei(x) and N(U)

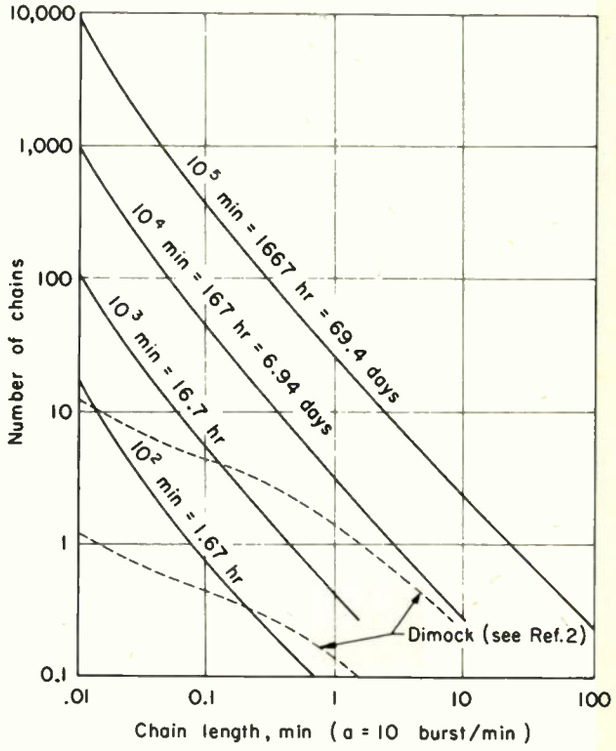


Fig. 3 — Distributions of burst chain lengths, for tests of various durations, (constant critical gap = 100 bits), compared with experience

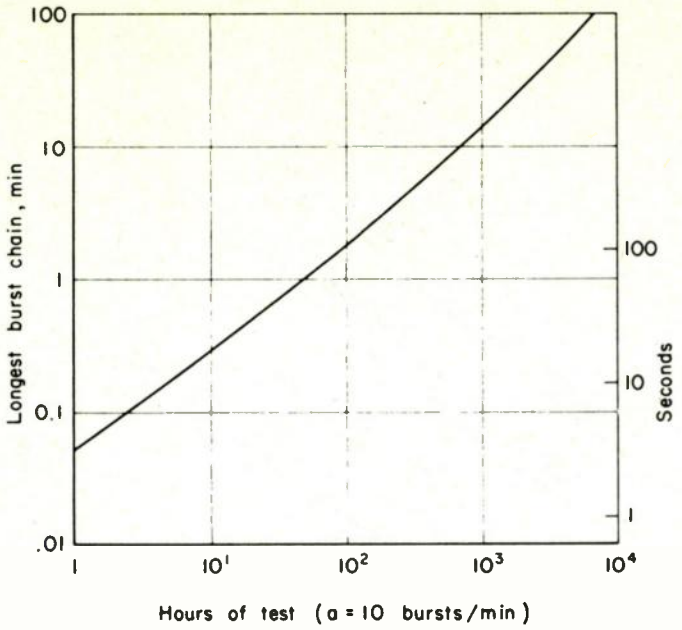


Fig. 4 — Longest expected burst chains versus test duration (constant critical gap = 100 bits)

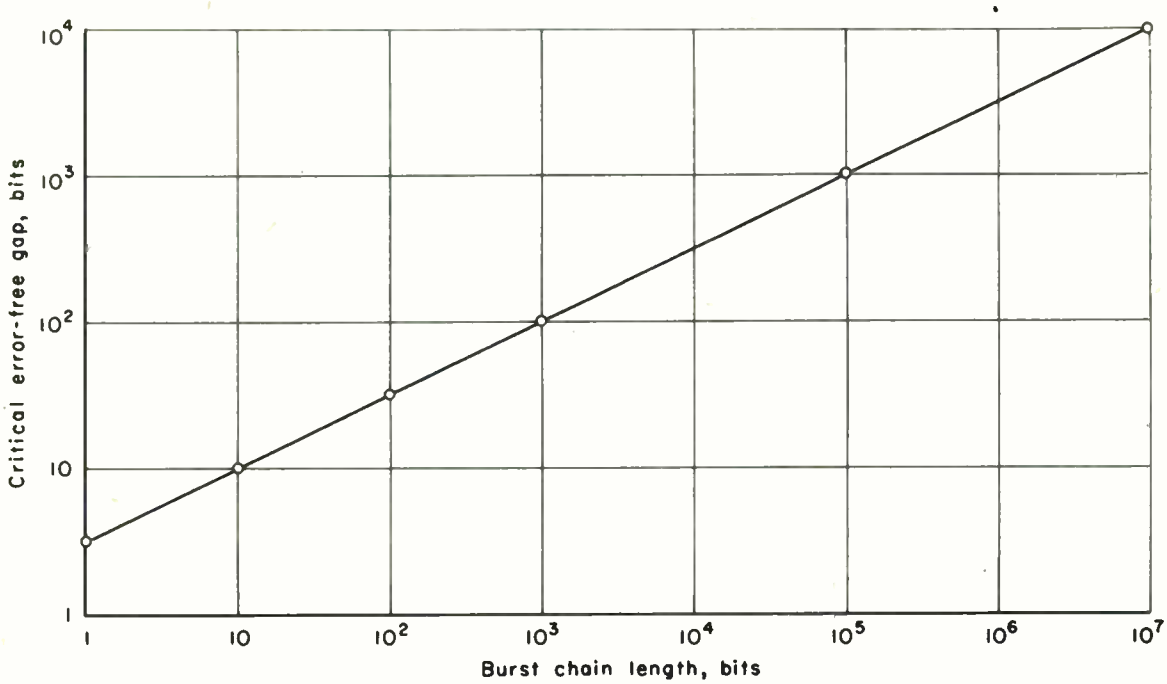


Fig. 5 — Critical gap varying with burst chain length

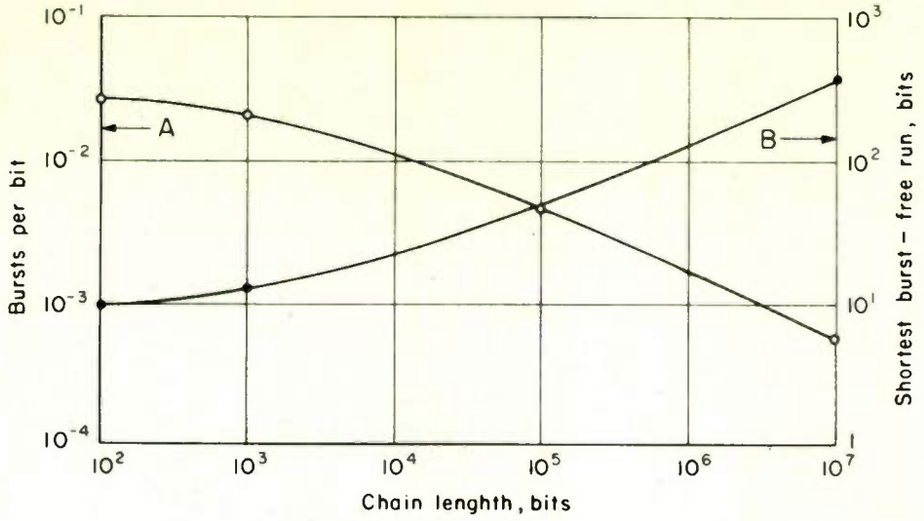


Fig. 6 — Burst per bit and shortest burst-free runs (critical gap as in Fig. 5)

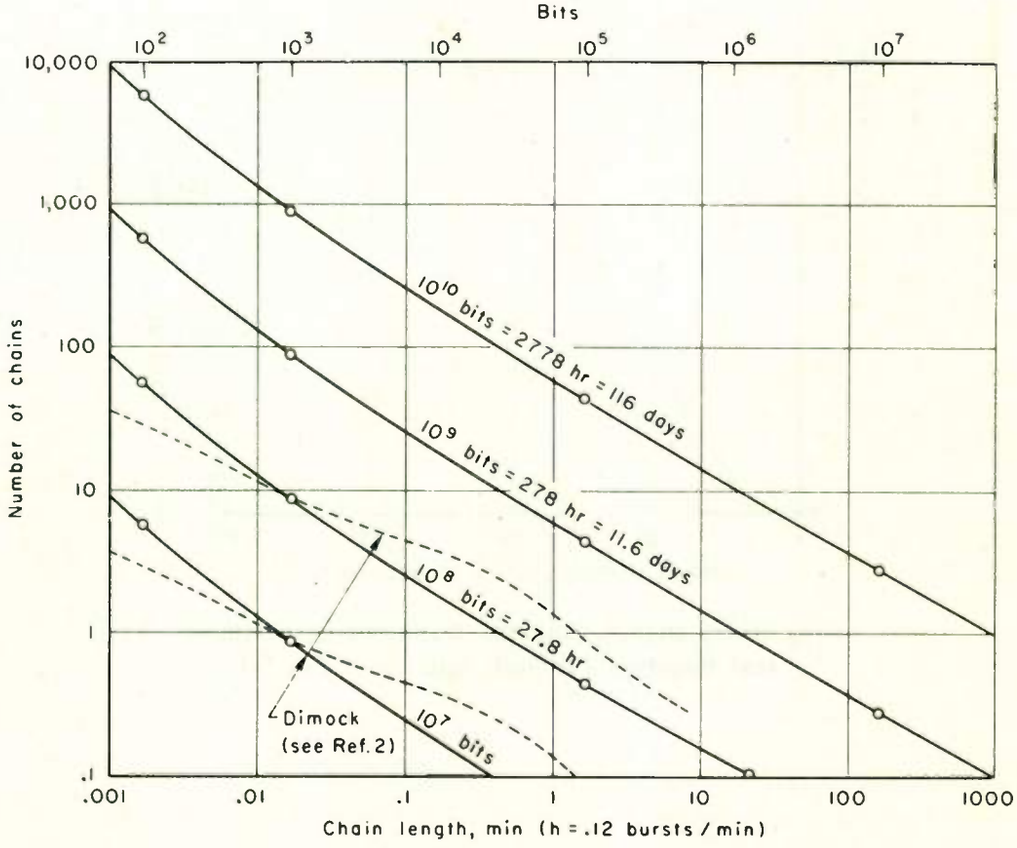


Fig. 7 — Distributions of burst chain lengths, for tests of various durations, (critical gap as in Fig. 5), compared with experience

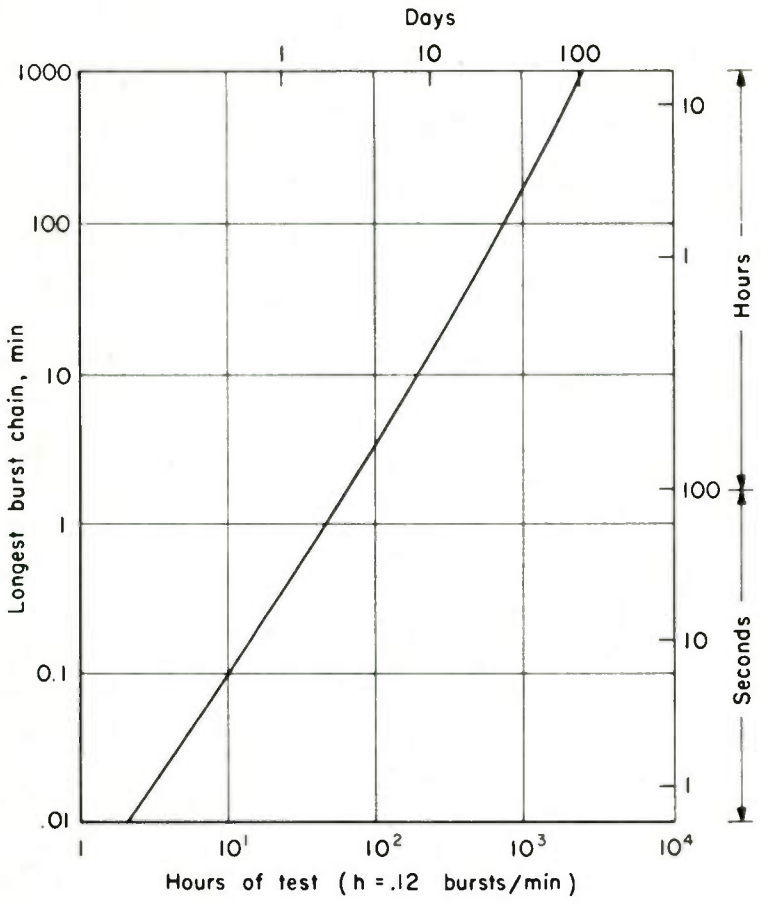


Fig. 8 — Longest expected burst chains versus test duration (critical gap as in Fig. 5)

GROUP SYNCHRONIZATION FOR DIGITAL TRANSMISSION SYSTEM

T. Sekimoto and H. Kaneko
Communication Research Laboratory
Nippon Electric Co., Ltd.
Kawasaki, Kanagawa, Japan

Summary

Group synchronization systems with digital feedback are studied, which are particularly applicable for PCM systems. Group synchronization is used to determine the time origin of a sequence of binary signals by finding a specified deterministic pattern from a randomly modulated binary sequence. A statistical analysis is made on the recovery process of synchronization, which is a discrete-time random process with finite state transitions. A new system with faster recovery characteristics is presented, and an optimal framing pattern is shown. Considerations on the minimization of the recovery time and on the improvement of the stability of synchronism are also made. Experimental results obtained by a recovery process simulator are shown together with the theoretical results.

1. Introduction

When digital information is transmitted through a noisy channel, the received signal should be sampled at a correct time so that the average probability of error is minimized. However, in many practical cases, the source transmission rate or the transmission medium is not strictly stable, and these factors cause a random delay of the time of signal arrival, thus necessitating synchronization, which we call "bit-rate synchronization".

Furthermore, in practical cases, these digital signals are serially encoded in blocks with a certain length of code or multiplexed in time-division basis. This block constitutes a frame, a unit of the sequence of digital signals. In such a case we have to determine the time origin of each sequence by inserting additional digits, or by affording a prefix condition* or a certain restriction on the information codes[2]. We call it "Group synchronization[3]" or "framing[4]".

Many attempts have been made to realize a stable synchronizer. We can classify them into two categories; one is step-by-step framing and the other is long-time framing in which synchronism is recovered by hunting when the collapse of synchronism has occurred. The former method is performed by detecting a specified framing pattern using sequential filters and so forth[3,5,8]. By this method the collapse will occur when an error occurs in the framing digits, or when the same pattern as the framing digits eventually occurs in the randomly modulated information pulses. However, the correct framing is regained at the next frame.

In a system using the latter method, the synchronizer usually monitors the specified framing pattern in the transmitted sequence and immediately hunts or shifts the digital phase of the synchronizer when a framing error is detected[4,7,11]. Therefore, the latter method is free from the occurrence of the same pattern, but it was thought that the recovery time was considerably long.

In this paper, the latter system with the synchronizer composed of digital circuits, which we call the "digital synchronizer[9]", is investigated from a statistical point of view, and a new system, "resetting sequence system", in which the recovery time is highly improved is presented with its stochastic analysis[10,11,12]. Also a few considerations on the optimal framing patterns, an optimization of this system and on the improvement of the stability of the synchronizer are discussed. Theoretical results are checked by an experiment using a recovery process simulator.

2. Digital Synchronizers

Two typical methods are considered to distribute the framing digits in a frame; one is the (a) "interlace method", in which the framing digits are distributed with equal specing as shown in Fig.1(a), and the other is the (b) " sequence method", in which all of the framing digits are located at the biginning of a frame as shown in Fig.1(b).

The synchronizer has a monitor circuit, by which the framing error is monitored at each framing digit, and it orders the whole system to hunt to recover synchronism when a framing error is detected. Also there are two methods of hunting; a (A) "bit-by-bit shifting method", by which the synchronizer shifts its digital phase by a one digit interval (hereinafter we use a "bit" unit) for each detection of a framing error, and a (B) "resetting method", in which the phase shift is performed with a variable length of digits; i.e., if a framing error is first detected at the x-th digit in the framing sequence, the phase shift is made by x bits.

* See, e.g., Ref.(1), p.67.

The synchronizer circuits according to (A) and (B) are illustrated in Figs.2 and 3 respectively. The bit-rate synchronizer produces a train of clock pulses. The channel separator, or the distributor, is composed of an N-bit shift register with a ring connection, if N denotes the number of digits in a frame, and is advanced bit by bit in accordance with the input clock pulses. By the channel separator and the decoder, the information signals on a time-division basis are transformed into a set of signals on a space-division basis. The outputs of the channel separator corresponding to the framing digits are lead to gate circuits AND_1 and AND_2, which select or indicate the time position corresponding to the framing digits. The output of AND_1 is the framing pattern in the transmitted sequence. The framing pattern generator, which is advanced by the output of AND_2, generates the same framing pattern as is used in the transmitted sequence.

The monitoring of the framing errors is made at the non-coincidence circuit, in which the transmitted framing pattern is compared with the locally generated framing pattern. The non-coincidence circuit is known as the modulo 2 adder or the "exclusive OR" circuit. When the system is in synchronism, these two patterns should coincide, and no trigger pulse is produced by the non-coincidence circuit. However, if the collapse of synchronism occurs, AND_1 selects wrong digits, and non-coincidence output will be produced.

In the system using the bit-by-bit shifting method, this output is fed back to an inhibitor through a delay circuit, which affords a one bit delay, and takes a clock pulse off, thus resulting in the retardation of the digital phase of the channel separator by one bit. Since the code of the mis-selected digits by AND_1 is probabilistic, the detection of a framing error and, hence, the retardation of the phase are probabilistic. When no error is found at a framing digit, the comparison is made at the same framing digit, which is itself shifted by one bit in time. This retardation is repeated and finally the synchronizer recovers synchronism.

On the other hand, in the system using the resetting method the phase shift is made by resetting the channel separator, together with the framing pattern generator, to its 0-th digit for every detection of non-coincidence. If we apply a sequence method for the resetting system, we can get greatly improved recovery characteristics.

As we see in Figs.2 and 3, the synchronizer is composed of a simple digital circuit elements, and we can modify it by logical transformations according to given conditions.

3. Analysis of the Recovery Process

We will show an analysis of the recovery process of the group synchronization for the two types of synchronizers: bit-by-bit shifting interlace system and resetting sequence system.

Now we define

- N number of digits in a frame
- r number of framing digits in a frame
- q probability of error of the transmission channel
- p probability that a framing error is detected at the non-coincidence circuit.

We will assume here that

- (A_1) bit-rate synchronization is correct*,
- (A_2) q is very small so that the collapse of synchronism does not occur during the recovery process,
- (A_3) the randomly modulated information pulses are independent and identically distributed,
- (A_4) for the interlace system, the framing pattern with length r has no periodicity with period less than r, because of the uniqueness of synchronism.

3.1 Bit-by-bit Shifting Interlace System

We will consider only the case where N is divisible by r**. Then, each of the framing digits is space by N/r bits. Suppose that the collapse of synchronism has occurred. If a framing error is not detected, then the comparison is made at the next framing digit space by N/r bits. On the other hand, if an error is detected, then the state of the channel separator is shifted by one bit, and the next comparison is made at that position. This procedure is repeated until the system is absorbed in the synchronizing state. Therefore, the recovery process is a Markov process, and its state transition is illustrated in Fig.4.

The system is characterized by a closed set of N states, $S_0, S_1, \ldots, S_{N-1}$, where S_0 is the synchronizing state. If we denote $\overline{S}_0 = \{S_1, S_2, \ldots, S_{N-1}\}$, then S_0 is the state that the system is in at the collapse of synchronism. If $\mathcal{E}$ denotes the event that the transition from S_0 to S_0 occurs, then $\mathcal{E}$ is clearly a recurrent event.

The transition probabilities are

---

* If we assume that the bit-rate synchronization is correct, then we will meet a contradiction in that no group synchronization is needed after the receiver has one attained the correct framing. However, in an actual situation, the bit-rate synchronization is not strictly correct due to the lack of stability of the transmitter, the receiver and the transmission medium. Here, in order to separate the bit-rate synchronization from our problem, we will employ this assumption.

** The results of the following analysis are, however, approximately applicable even for the case in which N is not divisible by r, if the framing digits are nearly equally spaced.

$$\left.\begin{aligned} \Pr(S_i \longrightarrow S_{i+1}) &= \begin{cases} q, & i = 0 \\ p, & i \neq 0 \end{cases} \\ \Pr(S_i \longrightarrow S_i) &= \begin{cases} 1-q, & i = 0 \\ 1-p, & i \neq 0 \end{cases} \end{aligned}\right\} \quad (1)$$

and the transition times are

$$\left.\begin{aligned} t(S_i \longrightarrow S_{i+1}) &= 1 \quad \text{bit} \\ t(S_i \longrightarrow S_i) &= N/r \quad \text{bits} \end{aligned}\right\} \quad (2)$$

irrespective of i.

Now we define the probability that the state transition from S_i occurs at the w_i-th trial by $\Pr(w_i)$, then

$$\Pr(w_i) = \left\{\begin{aligned} & q(1-q)^{w_o-1}, && i = 0 \\ & p(1-p)^{w_i-1}, && i \neq 0 \end{aligned}\right\} \quad (3)$$

Therefore, $\Pr(w_i)$ is the geometrical distribution, and its expectation and variance are

$$E(w_i) = \begin{cases} 1/q & i = 0 \\ 1/p, & i \neq 0 \end{cases} \quad (4)$$

$$V(w_i) = \begin{cases} (1-q)/q^2, & i = 0 \\ (1-p)/p^2, & i \neq 0 \end{cases} \quad (5)$$

If we denote

$$u = \sum_{i\ 1}^{N-1} (w_i - 1) \quad (6)$$

then, since the w_i's are mutually independent and identically distributed for $i = 1,2,\ldots,N-1$, the probability function of u is given by the (N-1)-st order convolution of $\Pr(w_i)$'s, and is the Pascal distribution (or the negative binomial distribution), i.e.,

$$\Pr(u) = \Pr^{(N-1)^*}(w)$$

$$= \binom{N+u-2}{N-2} (1-p)^u\, p^{N-1} \quad (7)$$

Therefore, we have

$$\left.\begin{aligned} E(u) &= (N-1)(1-p)/p \\ V(u) &= (N-1)(1-p)/p^2 \end{aligned}\right\} \quad (8)$$

Since u is the total number of framing digits where the system stays in the same state, the recovery time is Nu/r bits plus the time required for N states transitions, and if we express it in frame unit, then

$$f = \frac{u}{r} + 1, \qquad \text{frames} \quad (9)$$

Therefore, the average recovery time and its variance are

$$\left.\begin{aligned} E(f) &= \left(1 + \frac{N-1}{r}\,\frac{1-p}{p}\right), && \text{frames} \\ V(f) &= \frac{N-1}{r^2}\,\frac{1-p}{p} && (\text{frames})^2 \end{aligned}\right\} \quad (10)$$

The curves of $E_I(f) = E(f)$ and $\sigma_I(f) = \sqrt{V(f)}$ are shown in Figs.5 and 6 respectively, when N = 100 and p = 1/2. In these figures we see that $E_I(f)$ and $\sigma_I(f)$ decreases in inverse proportion to r.

As noted before, $\mathcal{E}$ is a recurrent event, and the recurrent time f_r is given by

$$f_r = f + (w_o - 1)/r, \qquad \text{frames} \quad (11)$$

Therefore, the average recurrent time of the collapse of synchronism is

$$E(f_r) = 1 + \frac{N-1}{r}\,\frac{1-p}{p} + \frac{1}{r}\,\frac{1-q}{q}, \qquad \text{frames} \quad (12)$$

The first and the second terms are negligible compared with the third one when $q \ll 1$, and

$$E(f_r) = 1/rq$$

It is important to note that the ratio

$$R = \frac{E(f)}{E(f_r)}$$

which is the measure of the disturbance due to the collapse of synchronism, is almost never improved by increasing r.

3.2 Resetting Sequence System

In the resetting sequence system, on the other hand, if a framing error is detected at the x-th digit in the framing sequence, the channel separator is shifted by x bits at a time, and the next comparison is always made from the first digit of the framing sequence. Therefore, the probability that no framing error is detected at any framing digit in a frame is very small when r is moderately large. The performance of this system is better understood by referring to the state transition graphs in Fig.7 for the examples with framing patterns (111), (101), (011) and (110). The same graphs would be given for the complementary patterns respectively.

This system has N elementary states, S_0, S_1, $S_2,\ldots,S_{N-1}$. Each of them is a set of sub-states s_{ij}'s, where $i = 0,1,2,\ldots,N-1$ and $j = 1,2,\ldots,r$. S_0 is the synchronizing state. Since $S_0,S_1,\ldots,S_{N-1}$ constitute a closed set of states, the complementary state $\bar{S}_0 = \{S_1, S_2, S_3, \ldots, S_{N-1}\}$ is the state that the system is in at the collapse of synchronism. Also an event $\mathcal{E}$ that the transition occurs from S_0 to $\bar{S}_0$ is a recurrent event.

In Fig.7(a) $\sim$ (d), we can see that the state transitions from such sub-states S_{ij} where

$$0 < (i+j) \bmod N < r, \qquad i \neq 0$$

are a little different from those occurred from other sub-states. Some state transitions are deterministic. These singular transitions are caused by the effect of the framing pattern itself. In order to avoid the difficulty in the analysis due to the singular transition mode, we assume that

(A_5) the singular transitions are replaced by the regular transitions as shown in Fig.7(e).

Then, the transition probabilities between the sub-states are

$$\left.\begin{aligned} \Pr(s_{ij} \longrightarrow s_{kl}) &= \begin{cases} q, & i = 0, \\ p, & i \neq 0, \end{cases} \quad k = (i+j) \bmod N \\ \Pr(s_{ij} \longrightarrow s_{im}) &= \begin{cases} 1-q, & i = 0, \\ 1-p, & i \neq 0 \end{cases} \quad m = j \bmod r + 1 \end{aligned}\right\} \quad (13)$$

and zero otherwise. The transition time according to each state transition is

$$\left.\begin{aligned} t(s_{ij} \longrightarrow s_{kl}) &= 1 \text{ bit}, \quad k = (i+j) \bmod N \\ t(s_{ij} \longrightarrow s_{i,j+1}) &= 1 \text{ bit}, \\ t(s_{ir} \longrightarrow s_{il}) &= (N-r+1) \text{ bits}, \end{aligned}\right\} \quad (14)$$

The remarkable features of this system arise from the fact that the probability of state transition ($s_{ir} \longrightarrow s_{il}$) with long transition time is exponentially reduced as r increases.

Now, suppose that the system is in a state S_i. Then the length of time w_i that the system stays in S_i is

$$w_i = x_i + r y_i, \quad \begin{cases} x_i = 1,2,\ldots,r \\ y_i = 0,1,\ldots \end{cases} \quad (15)$$

where

x_i = position of the framing digit where the transition occurs in state S_i

y_i = number of transitions ($s_{ir} \longrightarrow s_{il}$).

x_i and y_i are independent random variables. Since w_i is a random variable with the geometrical distribution, for $i \neq 0$,

$$\Pr(w_i) = p(1-p)^{w_i-1} = \Pr(x_i)\,\Pr(y_i), \quad (16)$$

$$\left.\begin{aligned} \Pr(x_i) &= p(1-p)^{x_i-1} / \{1-(1-p)^r\} \\ \Pr(y_i) &= \{1-(1-p)^r\}\,(1-p)^{r y_i} \end{aligned}\right\} \quad (17)$$

Accordingly the expectations and the variances of them are

$$\left.\begin{aligned} E(x_i) &= \frac{1}{p} - r\,\frac{(1-p)^r}{1-(1-p)^r}, \\ V(x_i) &= \frac{1}{p^2} - r^2\,\frac{(1-p)^r}{\{1-(1-p)^r\}^2}, \end{aligned}\right\} \quad i \neq 0 \quad (18)$$

$$\left.\begin{aligned} E(y_i) &= \frac{(1-p)^r}{1-(1-p)r}, \\ V(y_i) &= \frac{(1-p)^r}{\{1-(1-p)^r\}^2}, \end{aligned}\right\} \quad i \neq 0 \quad (19)$$

For x_0 and y_0, we can obtain similar expressions simply by substituting q instead of p in these formulae.

As we have seen in Fig.7(e), the state transitions are made by skipping some states, sometime even the synchronizing state S. The system is not always absorbed in the synchronizing state by the first scanning (or cycle), and it is absorbed in S_0 only when the sum of the x_i's first becomes equal to an integral multiple of N, which means, for a certain integer h,

$$\sum_{i=1}^{n} x_i = h N \quad (20)$$

where the number of transitions n and the number of cycles h are also random variables, and we can write n in the form

$$n = \sum_{j=1}^{h} m_j \quad (21)$$

m_j is the number of state transitions in the j-th cycle. Accordingly, eq.(20) is written as

$$\sum_{j=1}^{h} \sum_{i=1}^{m_j} x_{ij} = hN \quad (22)$$

We also define

$$u = \sum_{i=1}^{n} y_i = \sum_{j=1}^{h} \sum_{i=1}^{m_j} y_{ij} \quad (23)$$

Then, u is the total number of transitions through long paths ($s_{ir} \longrightarrow s_{il}$) until the system is absorbed in S_0.

We will further assume that

(A_6) $N \gg r$

Then, the state transition converges to a stationary process as the transition goes on.

Exactly speaking, all the m_j's are not statistically independent. In order to simplify our analysis, we will employ a rough approxi-

mation that

(A_7) each cycle of transitions starts from S_o, although it is not so for $j \geq 2$.

Under this assumption the m_j's are statistically independent, and also h is independent of the m_j's. If we denote the probability that the system is absorbed in S_o at the first cycle by P_o, the distribution of h is approximately the geometrical distribution, i.e.

$$\left.\begin{aligned} \Pr(h) &\fallingdotseq (1-p_o)^{h-1} p_o \\ E(h) &\fallingdotseq 1/p_o \\ V(h) &\fallingdotseq (1-p_o)/p_o^{\,2} \end{aligned}\right\} \quad (24)$$

where

$$p_o = \frac{1}{E(x)} = \frac{1}{\frac{1}{p} - r\,\frac{(1-p)^r}{1-(1-p)^r}} \quad (25)$$

The proof of this formula is given in the Appendix.

Suppose that the system is absorbed in S_o in the first cycle. Then the distribution of m_o is given as follows:
We sample a_x elements from the x-th group of r groups of infinite population with probability Pr(x), subject to

$$m_o = \sum_{x=1}^{r} a_x$$

$$N = \sum_{x=1}^{r} x\, a_x$$

Thereofre, we have

$$\Pr(m_o) \fallingdotseq \sum_{\forall a_x \{\sum x a_x = N\}} \frac{m_o!}{a_1!\, a_2! \cdots a_r!} \prod_{x=1}^{r} \{\Pr(x)\}^{a_x}$$

By the definition of p_o,

$$\sum_{\forall m_o} \Pr(m_o) = p_o$$

Under the assumption A_7 we can approximate the distribution of $m_j = m$ by normalizing the above formula as

$$\Pr(m) = \frac{1}{p_o} \sum_{\forall a_x \{\sum x a_x = N\}} \frac{m!}{a_1!\, a_2! \cdots a_r!} \prod_{x=1}^{r} \{\Pr(x)\}^{a_x} \quad \cdots\cdots (26)$$

It is difficult to find the moments of this distribution. However, since $N \gg r$, we have approximately

$$E(m) = \frac{N}{E(x)} = \frac{N}{\frac{1}{p} - r\,\frac{(1-p)^r}{1-(1-p)^r}} \quad (27)$$

The variance of m is, by approximating eq.(26) to a binomial distribution with population size $N(1 - 1/r)$,

$$V(m) = \frac{p(1-p)}{1-(1-p)^r}\left[1 - \frac{1-(1-p)^r}{r}\right] N \quad (28)$$

Now, since the m_j's and h are mutually independent under the assumptions, the distribution of u given by eq.(23) is the doubly compound distribution of y, i.e.,

$$\Pr(u) = \sum_{h=1}^{\infty} \Pr(h) \left\{ \sum_{\forall m} \Pr(m) \left\{ \Pr(y) \right\}^{m^*} \right\}^{h^*} \quad (29)$$

The expectation and the variance of the compound distribution are given from the moment generating function [13] and are

$$\begin{aligned} E(u) &= E(h)\, E(m)\, E(y) \\ V(u) &= V(h)\, E^2(m)\, E^2(y) \\ &\quad + E(h)\, V(m)\, E^2(y) \\ &\quad + E(h)\, E(m)\, V(y) \end{aligned} \quad (30)$$

Recalling eqs.(15) and (20), we have the recovery time

$$f = u + h, \qquad \text{frames} \quad (31)$$

Therefore, the expectation and the variance of the recovery time are

$$E(f) = E(h) + E(h)\, E(m)\, E(y)$$

$$E(x_1) = \frac{1}{p} + r\,\frac{(1-p)^r}{1-(1-p)^r}(N-r), \quad \text{frames} \quad (32)$$

$$V(f) = V(h) + V(u) \quad (33)$$

Computation of these values are performed for $N = 100$ and $p = 1/2$, and the resultant curves are shown in Figs.5 and 6 as $E_s(f)$ and $\sigma_s(f) = \sqrt{V(f)}$ respectively.

The average recovery time decreases exponentially by increasing the number of framing digits while r is not large. E(f) and V(f) converge to E(h) and V(h) respectively as r increases, and the improvement becomes less effective for excessively large r. We also see that $\sigma_s(f)$ when r = 2 is larger than that when r = 1, which is chiefly caused by V(h) in eq.(33).

$\mathcal{E}$ is a recurrent event and the average recurrent time is given by

$$E(f_r) = E(f) + E(y_o)$$

$$= \frac{1}{p} + \frac{(1-p)^r}{1-(1-p)^r}(N-r) + \frac{(1-q)^r}{1-(1-q)^r} \quad (34)$$

The first and the second terms are negligible compared with the third term when $q \ll 1$, and

$$E(f_r) = 1/rq \quad (35)$$

A remarkable feature appears here in that the ratio $R = E(f)/E(f_r)$ is almost exponentially reduced by increasing r, when r is not large, whereas this is not true in the bit-by-bit shifting interlace system.

We considered only two cases among four possible combinations of the methods (a), (b), (A) and (B). We can also analyze the rest of them; the bit-by-bit shifting sequence system and the resetting interlace system. However, these two systems have no additional advantages over the bit-by-bit shifting interlace system. The state transition graph, an example of the former system for a framing pattern (101), is shown in Fig.8. In this case, the probability of passing long paths is not different from that of passing short paths. Therefore, the average recovery time is approximately equal to that of the bit-by-bit shifting interlace system. Also for the resetting interlace system it is evident that the resetting of the framing digits, each of which is spaced by N/r bits, has nothing to contribute to the improvement of the recovery characteristics.

Note that if r = 1, all the four systems discussed above are identical.

4. Optimal Patterns for the Resetting Sequence System.

We have shown a stochastic analysis of the recovery process of synchronism for the resetting sequence system, derived particularly under A_6. This assumption is equivalent to neglecting the transient mode of the state transitions due to the framing pattern itself. We will find here a framing pattern which gives optimal recovery of synchronism.

As is seen in eqs.(32) and (33), the distribution of h has a great effect on the recovery characteristics, particularly for large r. On the other hand, we see in Fig.7 that there exists a framing pattern by which the system is necessarily absorbed in S_0 in the first cycle. This pattern is the one which is composed of all 1's (or all 0's). The expectation and the variance will be found in the same way. We have shown them in Figs.5 and 6 as $E_s'(f)$ and $\sigma_s'(f)$ with bloken lines, when this type of pattern is used.

However, in S_1 and S_{N-1} the number of substates where the framing error is to be detected is effectively one, and in S_2 and S_{N-2} it is effectively 2 and so forth. Existence of these states prevents the system from having a rapid recovery. This defect is rejected by composing a transmitting pattern $\underline{S}$ such that the complementary codes are put at the beginning and at the end of the all 1's (or all 0's) pattern, e.g., $\underline{S}$ = (01110). In this case the locally generated pattern $\underline{R}$ at the receiver should be $\underline{R}$ = (0*1110), and the channel separator and the framing pattern generator should be reset to the position marked by * in the framing sequence. Since the first 0 in $\underline{R}$ has a very small effect on the improvement of the recovery characteristics, we may use $\underline{R}$ = (*1110) instead. The state transition graph for this pattern is shown in Fig.9(a).

The expectation $E_s''(f)$ and the standard deviation $\sigma_s''(f)$ of the recovery time for this pattern are given approximately as follows;

$$E_s''(f) \fallingdotseq 1 + E(m')\,E(y)$$
$$V_s''(f) \fallingdotseq E^2(y)\,V(m) + V(y)\,E(m') \quad (36)$$

where E(m') is given by

$$E(m') = \frac{N-r-1}{E(x)} = \frac{N-r+1}{\frac{1}{p} - (r-1)\frac{(1-p)^{r-1}}{1-(1-p)^{r-1}}} \quad (37)$$

instead of eq.(27). Now, r is the total length of $\underline{S}$ and the effective number of r in $\underline{R}$ is accordingly (r-1). Therefore,

$$E_s''(f) = \frac{1 + (1-p)^{r-1}\{Np - 2(r-1)p - 1\}}{1 - (1-p)^{r-1}\{(r-1)p + 1\}} \quad (38)$$

Curves of $E_s''(f)$ and $\sigma_s''(f)$ are shown for N = 100 and p = 1/2 with chain lines in Fig.5 and in Fig.6 respectively. Since $E(y) \to 0$ and $V(y) \to 0$ as r increases, we have

$$\lim_{r\to\infty} E_s''(f) = 1$$
$$\lim_{r\to\infty} V_s''(f) = 0 \quad (39)$$

This result implies that the system recovers its synchronism almost certainly after one frame of disturbance, if we use a moderately large r. Although this property is attainable by a usual step-by-step synchronizing system, our system is better than that system because of the stability against an eventual occurrence of the same pattern as the framing pattern in the sequence of binary signals.

Finally, we will note that one can use $\underline{R}$ = (*111) instead of $\underline{R}$ = (*1110), when $\underline{S}$ = (01110). The effective number of framing digits

is reduced by 1 and is (r-2). However, the synchronizer circuit is simplified at the sacrifice of the reduction of effective r, because the locally generated pattern is composed of the same codes. The state transition graph for this pattern is shown in Fig.9(b).

5. An Optimization of the System

We have shown that the average recovery time of the resetting sequence system with an optimal framing pattern of the form $\underline{S}$ = (011....10) and $\underline{\bar{S}}$ = (*11....10) converges to one frame length as r increases. Now, we define a ratio K of the frame length to the length of the framing digits.

$$K = N/r \tag{40}$$

Then, K is a measure of the utility of the transmission channel, because 1/K is the ratio of digits spared for the framing purpose. Here, we will optimize the system with this type of optimal framing pattern so as to minimize the average recovery time for a given K and for a given transmission rate. The average recovery time μ in bits is, from eq.(38),

$$\mu = N E_s''(f) = Kr \frac{1-(1-p)^{r-1}\{(K-2)rp\ 2p-1\}}{1-(1-p)^{r-1}\{(r-1)p-1\}} \tag{41}$$

Curves of μ for various values of the parameter K are shown in Fig.10. Optimal r is found from the figure for each K. For example, if we design a system for K = 200, then the optimal N and r are 3,000 bits and 15 bits respectively, and 2985 bits are available for the information digits. In this case the minimized average recovery time is 3270 bits, or 1.09 frames.

The optimal r which minimizes eq.(41) also minimizes the time ratio R, because R is given by

$$R = \frac{E_s''(f)}{E(f_r)} = \frac{\mu rq}{N} = \frac{q}{K}\mu \tag{42}$$

6. An Improvement of Framing Stability

In general, the digital synchronizers loose their synchronism even when an error occurs in the framing digits, whether or not it is the actual collapse of synchronism. When it is the actual collapse, the system should recover its synchronism as fast as possible. However, it is not desirable that the system mis-hunts due to an eventual occurrence of an error in the framing digits.

We are considering very small average probability of error q. Suppose that the occurrence of an error is independent process, so the probability that at least one error is found in the framing digits in a frame is nearly rq. Hence, the probability that at least one error is found in each of the two consequtive frames is $(rq)^2$. Since rq is still very small, $(rq)^2$ is exceedingly small. On the other hand, when an actual collapse occurs, the monitor errors might be found in the consecutive frames. We can improve the stability of the synchronizer using this difference. That is, the hunting of the synchronizer is locked for one frame when a framing error is found, and the hunting is started only if an error is also found in the framing digits of the next frame.

An example of the locking circuit is shown in Fig.11, in which MM, Inh. and Del. denote a monostable multivibrator, an inhibitor and a delay circuit respectively. When a framing error is detected at the non-coincidence circuit, its output pulse triggers MM_1. The output pulse of MM_1, whose duration is one frame length, prevents the synchronizer from hunting by means of Inh_2 for one frame period. At the framing digits of the next frame Inh_2 is opened, and if an error is still found there, the system enters the recovery process. The operating time of MM_2 is adjusted to a little larger time than the average recovery time. The purpose of MM_2 is to prevent the misoperation of MM_1 during the recovery process, which results in a loss of recovery time. Del. is a delay circuit which is inserted to compensate the time delay through Inh_1, MM_1 and Inh_2. In this example the discrimination of the actual collapse from an eventual occurrence of an error is based on the statistical properties of the monitored errors in the two consecutive frames. However, in general, we can give various functions to it by using other circuits, such as sequential circuits, according to the statistical properties of the errors and the collapse[12].

We have given the formulae for the recurrent time in eqs.(34) and (35) in terms of q. However, if we use the locking circuit, the probability of mis-hunting is $(rq)^2$ and will be negligible compared with the probability of actual collapse NQ, which may be caused by incorrect bit-rate synchronization. The occurrence of the actual collapse may be equally probable at any digit in a frame. Therefore, if $N \gg r$, it will probably be found at the first digit of the framing sequence, and the average recurrent time is now

$$E(f_r) = 1/NQ = 1/K_rQ \tag{43}$$

instead of eq.(35).

Note that the optimal r which minimizes

$$R = \frac{E_s''(f)}{E(f_r)} = Q\mu \tag{44}$$

is also satisfied by the one which minimizes μ.

Hence, the optimal r is also found from Fig. 10.

If we use such locking circuits as we have explained, the recovery time is increased by one frame, which is the time interval for the locking of the synchronizer. However, the stability of the synchronizer is highly improved at the sacrifice of an increase in recovery time. Therefore, one may or may not use the locking circuit according to the values of q and Q.

7. Recovery Process Simulator

We made a recovery process simulator in order to check the theoretical analysis. The recovery process simulator is composed of the transmitter and the receiver as shown in the block diagram in Fig.12. Two dekatrons are used for the channel separators in each. The frame length is 100 and the framing pattern is arbitrarily programmed within $r = 1 \sim 5$ on the program boards. The clock rate is 10 kc and a thyratron noise generator is used as an information source which modulates the information pulses into a random sequence.

The synchronizer can be used in either the bit by bit shifting or the resetting method. Collapse of synchronism is triggered manually by pushing a key at the transmitter, which inhibits one of the framing digits in the transmitting sequence. Starting time of the collapse is also sent to the reciver, and the recovery time is measured by counting clock pulses by a counter during the time interval between the starting time and the end of the recovery process. The end of the recovery process is detected by monitoring the complete coincidence of the transmitted framing pattern and the locally generated pattern.

In this experiment we have used a framing pattern composed of all 1's. The results of experiment are shown in Fig.5 and 6, both for the bit-by-bit shifting interlace system and for the resetting sequence system. Since we have used a pattern of all 1's, the experimental results for the resetting sequence system correspond to $E_s'(f)$ and $\sigma_s'(f)$ in those figures. We can see that the experimental values are very close to the theoretical curves.

8. Conclusions

We have presented an analysis of the digital synchronizers, which can be composed of simple digital circuit elements. We summarize our conclusions as follows.

(1) Digital synchronizers are classified into the following two typical systems;
bit by bit shifting interlace system, and
resetting sequence system.
For the former the recovery time is improved almost in inverse proportion to the number of framing digits in a frame. However, the time-ratio of the channel disturbance due to the collapse can not be improved by doing so.

(2) For the resetting sequence system, the average recovery time is improved almost exponentially by increasing the number of framing digits, and so is the time-ratio of the channel-disturbance.

(3) The optimal framing pattern for the resetting sequence system is composed of all 1's codes with two 0's at the beginning and at the end of the framing sequence (or the complementary pattern of it). In this case the system will almost always recover synchronism in one frame, if the number of the framing digits is moderately large.

(4) When the utility of the transmission channel is given, one can find the optimal number of the framing digits, by which the average recovery time is minimized.

(5) The stability of the synchronizer against the mis-hunting is improved by using a circuit to lock the hunting for one frame.

(6) The results of an experiment with a recovery process simulator have well proved the statistical analysis of the recovery process.

(7) The resetting sequence system is particularly suitable for use in a long-frame Pulse Code Modulation system, or a long frame digital data transmission system.

These results presented in this paper can be modified according to the statistical properties of the information signals, the occurrence of errors and the collapse of synchronism. In the above presentation we have not considered any practical restrictions or any boundary conditions on the actual system. In an actual design of systems, these conditions should be taken into consideration.

Acknowledgement

Authors wish to express their thanks to Drs. M. Kobayashi, O. Harashima, T. Suzuki and Y. Degawa of the Nippon Electric Co., Ltd., and Mr. Y. Nakamaru of the Hokkaido University, for their instructive suggestions and the constant encouragement for our research. We are also indebted to the members of the committee on PCM systems sponsored by the Ministry of Education of Japan. Our thanks is also due to Mr. A. Larsen for reading through this paper.

Appendix

We will derive eq.(25). If $N \gg r$, the transition process converges to a stationary process, and the probability that the system is in one of the sub-states of a state S_i is neary constant irrespective of the initial probabilities as $i \rightarrow N$. If we consider a cross-section between the states S_{N-1} and S_N $(=S_0)$, the transition probability across this cross-section should be equal to unity, because all the states constitute a closed set of states and there is neither absorbing state nor transient state. Furthermore,

since the probability that the system is in the sub-state $s_{N-1,1}$ is p_o, we have

$$\sum_{x=1}^{r} p_o \cdot x \cdot Pr(x) = 1$$

Therefore, we have

$$p_o = \frac{1}{E(x)}$$

References

(1) M.Fano, "Transmission of information", MIT press and John Wiley & Sons, 1961

(2) E.N.Gilbert, "Synchronization of binary messages", IRE, PGIT, vol.IT-6, no.4, Sept. 1960, pp.470-477

(3) R.H.Barker, "Group synchronizing of binary digital systems", W.Jackson, Communication Theory, Butterworth Sci. Pub., London, 1953, pp.273 - 287

(4) J.M.Manley, "Synchronization for the PCM receiver", Bell Lab. Rec., vol.27, no.2, Feb.1949, p.62

(5) D.A.Huffman, "The synthesis of linear sequential coding networks", C.Cherry, Information Theory, Butterworth Sci. Pub., London, 1955, p.77.

(6) B.Elaspas, "The theory of autonomous linear sequential networks", IRE, PGCT, vol.CT-6, no.1,1959, pp.45 - 60.

(7) S.M.Shreiner, A.R.Vallarino, "48-channel PCM system", IRE Conv. Rec., Pt. 8 1957.

(8) H.Kaneko, "A statistical analysis of the synchronization for digital receivers", Master Thesis, University of California, Berkeley, Jan. 1962.

(9) T.Suzuki, H.Kaneko, "Synchronizing system for multi-channel PCM", IECEJ. Conv. Rec., 1958, p.265, (in Japanese).

(10) H.Kaneko, "Synchronizing system for multi-channel PCM (2)", Joint Conv. Rec., 1958, p.265, (in Japanese).

(11) Y.Nakamaru, H.Kaneko, "Synchronization system for digital transmission", J. of IECEJ, vol.43, no.12, Dec. 1960, pp.1388 - 1396, (in Japanese).

(12) Y.Nakamaru, H.Kaneko, "Synchronization for multi-channel PCM", IECEJ, PGIT, April 22, 1960, (abstract in English; IRE, PGIT, vol. IT-6, no.5, Dec. 1960, p.563).

(13) W.Feller, "An introduction to probability theory and its applications, vol.I", John Wiley & Sons, 1950

(14) H.Cramer, "Mathematical methods of statistics", Princeton University Press, 1946.

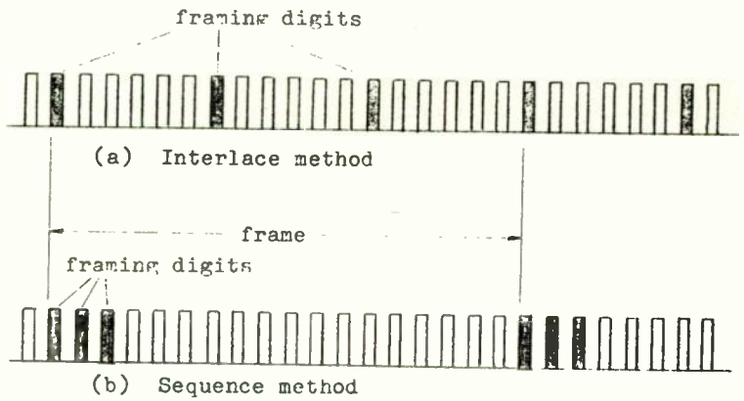


Fig. 1. Methods of distributing the framing digits in a frame.

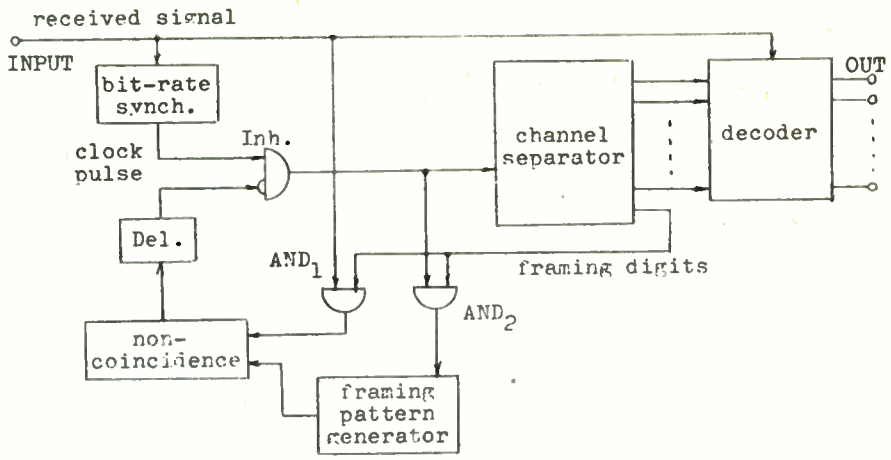


Fig. 2. A digital synchronizer using bit-by-bit shifting method.

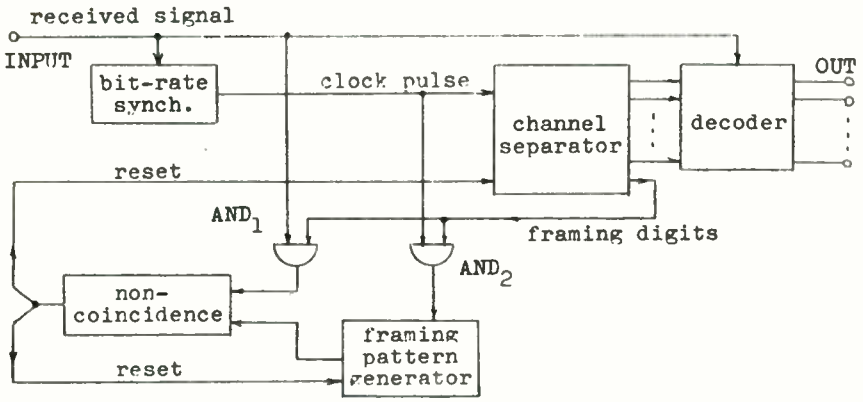


Fig. 3. A digital synchronizer using resetting method.

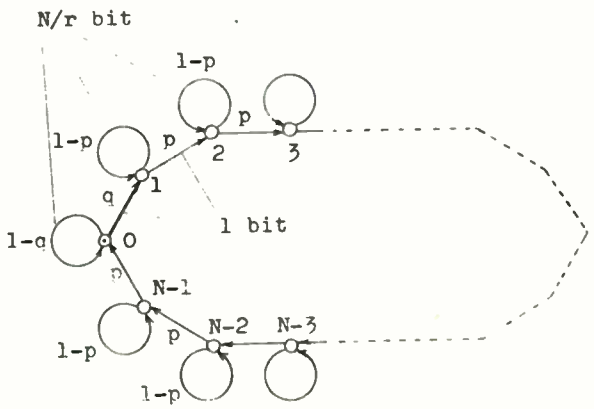


Fig. 4. State transition graph of the bit-by-bit shifting interlace system.

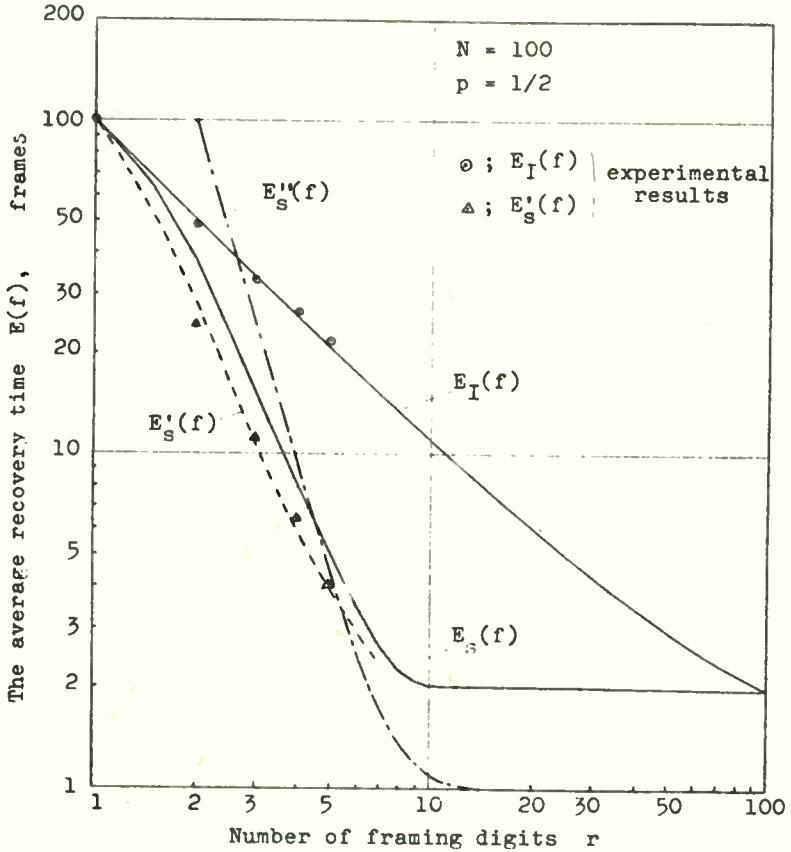


Fig. 5. The average recovery time of the synchronizers.

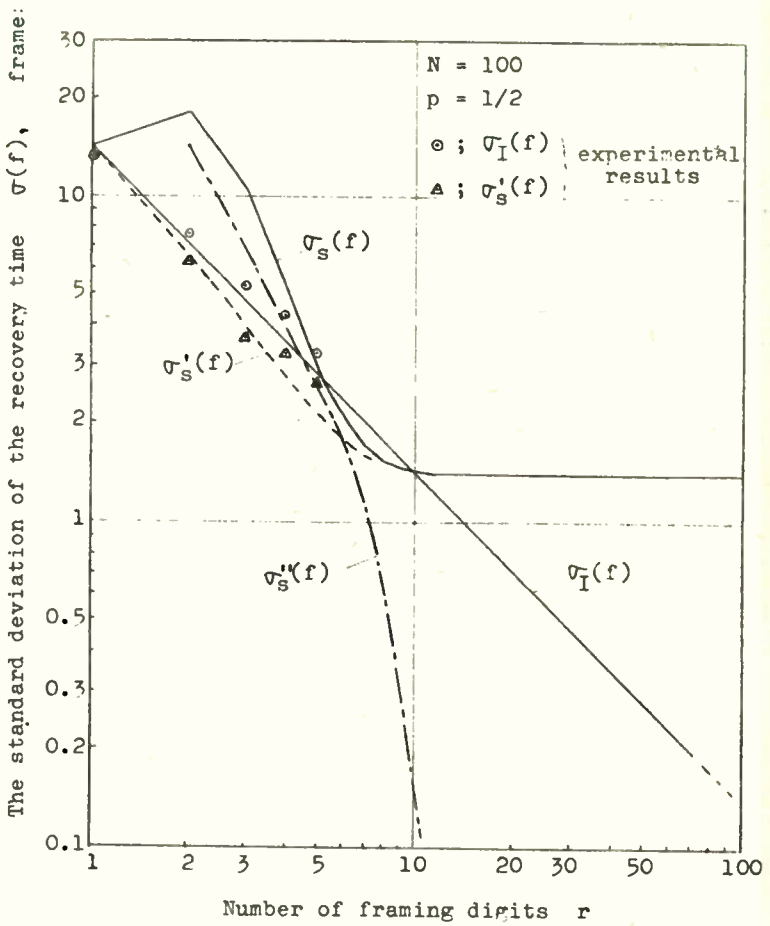


Fig. 6. The standard deviations of the recovery time of the synchronizers.

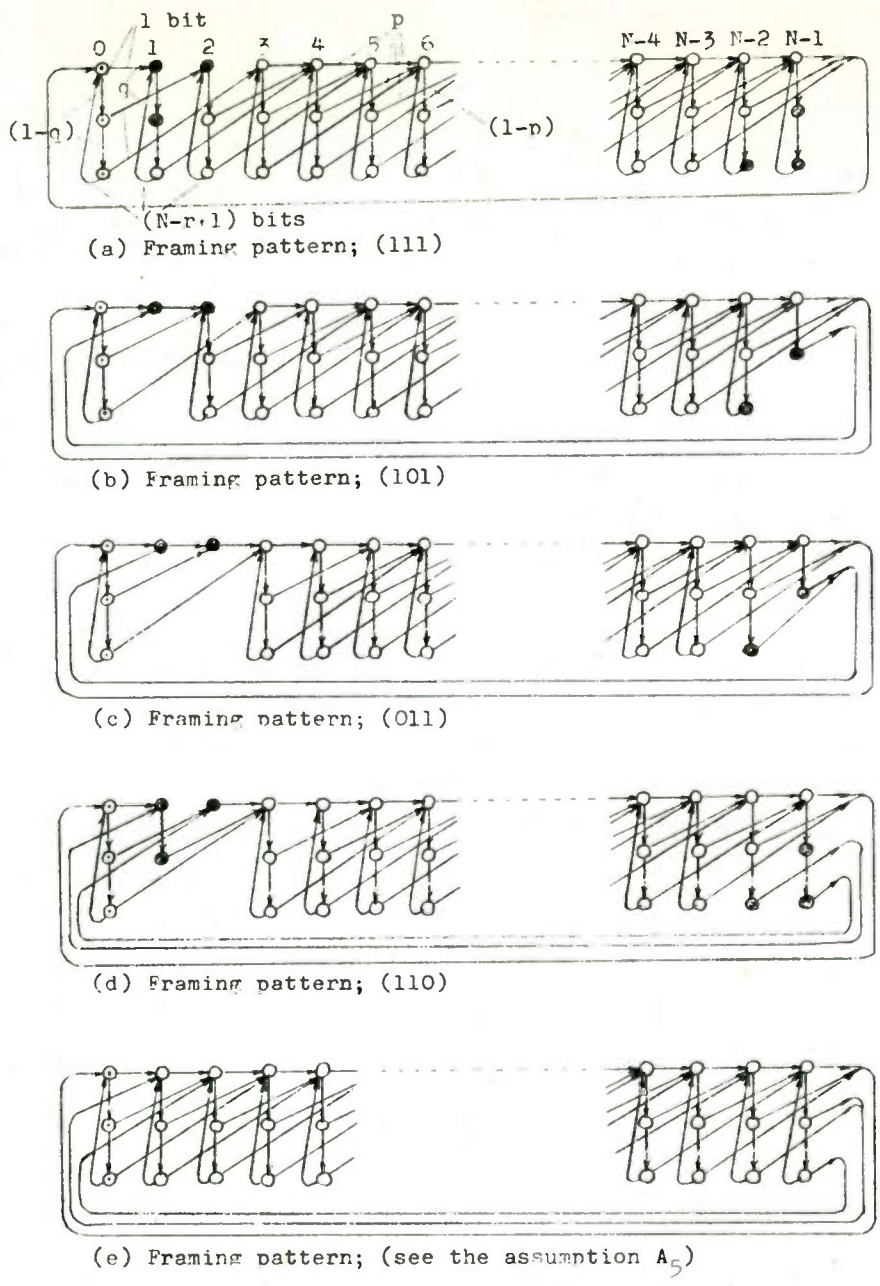
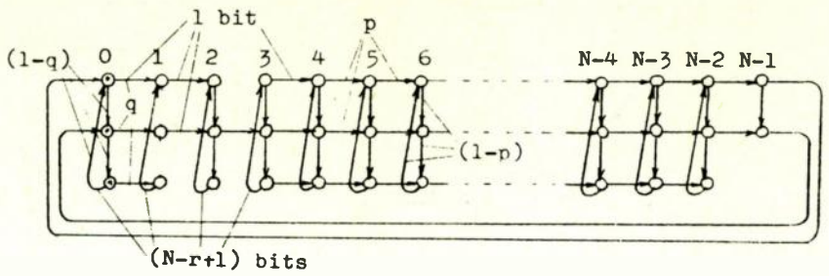
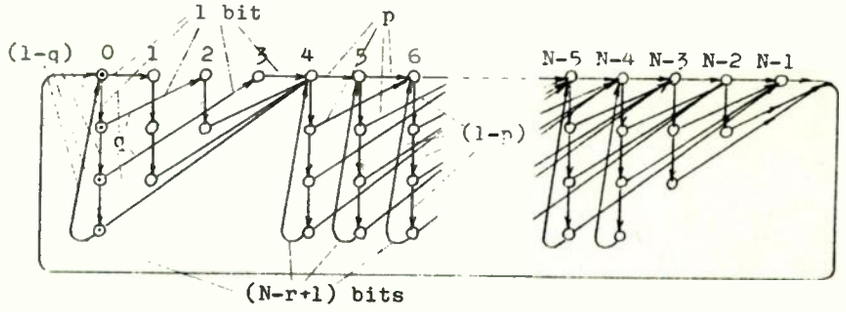


Fig. 7. State transition graphs of the resetting sequence system.

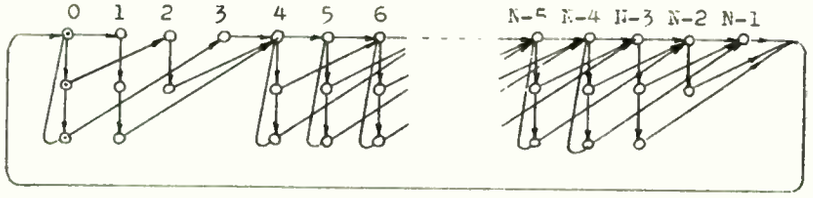


Framing pattern; (101)

Fig. 8. State transition graph of the bit-by-bit shifting sequence system.



(a) Framing patterns;
$\underline{S}$ = (01110), $\underline{R}$ = (*1110)



(b) Framing patterns;
$\underline{S}$ = (01110), $\underline{R}$ = (*111)

Fig. 9. State transition graphs for the optimal framing patterns of the resetting sequence system.

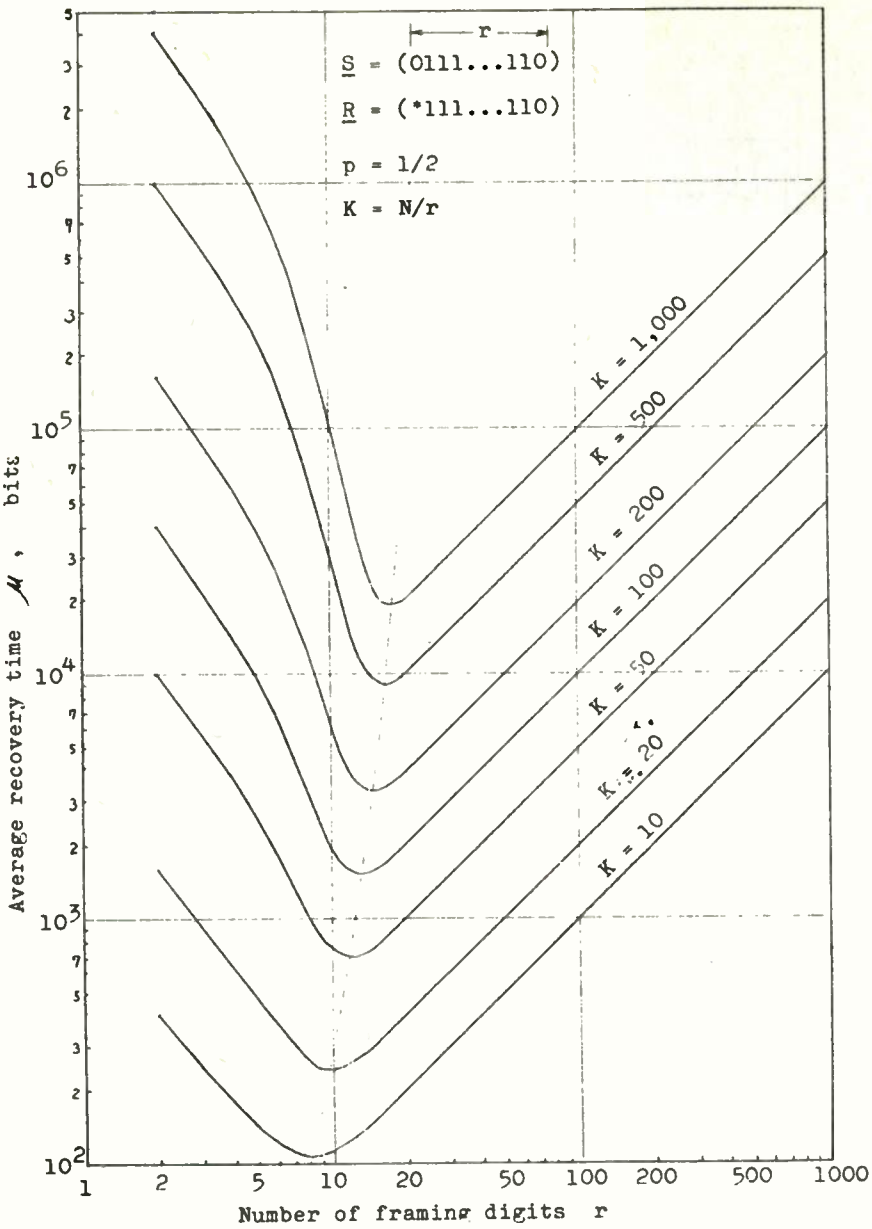


Fig. 10. Minimization of the average recovery time.

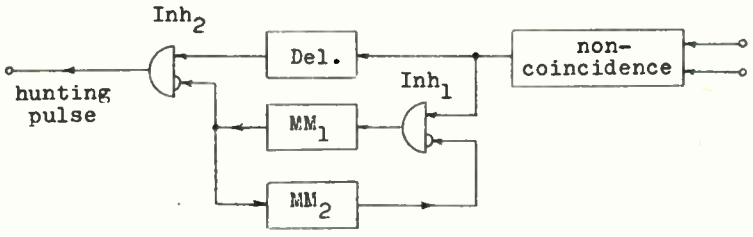


Fig. 11. An example of a locking circuit.

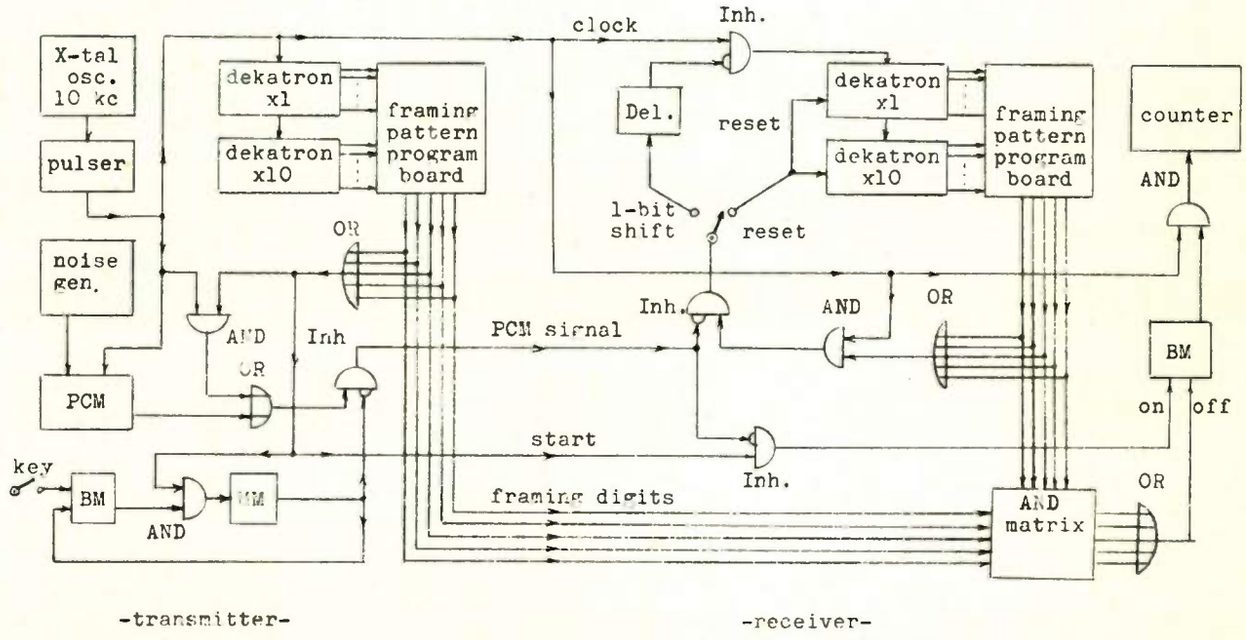


Fig. 12. Block diagram of the recovery process simulator.

COMPARATIVE PERFORMANCE OF DIGITAL DATA TRANSMISSION SYSTEMS IN THE PRESENCE OF CW INTERFERENCE*

Frank G. Splitt
Cook Technological Center
Division of Cook Electric Company
Morton Grove, Illinois

Summary

During the last decade there has been a considerable amount of work done on the analysis of digital data transmissions in an interference environment. With few exceptions, the type of additive interference considered is normal noise. This paper treats the case where the interference consists of CW that falls within the passband of the receiver. The results obtained can also be utilized to ascertain system performance in the presence of interference from similar systems and for certain classes of interrupted CW (ICW) interference.

The performance of differential phase-shift, coherent phase-shift, amplitude keyed, frequency shift and time-shift binary digital data transmissions is determined. The average digit error probability is taken as the measure of system performance. The element of uncertainty involved in the CW interference analysis is due to random variation in the phase angle between the signal and interference. The performance of systems employing angle modulation is found to be independent of this phase angle for certain values of the interference frequency. In general, it is shown that the digit error probability varies as the arc cosine of a function of the signal-to-interference ratio, the function being dependent on the system under consideration.

It is shown that, of the systems investigated, coherent phase-shift keying is the best transmission technique in the face of CW interference. The digit error probability in a differential phase-shift keyed transmission is found to be a sensitive function of the frequency of the interfering tone, and depending on this frequency, can be quite different from that obtained in a coherent phase-shift keyed transmission. Frequency-shift keying is found to exhibit the lowest threshold of the non-coherent systems considered, while time-shift keying is found to give an improvement of approximately 4 db over a simple amplitude-keyed system.

*Portions of this paper are based on material contained in a dissertation to be submitted to the Graduate School of Northwestern University in partial fulfillment of the requirements for the Ph. D. degree.

Introduction

Recent years have seen an ever increasing interest in the art of digital data transmission. Since one of the major objectives of these systems is high reliability, there has arisen the question of system performance in the face of interference. This question has received considerable attention in the field of statistical communication theory.[1-5] A review of the cited references as well as other work in this area, indicates that the type of additive interference* considered is almost always limited to that of normal noise. Though most authors admit that interference other than normal noise must be considered in choosing a digital transmission, the fact remains that there has been little, if any, information available upon which to base this consideration. Thus, there has been a tendency in some circles to consider system performance only in terms of a normal noise background, when in practice normal noise is only one of many forms of additive interference confronting data transmission systems.

This paper determines the effects of another type of interference, namely, CW, so as to provide a broader view of the relative performance characteristics of several binary data transmission systems. The systems considered are differential phase-shift keying, ΔPSK; coherent phase-shift keying, PSK; amplitude keying, AK; frequency-shift keying, FSK; and time-shift keying, TSK.** The latter transmission technique is of particular interest since it permits the use of an optimum threshold receiver with an "on-off" type transmission. Furthermore, the time structure of the detected TSK signal is such that errors due to pulse sag would be virtually eliminated.

*In this paper attention will be restricted to additive types of interference.

**This transmission technique is also known as "diphase" and binary pulse position coding.

General System Considerations

Selectivity

In the following analysis, consideration will be given to both coherent and noncoherent systems. As is well known, coherent systems are capable of achieving all of their selectivity after the detection process while noncoherent systems must resort to predetection filtering if the system is to be at all efficient in the face of wide band interference. It is important to note that differentially coherent systems also require predetection filtering.

It will be assumed here that both the PSK and the ΔPSK systems use a matched filter for selectivity purposes. This choice is not critical. However, this filter characteristic is selected since it would normally be used to enhance the performance of these systems in a white noise background. The matched filter has the property that the signal-to-noise ratio is maximized at its output at some given instant of time. This time will be taken as the time corresponding to the termination of the input digit.

In the case of noncoherent systems, it will be assumed that a linear bandpass filter, having a "selectivity function," $G(\omega)$, is utilized prior to the detector. The selectivity function is nothing more than the familiar normalized magnitude of the voltage transfer ratio,* i.e.,

$$G(\omega) = \left| \frac{E_o(\omega)}{E_i(\omega)} \right| \div \left| \frac{E_o(\omega)}{E_i(\omega)} \right|_{max} \qquad (1)$$

where $|E_o(\omega)|$ is the peak amplitude of the filter output, given a sinusoidal input of angular frequency ω and peak amplitude $|E_i(\omega)|$. The fact that the midband "gain" of the IF filter is thus taken as unity, will not result in a loss of generality in the analysis of incoherent systems.

Ideally, the bandpass filter would be chosen to provide maximum discrimination against undesired signals and noise while still preserving the essential qualities of the desired signal. Except for the fact that the filter type selected should not lead to intersymbol influence, the exact nature of the filter is not really pertinent to the ensuing discussion. However, the filter type must approximate, to a reasonable degree, the selectivity function that one might anticipate in practice if the final results are to be at all meaningful. One such filter type[6] is the so-called gaussian filter for which

$$G(\omega) = \exp\left[-\left(\frac{\omega - \omega_o}{\omega_c} \right)^2 \right] \qquad (2)$$

where $\omega_o = 2\pi$ times the center frequency of the filter, $\omega_c = \frac{2\pi}{T}$ and T is the duration of a message digit.

Detectors

The phase information in the PSK and ΔPSK systems is extracted by means of an ideal product detector. This detector is characterized by the following input-output relation,

$$X_o(t) = X_{i_1}(t) X_{i_2}(t) \qquad (3)$$

where $X_o(t)$ is the detector output and $X_{i_1}(t)$, $X_{i_2}(t)$ are its two inputs.

The detectors utilized in the noncoherent systems are considered to be of the instantaneous linear envelope type, i.e., the detector output is just the envelope of the input. This, of course, implies the existence of a low-pass zonal filter having a cutoff frequency less than the IF but high enough to preserve the essential time structure of the detector output.

Decision Device

All systems under consideration are assumed to be fully synchronous. By this is meant that a device called the Decider can sample the output of the product detector or the linear envelope detector at precisely the instant of maximum signal. The Decider bases its decision, as to whether a Mark or a Space has been transmitted on the value of these samples relative to a reference level. It will be assumed that the reference level is that which would normally be used in practice. With the exception of the AK system this level is also optimum from the statistical decision theory point of view when the interference is white gaussian noise. Optimality is here taken to mean the minimization of the digit error probability.[7, 8] The threshold level selected for the AK receiver is the average value of the detector output waveform. This threshold level is generally used with such systems since the optimum threshold would be very difficult to implement.

Differential Phase - Shift Keying

The ΔPSK System

In this paper the following ΔPSK transmission convention is adopted:

*Since the noncoherent systems will utilize detectors that are insensitive to phase variations, only the magnitude of the voltage transfer ratio is of interest.

(1) No phase change corresponds to a Space
(2) Phase change of π radians corresponds to a Mark.

So that each pulse can act as a phase reference for its successor, it is required that the product of the angular carrier frequency, ω_o, and the duration of a data digit, T, be an even multiple of π radians. The receiver is assumed to consist of two channels both of which contain an extremely high Q filter tuned to the center frequency of the transmission. The impulse response, h(t), of such a filter is well approximated by

$$h(t) = \left(\frac{2}{T}\right) U(t) \cos \omega_o t \qquad (4)$$

where

T = duration of a data digit
ω_o = (angular) RF carrier frequency
$U(t)$ = unit step function

and where the factor $\left(\frac{2}{T}\right)$ has been selected for mathematical convenience. Alternate data pulses are gated to the two filters. Each of the filters is de-energized 2T seconds after its corresponding data pulse is first applied to its input terminals. The outputs of both filters are fed to an ideal product detector. The output of the product detector is assumed to be stripped of its high frequency content by an ideal zonal filter and then passed on to a synchronous decision device; i.e., the Decider. The threshold level for the Decider is taken as zero volts.

CW Error Analysis

Since we are dealing with a differential phase-shift system there are two ways in which a Space can be transmitted and two ways in which a Mark can be transmitted. The phase relations corresponding to these possibilities are (0, 0) or (π, π) for a Space and (0, π) or (π, 0) for a Mark. It can be readily demonstrated that each of the two Space and Mark phase combinations lead to the same error probabilities; therefore, only one of the Space cases and one of the Mark cases need be considered.

We first consider the case of a (0, 0) type Space transmission that has been perturbed by CW interference. The gated input to the first predetection filter is thus assumed to be of the form

$$e_1(t) = \left[U(t) - U(t-T)\right] \cdot \left\{A \cos \omega_o t + B \cos\left[(\omega_o + \Delta\omega)\, t + \phi\right]\right\}$$

where

A = peak signal amplitude
B = peak interference amplitude
$\Delta\omega$ = angular frequency displacement of the interfering tone with respect to the RF carrier, ω_o
ϕ = RF phase angle between the signal carrier and the interference at the start of the reference pulse

The input to the second filter would then be given by:

$$e_2(t) = \left[U(t-T) - U(t-2T)\right] \cdot \left\{A \cos \omega_o t + B \cos\left[(\omega_o + \Delta\omega)\, t + \phi\right]\right\}$$

The output of the first filter during the time interval $T \le t \le 2T$ can be written as

$$y_1(t) = A \cos \omega_o t + B \left[\frac{\sin \frac{\Delta\omega}{2} T}{\frac{\Delta\omega}{2} T}\right] \cdot \cos\left(\omega_o t + \phi + \frac{\Delta\omega}{2} T\right) \qquad (5)$$

while the output of the second filter during the same time interval is given by

$$y_2(t) = \frac{A(t-T)}{T} \cos \omega_o t + \frac{B(t-T)}{T} \left[\frac{\sin \frac{\Delta\omega}{2}(t-T)}{\frac{\Delta\omega}{2}(t-T)}\right] \cdot \cos\left[\left(\omega_o + \frac{\Delta\omega}{2}\right)(t-T) + (\omega_o + \Delta\omega T) + \phi\right] \qquad (6)$$

In the above, use has been made of the fact that the output of a filter, having an impulse response given by Equation (4), to a time gated input signal,

$$e_i(t) = B\left\{U\left[t - (k-1)T\right] - U(t-kT)\right\} \cos\left[(\omega_o + \Delta\omega)\, t + \phi\right],$$

can be written as

$$y_o(t) = \frac{BL}{T}\left\{\left[\frac{\sin \frac{\Delta\omega}{2} L}{\frac{\Delta\omega}{2} L}\right] \cdot \cos\left[\omega_o t + \frac{\Delta\omega}{2} L + \Delta\omega\,(k-1)\, T + \phi\right]\right\}$$

$$L = \begin{cases} 0 & , \quad t < (k-1)T \\ t-(k-1)T & , \quad (k-1)\,T \le t \le kT \\ T & , \quad t > kT \end{cases},$$

when $\Delta\omega << \omega_o$, $\omega_o T >> 1$. The product of $y_1(t)$ and $y_2(t)$ after elimination of high frequency terms and sampling at time, $t = 2T$, is given by

$$Y_S = \frac{A^2}{2} + \frac{B^2}{2}\mathcal{L}^2 \cos \Delta\omega T$$

$$+ AB\mathcal{L}\cos\frac{\Delta\omega}{2}\,T\cos(\Delta\omega T + \phi). \qquad (7)$$

where

$$\mathcal{L} \equiv \frac{\sin\frac{\Delta\omega}{2}\,T}{\frac{\Delta\omega}{2}\,T}$$

Since it has been assumed that a Space has been transmitted the Decider will err if Y_S is less than zero. Clearly, when $\Delta\omega T$ is an even number (including zero) of π radians Y_S will always be greater than zero. Consequently, the probability of error on a transmitted Space would be zero for these interfering frequencies. When $\Delta\omega T$ is an odd multiple of π radians ($\Delta\omega T = (2m+1)\pi$, $m = 0, 1, 2, \ldots$) one finds that an error will occur with probability equal to one if the peak signal-to-interference ratio, A/B, is less than $2\left[(2m+1)\pi\right]^{-1}$. For other values of $\Delta\omega T$ consideration must be given to the initial RF phase angle ϕ corresponding to the pair of digits in question. This phase angle will, in general, be different for each set of consecutive digits considered. These variations in the initial digit-phase angle can, to a certain degree, be attributed to the fact that the interfering tone is "off" frequency. However, in any practical situation, these variations would still be present if the tone was "on" frequency and can be ascribed to either phase jitter in the interfering tone or other transmission anomalies not considered here.*

The effect of this phase angle on the digit error probability for the case of a transmitted Space is illustrated by means of phasor diagrams in Fig. 1. In these diagrams, the components of the signal-plus-interference phasor corresponding to the reference pulse are $\hat{y}_{1S}$ and $\hat{y}_{1I}$. The resultant phasor is designated as $\hat{y}_1$. Similarly, the components of the signal-plus-interference phasor which corresponds to the data pulse are $\hat{y}_{2S}$ and $\hat{y}_{2I}$. The resultant of these two phasors is $\hat{y}_2$. In this illustration the peak signal-to-peak interference ratio (A/B) is 1.2 and the angular frequency difference between the center frequency of the transmission and that of the interfering tone is such that $\Delta\omega T$ is equal to 1.1 radians. The Decider will err if the absolute value of the angle between the phasors $\hat{y}_1$ and $\hat{y}_2$ exceeds $\pi/2$ radians. Figure 1(a) portrays a situation where ϕ is such that the Decider will err while Fig. 1(b) illustrates a case where the correct decision will be made.

Since there is no reason why one initial phase angle should be any more likely than another it is treated as a random variable having a uniform distribution over the range 0-2π. Having defined the distribution for ϕ, the digit error probability can now be determined. From Equation (7) and the defined threshold it is seen that the probability of error, given that a Space has been transmitted is given by

$$P_S(E) = \Pr\left\{\cos\phi > \frac{A^2 + B^2\mathcal{L}^2}{2\left|\mathcal{L}\cos\frac{\Delta\omega}{2}\,T\right|}\right\} \qquad (8)$$

when

$$\Delta\omega T \neq m\pi, \quad m = 1, 2, 3, \ldots$$

Now,

$$\Pr\left\{\cos\phi > a\right\} = \begin{cases} 0 & ; a > 1 \\ 1/\pi \arccos a & ; -1 \le a \le 1 \\ 1 & ; \text{otherwise} \end{cases} \qquad (9)$$

when ϕ is uniformly distributed over the range 0 - 2π. Thus the conditional Space error probability is given by,

$$P_S(E) = \begin{cases} 0 & ; \Delta\omega T \neq (m+1)\pi,\ a > 1 \\ & \Delta\omega T = 2(m+1)\pi,\ Z > 0 \\ & \Delta\omega T = (2m+1)\pi,\ Z > \frac{2}{(2m+1)\pi} \\ \frac{1}{\pi}\arccos a_S & ; \Delta\omega T \neq (m+1)\pi,\ -1 \le a \le 1 \\ 1 & ; \text{otherwise} \end{cases} \qquad (10)$$

where $Z = \frac{A}{B}$, $m = 0, 1, 2, 3, \ldots$

and

$$a_S = \frac{Z + Z^{-1}\mathcal{L}^2}{2\left|\mathcal{L}\cos\frac{\Delta\omega}{2}T\right|}$$

In the case of a Mark transmission, the sampled value of the detector output after zonal filtering is

---

*Doppler frequency shifts or multipath disturbances are two such effects.

$$Y_M = -\frac{A^2}{2} + \frac{B^2}{2}\mathcal{L}^2 \cos \Delta\omega T$$

$$-AB\mathcal{L} \sin \frac{\Delta\omega}{2} T \sin (\Delta\omega T + \phi) \qquad (11)$$

Following the same argument as in the case of a Space transmission, it is found that the probability of error, given that a Mark has been transmitted, is

$$P_M(E) = \begin{cases} 0 & ; \Delta\omega T \neq 2m\pi,\ a > 1 \\ & \Delta\omega T = 0,\ Z > 1 \\ & \Delta\omega T = 2(m+1)\pi,\ Z > 0 \\ \frac{1}{\pi} \arccos a_M & ; \Delta\omega T \neq 2m\pi,\ -1 \leq a \leq 1 \\ 1 & ; \text{otherwise} \end{cases} \qquad (12)$$

where

$m = 0, 1, 2, 3, \ldots$

and

$$a_M = \frac{Z - Z^{-1}\mathcal{L}^2 \cos \Delta\omega T}{2\left|\mathcal{L} \sin \frac{\Delta\omega}{2} T\right|},$$

If Marks and Spaces are taken to be equally likely, then the average digit error probability, $\overline{P(E)}$, is equal to one-half of the sum of the Mark and Space digit error probabilities. The average digit error probability is shown in Fig. 2 as a function of the signal-to-interference ratio for two values of the angular frequency difference between the RF carrier and CW interference. One of the values, namely, $\Delta\omega = \frac{1.3}{T}$ is the angular frequency displacement to which the ΔPSK system is most sensitive.

Coherent Phase Shift Keying

The coherent PSK system analysis is identical to that for ΔPSK except that the phase reference is locally generated and is, therefore, free from interference. It is easy to show that the Mark and Space digit error probabilities are equal and that the average digit error probability is given by

$$\overline{P(E)} = \begin{cases} 0 & ; \Delta\omega T \neq 2m\pi,\ a > 1 \\ & \Delta\omega T = 2m\pi,\ Z > 0 \\ \frac{1}{\pi} \arccos a & ; \Delta\omega T \neq 2m\pi,\ -1 \leq a \leq 1 \\ 1/2 & ; \text{otherwise} \end{cases}$$

where

$a = Z\left|\mathcal{L}\right|^{-1}$, Z = peak signal-to-peak interference ratio

and

$m = 1, 2, 3, \ldots$ This expression is shown plotted in Fig. 2. The curve for $\Delta\omega = 0$ corresponds to the case of most effective CW interference while that for $\Delta\omega = \frac{1.3}{T}$ is shown for the sake of comparison with ΔPSK.

Noncoherent Amplitude Keying

The AK System

The AK receiver considered here utilizes a predetection filter of selectivity, $G(\omega)$, centered about the IF. The detector is of the linear envelope type and the threshold level is taken as the average value of the detected output signal. As mentioned previously, this threshold level is commonly used because of the difficulty of implementing the optimum threshold mechanism. In this case the Decider registers a Mark if the sampled signal is greater than the threshold and a Space if it is less than the threshold. Since it will be of interest to compare all systems on the basis of the same mean-square transmitted power, the peak signal in the AK transmission is taken as $\sqrt{2}$ A rather than A.

CW Error Analysis

Consider first the case of a transmitted Mark. Here the input to the Decider is

$$Y_M = \left\{ 2A^2 + \left[BG(\omega_I) \right]^2 + 2\sqrt{2}\, ABG(\omega_I) \cos \phi \right\}^{1/2} \qquad (14)$$

where ϕ is a random phase angle having a uniform distribution between zero and 2π radians and $G(\omega_I)$ is the value of the selectivity function at the interference frequency. If a Space is transmitted, the Decider input would be given by

$$Y_S = BG(\omega_I) \qquad (15)$$

These inputs can be put in normalized form by dividing by $\sqrt{2}$ A giving,

$$Y'_M = \left[1 + X^2 + 2X \cos \phi \right]^{1/2} \qquad (16)$$

and

$$Y'_S = X \qquad (17)$$

where

$X = \frac{BG(\omega I)}{\sqrt{2}\,A}$. The average value of the detected signal is thus given by

$$\overline{Y} = \frac{1}{2}\left\{ X + \frac{1}{2\pi} \int_0^{2\pi} \left[1 + X^2 + 2X \cos \phi \right]^{1/2} d\phi \right\} \qquad (18)$$

The indicated integration can be carried out by means of a term by term integration of the appropriate series expansion for the radical. The resulting series are rapidly convergent and can, therefore, be approximated by their first few terms. Proceeding on this basis it can be shown that the probability of making an error when a Mark is transmitted is

$$P_M(E) = \begin{cases} 0 & ; a > 1;\ X = 0 \\ \frac{1}{\pi} \arccos a & ; 1 \le a \le 1 \\ 1 & ; \text{otherwise} \end{cases} \qquad (19)$$

where

$$a = \begin{cases} \frac{3}{8X} - \frac{1}{4} + \frac{5X}{16} - \frac{X^2}{16} - \frac{3X^3}{256} - \frac{X^4}{256} - \frac{X^5}{512} \cdots ; \\ X \le 1, \quad X \neq 0 \\ \frac{3}{8X} - \frac{1}{64X^3} - \frac{3}{1024X^5} - \cdots ; X > 1 \end{cases}$$

Suppose now that a Space is transmitted. The digit error probability is given by

$$P_S(E) = \Pr\left\{ X \ge \overline{Y} \right\} \qquad (20)$$

which can be shown to be zero for $X > 0$.

It is seen, then, that the average digit error probability in an AK transmission with an average threshold is simply given by one-half the Mark error probability under the CW interference condition. This expression is shown plotted in Fig. 3 for the case where the interference lies in the center of the predetection filter passband. It is to be noted that when the frequency of the interfering tone deviates from the center of the passband, the error threshold* will decrease in a manner dictated by the selectivity function.

Noncoherent Frequency-Shift Keying

The FSK System

In the system to be considered, two predetection filters are utilized, one centered about the Mark frequency, ω_M, and one centered about the Space frequency, ω_S. The selectivity functions for these filters are designated as $G_M(\omega)$ and $G_S(\omega)$, respectively. These selectivity functions are assumed to differ only in the fact that they are tuned to different frequencies. For the sake of generality, these filters will here be considered to overlap at the center frequency of the transmission, i.e., at $\omega_o = (1/2)(\omega_M + \omega_S)$. Both the Mark and Space filters are followed by a linear envelope detector. The output of the Mark detector is subtracted from that of the Space detector. The difference voltage is then passed on to a synchronous Decider which judges a digit to be a Mark if the sample is of negative polarity and a Space if the sample is of positive polarity.

CW Error Analysis

Suppose that a Space in the form of a sinusoid of amplitude A, angular frequency ω_S and duration T is received and that an interfering tone of peak amplitude B and angular frequency $\omega_I = \omega_o \pm \Delta\omega$ falls within the passband of the receiver. The sampled difference voltage would then be given by

$$Y_S = \left\{ A^2 + \left[BG_S(\omega_I) \right]^2 + 2\, ABG_S(\omega_I) \cos \phi \right\}^{1/2} - BG_M(\omega_I) \qquad (21)$$

where ϕ is a random phase angle having a uniform distribution between zero and 2π radians. An error will occur if $Y_S < 0$, i.e., if

$$\cos \phi < \frac{Z^{-1}\left[G_M^2(\omega_I) - G_S^2(\omega_I) \right] - Z}{2\, G_S(\omega_I)} \qquad (22)$$

where $Z = \frac{A}{B}$. Thus the digit error probability given that a Space is transmitted, would be given by Equation (9) with

$$a = \frac{Z - Z^{-1}\left[G_M^2(\omega_I) - G_S^2(\omega_I) \right]}{2\, G_S(\omega_I)} \qquad (23)$$

A similar result is obtained for $P_M(E)$ if one interchanges $G_M(\omega_I)$ and $G_S(\omega_I)$ in the above expression for a. It is seen, then, that Mark and Space error probabilities are, in general, different so that the average digit error probability must be given by the weighted sum of these two probabilities.

Figure 3 shows this digit error probability for the worse case; i.e., when the interfering tone is coincident with either the Mark or Space frequency. This curve is also valid for interference that is slightly off these frequencies. As the interference deviates from either ω_M or ω_S toward ω_o, the error threshold drops to a value determined by the filter crossover point.

*The error threshold is here taken to mean the signal-to-interference ratio at which errors first appear.

Noncoherent Time Shift Keying

The TSK System

In a simple time shift keyed transmission, the time interval corresponding to an information digit is divided into two equal intervals. A convention as to the representation of either of the two possible digit states is then selected by assigning a pulse to one of the two time intervals. Here a pulse in the first interval will correspond to a Mark. The resulting waveform is then transmitted by "on-off" keying, i.e., AK. If the baseband signal were a continuous "on" or a continuous "off," the TSK baseband signal would consist of a square wave having a period, T, where T is the duration of a message digit. In general, the TSK baseband waveform would appear as is shown in Fig. 4.

Upon reception, the TSK signal is passed through a single predetection filter which, of course, would be twice as wide as that required to pass an ordinary "on-off" transmission. As in the case of previous noncoherent systems this filter will be characterized by a selectivity function $G(\omega)$. The filtered signal is detected by an instantaneous linear envelope detector. The message function, i.e., the original baseband signal is then recovered by passing the detected signal through a synchronous Decider. The Decider observes the level of the signal in each of the two time intervals corresponding to a baseband digit. A Mark is declared if the level in the first interval is greater than that in the second. A Space is declared if these levels are reversed.

This method of transmission has an obvious advantage over ordinary "on-off" keying since the optimum threshold level is independent of the signal strength. This transmission technique has the further advantage (over other digital data transmission techniques) that the detected signal is the TSK baseband waveform which contains very little low frequency energy. Thus the signal can be routed by a network having a poor low-frequency response without incurring errors due to pulse sag. The importance of this consideration in the design of military PCM equipment was pointed out by Brogle[9] in a discussion of a biternary transmission technique.

CW Error Analysis

Since the TSK transmission has a 50% duty cycle the peak of the TSK "on" pulse is taken as $\sqrt{2}$ A as was done in the case of AK. Now from the symmetry of the TSK system it can be seen that the digit error probabilities for Marks and Spaces are identical so that only one case need be considered. Suppose that a Space is transmitted and as before the interfering tone has an amplitude B and is of arbitrary frequency ω_I. The two samples upon which the Decider must judge are,

$$Y_1 = BG(\omega_I) \tag{24}$$

for the first interval and

$$Y_2 = \left\{2A^2 + \left[BG(\omega_I)\right]^2 + 2\sqrt{2}\,ABG(\omega_I)\cos \phi\right\}^{1/2} \tag{25}$$

for the second interval, where ϕ is a random variable having a uniform distribution in the range $0 - 2\pi$. An error will occur if $Y_1 > Y_2$, i.e., if

$$\cos \phi > \frac{Z}{\sqrt{2}\ G(\omega_I)}$$

where $Z = A/B$.

Thus, the average digit error probability in a CW perturbed TSK transmission is

$$\overline{P(E)} = \begin{cases} 0 & ; \quad \dfrac{Z}{\sqrt{2}\ G(\omega_I)} > 1 \\ \dfrac{1}{\pi} \arccos \dfrac{Z}{\sqrt{2}\ G(\omega_I)} & ; \quad 0 \le \dfrac{Z}{\sqrt{2}\ G(\omega_I)} \le 1 \end{cases} \tag{26}$$

This result is portrayed in Fig. 3 for the case where the interference falls in the center of the predetection passband. If the interference is moved off the center frequency, the digit error probability curves will be of the same form, however, the threshold will vary in accordance with the first condition given in Equation 26.

System Comparison

Reference to Figs. 2 and 3 will show that the maximum error threshold for PSK and FSK systems is at zero db. This threshold is the lowest of any of the systems considered here.* Since FSK requires at least three-halves the bandwidth needed for PSK, it would be exposed to more

---

*It is to be noted, e.g., that the effect of a single source of narrow-band interference can be eliminated in an FSK transmission of the type considered here. This can be done by incorporating a "look-through" device which functions to determine the channel perturbed by the interference. Once this has been done the output of the interference free channel can be detected as if the transmission were AK.

sources of narrow band interference. Based on this consideration it is concluded that, of the systems considered, PSK gives the best performance in the presence of CW interference.

Reference to Fig. 2 shows that there is a strong dependence of the ΔPSK digit error probability on the frequency of the interfering tone. Though this dependence can be attributed, in part, to the selectivity of the matched filter, it is primarily due to the frequency sensitivity of the phase comparison device. This frequency sensitivity is the result of the fact that the phase reference is contaminated by the interfering tone and that the phase shift incurred by the interfering tone, as it is "passed through" the delay device, depends on the frequency of the tone. Comparison of the curves for coherent PSK and ΔPSK for the same frequency of the interfering tone will serve to illustrate this effect since the only difference between these two systems lies in the type of phase reference utilized. It is to be noted that the error threshold level, i.e., the signal-to-interference ratio at which errors first appear, in both the PSK and ΔPSK systems varies with the frequency of the interfering tone. Because of the sharp thresholds involved with CW interference (see Figs. 2 and 4) these threshold levels can be used as a measure of relative system performance.

Reference to Equations (10), (12), and (13) will show that ΔPSK and PSK systems of the type discussed will be error free if the interfering tone is displaced from the center frequency of the transmission by an integral multiple of the bit rate, i.e., if $\Delta f = m/T$ where m is an integer other than zero. This is attributable to the fact that the matched filters employed by these systems exhibit nulls in their transfer characteristics for such tones. It is this fact that allows the use of 1/T frequency spacing in FDM-PSK and FDM-ΔPSK systems. When the interference falls exactly in the center of the receiver passband, $\Delta f = 0$, then both PSK and ΔPSK systems exhibit a threshold at a signal-to-interference ratio of zero db (see Fig. 2); however, the average error probability curve for ΔPSK is of the "go - no go" type while that for PSK is much less abrupt. This difference in the shape of the error probability curves of ΔPSK and PSK systems is also displayed when Δf is an odd multiple of the bit rate. As the interfering tone is pulled off the center frequency of the transmission the threshold for PSK decreases while that for ΔPSK increases and then decreases after a peak at $\Delta f = 0.206/T$. It can be shown that the threshold for ΔPSK is always greater than or equal to that for PSK and that the difference in thresholds reaches a maximum value of 3 db when $\Delta f = \frac{0.25}{T}$.

Lawton[3] has shown that in the presence of white gaussian noise, ΔPSK is characterized by an average digit error probability that is quite close (within 2 db) to that of PSK for the range of error probabilities of general interest (less than 10^{-2}). Lawton has also shown that the difference between the two error probabilities approaches zero for large signal-to-noise ratios. Thus, in contrast to the case of white noise interference, CW interference can lead to a 3 db difference in the performance characteristics of PSK and ΔPSK transmission systems even when the digit error probabilities are very low.

In the case of the three noncoherent systems considered, reference to Fig. 3 will show that FSK has an error threshold level that is 3 db less than that of TSK and 6.8 db less than that of AK. Each of the curves shown in this illustration corresponds to interference at a frequency leading to the highest error threshold. In comparing these noncoherent systems it must be remembered that both the FSK and TSK transmissions occupy twice as much spectrum space as does an AK transmission and would, therefore, be exposed to more sources of narrow-band interference.

It can be shown that for white noise interference the digit error probability for FSK is identical to that for TSK.* As was pointed out above, FSK is 3 db superior to TSK in the case of CW interference. Thus, there is also a difference between the relative performance of FSK and TSK systems in the presence of an interfering tone and the relative performance in a white noise background.

Conclusions

This investigation has shown that in the face of CW interference, all of the digital data transmission systems considered exhibit a well-defined error threshold. When the signal-to-

---

*The formal derivations of the digit error probabilities for each of these transmissions are essentially the same. In both cases, the problem resolves itself to that of computing the probability that the value of a noise variate exceeds that of a signal plus noise variate. The mean square value of the noise in the TSK case is twice that of the FSK case because of the difference in predetection filter bandwidths. This difference in noise power is, however, compensated by the fact that the peak signal in the TSK transmission would be $\sqrt{2}$ times that in an FSK transmission when the systems are compared on an equal average power basis.

interference ratio is greater than this threshold no errors can occur. (This assumes, of course, that the system is not overloaded and that the required synchronism has been established.) Once this threshold is reached an increase in the CW interference power level by a fraction of a db can lead to very high ($> 10^{-1}$) average digit error probabilities. This is in marked contrast to the performance of these systems in a white noise environment wherein it is always (theoretically) possible for errors to occur. Thus, it can be seen that the relative performance characteristics of digital data transmission systems depends, to a large measure, on the type of interference considered. These differences in performance characteristics as a function of the type of interference serve to point out that due attention must be paid to the total signal environment before selecting the modulation technique for a digital data transmission system. This fact cannot be too strongly emphasized in this age where radio frequency interference is one of the more vexing problems facing the communication engineer.

As in the case of white noise interference, PSK was found to be the best of the basic digital data transmission techniques. The performance of a simple "on-off", i.e., AK transmission was found to be inferior to all systems considered; however, the use of TSK allows a worthwhile improvement in the performance characteristics of "on-off" type transmissions. This improvement, which is in terms of reduced error probabilities for the same average transmitted power, is due to the receiver's ability to utilize an "optimum" decision level. The TSK transmission offers the further advantage of compatibility with a transmission network having a poor low-frequency response.

Finally, it is to be noted that the results obtained in this paper can also be utilized to ascertain the performance of the various systems considered for several important classes of interrupted continuous wave (ICW) interference. Among these ICW interference signals are the following:

(1) An RF carrier modulated by digit streams that are bit rate synchronous with the desired signal and have bit rates that are the same as or are submultiples of that of the desired transmission

(2) An RF carrier modulated by digit streams that have a bit rate much less than that of the desired signal.

An important example of a type of interference falling in the latter category is the signal from certain types of pulsed radar when the perturbed data transmission system is operating at a very high bit rate.

Acknowledgement

The investigation described in this paper was performed at Northwestern University, Evanston, Illinois. The author would like to acknowledge the assistance of Prof. R.E. Beam of the Department of Electrical Engineering, under whose direction this work was done.

Appreciation is also expressed to the management of the Cook Research Laboratories for support received during the course of this work. Here, special thanks are due Messrs. A.D. Anderson and C.E. Durkee. In addition, the careful review of the final manuscript by Mr. D. Erickson is gratefully acknowledged.

References

1. S.H. Rieger, "Error Probabilities of Binary Data Transmission Systems in the Presence of Random Noise," IRE National Conv. Record, 1953, Part 8, pp 72-79.

2. G.F. Montgomery, "A Comparison of Amplitude and Angle Modulation for Narrow-Band Communication of Binary-Coded Message in Fluctuation Noise," Proc. IRE, Vol. 42, pp 447-454, February 1954.

3. J.S. Lawton, "Comparison of Binary Data Transmission Systems," Proc. IRE Conv. on Military Electronics, 1958, pp 54-61.

4. L.L. Campbell, "Error Rates in Pulse Position Coding," IRE Trans. on Information Theory, IT-3, No. 1, pp 18-24, March 1957.

5. A.B. Glenn, "Performance Analysis of a Data Link System," IRE Trans. on Communications Systems, CS-7, No. 1, pp 14-24, May 1959.

6. L.A. Meacham and E. Peterson, "An Experimental Multichannel Pulse Code Modulation System of Toll Quality," BSTJ, Vol. 27, No. 1, pp 1-43, January 1948.

7. J. Lawson and G. Uhlenback, Threshold Signals, McGraw-Hill Book Company, Inc., New York, N.Y., Section 7.5, 1950.

8. C.W. Helstrom, Statistical Theory of Signal Detection, Pergamon Press, Ltd., London, 1960.

9. A.P. Brogle, "A New Transmission Method for Pulse-Code Modulation Communication Systems, "IRE Trans. on Communication Systems, CS-8, No. 3, pp 155-160, Sept. 1960.

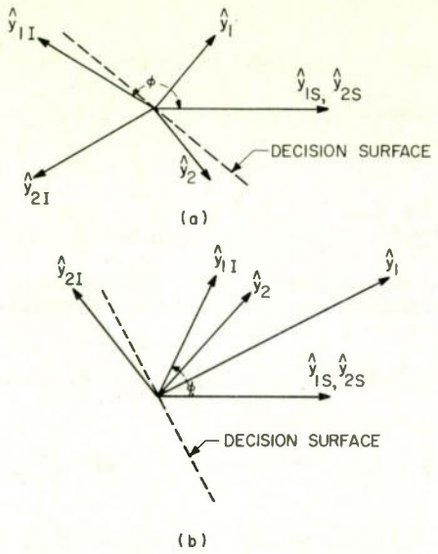


Fig. 1. Phasor diagrams illustrating the effect of the initial digit phase angle, ϕ, on the decision in a Δ PSK transmission. (a) Initial phase angle such that an incorrect decision will be made. (b) Initial phase angle such that a correct decision will be made.

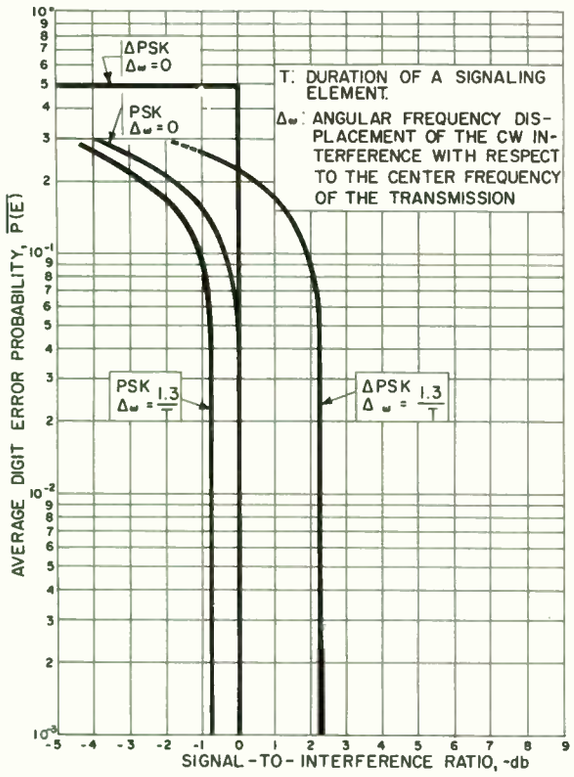


Fig. 2. Digit error probabilities for PSK and Δ PSK transmission systems for two values of the interference frequency.

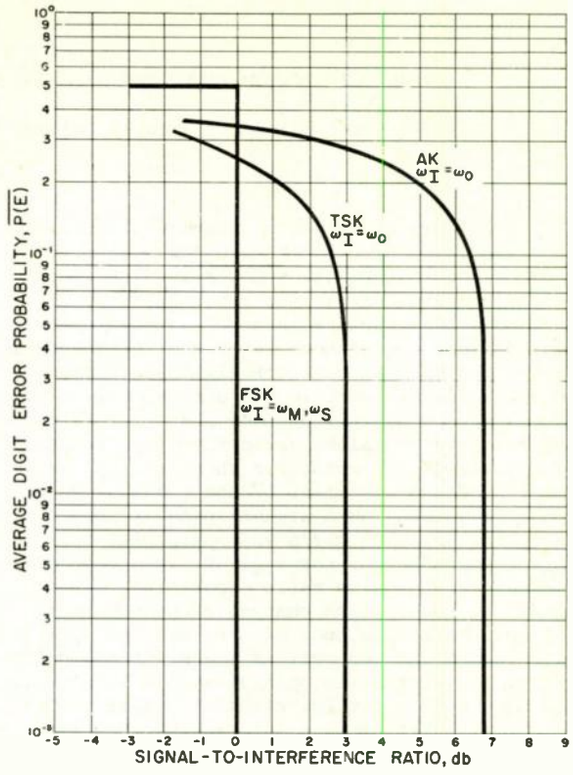


Fig. 3. Digit error probabilities for AK, FSK and TSK transmission systems for selected values of the CW interference frequency, ω_I.

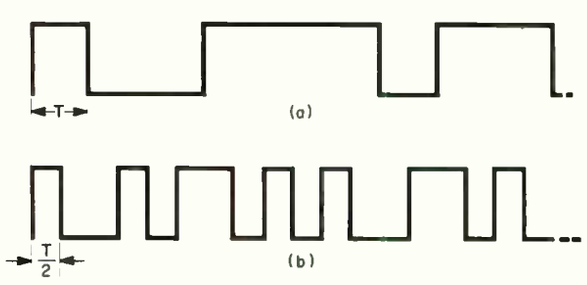


Fig. 4. Waveforms in a TSK transmission system. (a) Information function (baseband signal). (b) Information function after time-shift coding (TSK baseband signal).

ON THE OPTIMUM PERFORMANCE OF N-ARY SYSTEMS HAVING TWO DEGREES OF FREEDOM

R.W. Lucky and J.C. Hancock, Purdue University, School of E.E., Lafayette, Inc.

Abstract

A digital transmission system with n possible transmitted symbols is considered. If the time between transmitted symbols is T_o seconds and the bandwidth is W, the n possible symbols correspond to n vectors in a $2WT_o$ dimensional signal space. This paper considers the theoretical properties of a class of digital systems where the signal space is two dimensional. Such systems are both amplitude and phase modulated. Approximate expressions are derived for the average probability of error for these systems as a function of the placement of the n symbol vectors in the two-dimensional signal space. Optimum placements are then given which minimize this probability of error for a given average or peak power signal-to-noise ratio constraint. It is shown that the optimum channel structure is a function of the alphabet size n, and the type of power constraint, as well as the signal-to-noise ratio. In general the optimum system is a phase modulated system for low signal-to-noise ratios and for alphabet sizes $n \leq 16$ in the high signal-to-noise ratio region. The performance of this optimum system in terms of channel capacity and probability of error is then compared with the performance of one-dimensional systems, AM-only and PM-only, in a complete set of curves for both peak and average power.

Introduction

A generalized communication system employing a discrete channel for the transmission of digital data is shown in Figure 1. The message source is either a function of time or a sequence of symbols. In the event that the message is a function of time, sampling and quantization is assumed. In any case the output of the encoder is a sequence of symbols $\{A_i\}$ which we shall assume is constructed by an independent choice of each successive symbol from a set of n possible symbols A_i. At the transmitter each symbol A_i is made to correspond to a signal waveform $s_i(t)$ suitable for transmission through a noisy medium to the receiver. If the symbols are to be regularly spaced T_o seconds apart, the transmitted waveform is the superposition.

$$s(t) = \sum_k s_i(t - kT_o) \qquad (1)$$

The selection of the optimum set of n possible transmitted waveforms $s_i(t)$ corresponding to the n possible symbols A_i at the input to the transmitter is a difficult problem in communication theory and is not considered in this paper. In this paper we consider a class of communication systems wherein there are two or less degrees of freedom associated with each message waveform $s_i(t)$. In this class will be contained digital amplitude modulation (AM), phase modulation (PM) and combined amplitude and phase modulation (AM-PM).

The narrow band signal s(t) will be written as

$$s(t) = r(t) \cos[\omega_c t + \phi(t)] \qquad (2)$$

or in terms of quadrature components as

$$s(t) = a(t) \cos \omega_c t - \beta(t) \sin \omega_c t \qquad (3)$$

A well-known result of the sampling theorem is that a signal bandlimited to W cps may be completely specified by 2W parameters per second. A particularly useful form of this theorem expresses the signal (2) in terms of samples of phase and amplitude of the carrier every $\frac{1}{W}$ seconds.(1)

$$s(t) = \sum_n r\left(\frac{n}{W}\right) \frac{\sin \pi (Wt - n)}{\pi (Wt - n)} \cos \left[\omega_c t + \phi\left(\frac{n}{W}\right)\right] \qquad (4)$$

Alternatively, both quadrature components of the signal may be specified every $\frac{1}{W}$ seconds.

$$s(t) = \sum_n \alpha\left(\frac{n}{W}\right) \frac{\sin \pi (Wt - n)}{\pi (Wt - n)} \cos \omega_c t$$

$$- \sum_n \beta\left(\frac{n}{W}\right) \frac{\sin \pi (Wt - n)}{\pi (Wt - n)} \sin \omega_c t \qquad (5)$$

In any event there are $2WT_o$ samples of the transmitted signal specified in a message of length T_o seconds. This means that each message waveform $s_i(t)$ has $2WT_o$ degrees of freedom associated with it. The selection of the waveforms $s_i(t)$ is equivalent to the selection of n vectors in a $2WT_o$ dimensional signal space. The probability of incorrect reception of any individual waveform $s_i(t)$ is a function of its distance in signal

space from the other waveforms $s_j(t)$. If the dimensionality of the signal space is greater than or equal to the alphabet size, n, an optimal distribution of the signal vectors may be effected by a method discussed by Kotelnikov(2). In this scheme the vectors are of equal length, equidistant from one another, and have a zero average along any coordinate axis.

In practice there as yet has been little use of signal spaces of greater than a single dimension. Frequency shift keying, phase shift keying, amplitude keying, and time shift keying are systems wherein a single degree of freedom is ascribed to each transmitted symbol. A phase shift keyed system is obtained from the signal, Equation (4), by keeping the amplitude samples $r(n/W)$ constant. Each symbol A_i is represented by a discrete value of the phase sample, ϕ_i. An amplitude keyed system is obtained in like manner by maintaining constant the phase samples and assigning each symbol A_i to a discrete amplitude value r_i. In either of these systems a degree of freedom of the signal is unused at each sample point.* The combined AM-PM system presents a two dimensional problem where each symbol A_i is designated by a phase ϕ_i and an amplitude r_i (or by quadrature components α_i and β_i). A comparison of the three systems, AM, PM, and AM-PM is shown in Figure 2 in terms of sample arrangements of 5 vectors in the two dimensional signal space. (The alphabet size of 5 has been chosen only for pictorial simplicity.) The peak power of the carrier at a sample point in any of these systems is $r^2_m/2$.

In the remainder of this paper we will first derive optimum phasor placements for AM-PM systems and then compare the performance of this optimum AM-PM system with the performances of AM-only and PM-only systems. In deriving optimum systems we will assume ideal maximum likelihood detection at the receiver and use average probability of error per symbol as a performance index. Ideal detection constitutes correlation detection by a coherent carrier in the cases of AM and of PM. For the combined AM-PM system coherent detection of the quadrature components of the received signal is optimal.

Asymptotic Expressions for the Average Probability of Error as a Function of Phasor Placement

The received signal will be written as

$$z(t) = v(t) \cos[\omega_c t + \phi(t)] \qquad (6)$$

or

$$z(t) = x(t) \cos \omega_c t - y(t) \sin \omega_c t \qquad (7)$$

With Gaussian, white noise as the perturbing influence in the channel, the joint probability density of a received sample is easily written in terms of quadrature components and the noise power σ^2 contained in the bandwidth W.

$$p_{A_i}(x,y) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x-\alpha_i)^2}{2\sigma^2}} e^{\frac{-(y-p_i)^2}{2\sigma^2}} \qquad (8)$$

This is the conditional density given the symbol A_i has been sent that the quadrature samples x and y are received.

The receiver picks as its estimate of the transmitted symbol the symbol A_i for which the a posteriori probability

$$p(A_i)p_{A_i}(x,y)$$

is greatest. (The a priori probability of the symbol A_i is designated $p(A_i)$.) The situation at the receiver is shown in Figure 3 for the example system of five phasors. The receiver x-y plane is divided into n unique decision regions R_i, each region R_i containing all points (x,y) for which

$$p(A_i)p_{A_i}(x,y) \geq p(A_j)p_{A_j}(x,y) \qquad (9)$$

for all $j \neq i$. When the received phasor falls in the region R_i the decision is made that the symbol A_i has been transmitted. The decision boundaries, such as DD' in Figure 3, are determined by the equalities of Equation (9). Using Equation (8) for the conditional probabilities

$$p_{A_i}(x,y)$$

one can show that the decision boundaries are straight lines perpendicular to the lines r_{ij} connecting states A_i and A_j and at a distance

$$d = \frac{r_{ij}}{2}\left[1 + \frac{2\sigma^2}{r_{ij}^2} \log \frac{p(A_i)}{p(A_j)}\right] \qquad (10)$$

---

*In an AM system this redundancy may be removed by filtering the carrier for SSB transmission.

from the state A_i. If the states A_i and A_j are equally probable at the transmitter, then the decision boundary between these states is the perpendicular bisector of the line r_{ij}. The probability of correct reception of the symbol A_i after it has been transmitted becomes

$$p_{A_i}(A_i) = \int_{R_i} \frac{1}{2\pi\sigma^2} e^{\frac{-(x-\alpha_i)^2}{2\sigma^2}} \cdot e^{\frac{-(y-\beta_i)^2}{2\sigma^2}} dxdy \qquad (11)$$

and the average probability of error is

$$p_e = 1 - \sum_{i=1}^{i} p(A_i) p_{A_i}(A_i) \qquad (12)$$

Thus the probability of error is a function of 3n variables - 2n variables associated with the placements (α_i, β_i) of the phasors and n a priori probabilities $p(A_i)$. We would like to find the values of these 3n variables that minimize the probability of error for a given average or peak signal-to-noise ratio. This minimization becomes meaningless, however, since the probability of error is zero when all a priori probabilities except one are zero. Since only one symbol is being transmitted, the receiver can always detect this symbol with certainty. Thus, the probability of error criterion fails to take into account the information content of the transmitted symbols. One way around this difficulty is the use of channel capacity as a performance criterion to effect an optimization. Unfortunately, this leads to a great amount of mathematical complexity. An alternative approach to the optimization process is to consider only an equiprobable channel where all symbols have the a priori probability $p(A_i) = 1/n$. This conforms to the usual practice and greatly simplifies the mathematical analysis.

Another obstacle in the way of an analytical optimization is the evaluation of the integrals in Equation (11). Even for extremely simple decision regions these integrals may not be evaluated in closed form as functions of α_i and β_i. Therefore, we shall use asymptotic approximations for the probability density in the high and low signal-to-noise ratio regions.

Low Signal-to-Noise Ratios

What we wish to do now is to derive a simple, approximate expression for the average probability of error in terms of the phasor placements. Since all the decision regions are either closed convex regions (e.g., R_5 in Figure 3) or open regions (e.g., R_1 in Figure 3), our method of attack will be to start by finding the probabilities of certain "building block" types of regions - triangles with the transmission state A_i at one vertex and infinite "wedges" or angles with the transmission state A_i within the wedge. These building block regions are shown in Figure 4b and 4c. The probability of these regions may be found from use of the marginal phase density of the received carrier. This density was derived by Rice[2] and may be obtained by changing the density, Equation (8), into polar coordinates, r, ϕ, v, θ, and integrating over-all v. We obtain

$$p_{r,\phi}(\theta) = \frac{1}{2\pi} e^{-r^2/2\sigma^2} + \frac{r}{\sqrt{2\pi}} \cos(\theta - \phi)\, e^{-r^2/2\sigma^2} \sin^2(\theta - \phi)\, V\left[\frac{r}{\sigma}\cos(\theta - \phi)\right] \qquad (13)$$

where V(x) is the normal probability integral

$$V(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt \qquad (14)$$

For low signal-to-noise ratios an expansion of the Equation (13) has been given by Cahn[3].

$$p_{r,\phi}(\theta) = \frac{1}{2\pi}\left[1 + \sqrt{\frac{\pi}{2}}\frac{r}{\sigma}\cos(\theta - \phi) + \frac{r^2}{2\sigma^2}\cos 2(\theta - \phi) + \ldots\right] \qquad (15)$$

The marginal phase distribution may be obtained by integrating Equation (15) so that the probability of the received phasor falling in the angle $\gamma = \theta - \phi$ of Figure 4a is

$$p_r(\gamma) = \frac{1}{2\pi}\left[\gamma + \sqrt{\frac{\pi}{2}}\frac{r}{\sigma}\sin\gamma + \frac{r^2}{4\sigma^2}\sin 2\gamma + \ldots\right] \tag{16}$$

For our purposes we use only the first two terms of this expansion. Application of geometry shows that to this degree of approximation the probability of the triangle, Figure 4c, is zero. This same result obviously holds for all closed or bounded decision regions. For the open wedge in Figure 4b the probability becomes

$$p_{A_i}(\text{wedge}) \cong \frac{1}{2\pi}\left[\gamma_1 + \sqrt{\frac{\pi}{2}}\left(\frac{r_{12}}{4\sigma} + \frac{d_2}{2\sigma}\right)\right] \tag{17}$$

Using these results we may now examine the average probability of correct reception in the low signal-to-noise ratio region for the example system previously depicted. Notice that there is approximately zero probability of correct reception of the symbol A_5 since this region is bounded. The probabilities of the other regions R_1 to R_4 may be computed using Equation (17), since the difference between the shapes of these regions and the wedge is only a closed region with negligible probability. For example, the probability of region R_1 is

$$p_{A_i}(R_1) \cong \frac{1}{2\pi}\left[\gamma_1 + \sqrt{\frac{\pi}{2}}\left(\frac{r_{12}}{4\sigma} + \frac{r_{14}}{4\sigma}\right)\right] \tag{18}$$

It is not necessary to calculate the value of the angle γ_1 since the sum of these angles for all the decision regions must be 2π. The average probability of error for this example becomes

$$p_e = 1 - \frac{1}{5}\left[1 + \frac{r_{12} + r_{23} + r_{34} + r_{41}}{2\sigma\sqrt{2\pi}}\right] \tag{19}$$

This result is easily generalized to

$$p_e = 1 - \frac{1}{n}\left[1 + \frac{L}{2\sigma\sqrt{2\pi}}\right] \tag{20}$$

where L is the perimeter of the largest convex polygon consisting of the lines r_{ij}. Figure (5) shows this perimeter for the example system. Notice that the probability of error is independent of the position of A_5 in its present location.

High Signal-to-Noise Ratios

In the low signal-to-noise ratio region we found that the important factor in determining probabilities was the access of a state to an unbounded decision region. The high signal-to-noise ratio region is completely different in that now there is very little probability of an appreciable deviation of the transmitted phasor so that the short range factors become dominant.

The expansion of the marginal phase density for high signal-to-noise ratios is given by[3]

$$p_{r,\phi}(\theta) = \frac{r}{\sigma\sqrt{2\pi}}\cos(\theta - \phi)\, e^{\frac{-r^2\sin^2(\theta - \phi)}{2\sigma^2}} \tag{21}$$

Again referring to Figure 4a, the probability of the angle γ is

$$p_r(\gamma) \cong V\left(\frac{r}{\sigma}\sin\gamma\right) - \frac{1}{2} \quad \text{for} \quad \gamma \leq \frac{\pi}{2}$$

$$= \frac{1}{2} \quad \text{for} \quad \gamma > \frac{\pi}{2} \tag{22}$$

The triangle of Figure 4c is the basic unit to be used and its probability is found to be

$$p_{A_i}(\Delta) \cong V\left(\frac{d}{\sigma}\right) - 1 + \frac{\gamma_1}{2\pi} \tag{23}$$

if the angles facing the vertex A_i are acute and

$$p_{A_i}(\Delta) \cong \frac{\gamma_1}{2\pi} \tag{24}$$

otherwise.

For any given decision region the angles γ_i sum to 2π, while the altitudes d are equal to $r_{ij}/2$. The probability of region R_5 in Figure 3, for example, becomes

$$p_{A_5}(R_5) \cong V\left(\frac{r_{25}}{2\sigma}\right) + V\left(\frac{r_{15}}{2\sigma}\right) + \left(\frac{r_{35}}{2\sigma}\right) + V\left(\frac{r_{45}}{2\sigma}\right) - 3 \tag{25}$$

The probability of the region R_1 is

$$P_{A_i}(R_1) \cong Q\left(\frac{r_{12}}{2\sigma}\right) + Q\left(\frac{r_{14}}{2\sigma}\right) + Q\left(\frac{r_{15}}{2\sigma}\right) - 2 \qquad (26)$$

In general the probability of any decision region R becomes approximately

$$P_{A_i}(R_i) = 1 + \sum_{j\alpha\Omega_i} \left[Q\left(\frac{r_{ij}}{2\sigma}\right) - 1 \right] \qquad (27)$$

where the index "j" is summed over the set Ω_i which includes all the "immediate neighbors"[1] of state A_i. "Immediate neighbors" is taken to mean the set of states that influence the shape of the decision region R_i.* For example, in Figure 3 states A_2, A_5, and A_4 belong to Ω_1, since they affect the shape of region R_1. State A_3 obviously has no effect whatsoever on the probability of correct reception for the symbol A_1. The expression for the average probability of error becomes

$$P_e = 1 - \frac{1}{n}\sum_{i=1}^{n} \left\{ 1 + \sum_{j\alpha\Omega_i} \left[Q\left(\frac{r_{ij}}{2\sigma}\right) - 1 \right] \right\} \qquad (28)$$

which looks difficult but in reality is quite easy to apply. Figure 6 demonstrates its use in the example system.

Optimum Channel Structures

Regardless of the dimensionality of the signal space, the probability of error is a function of the relative distances between the vectors representing the symbols A_i and not of the absolute lengths of these vectors from the origin. Since these lengths represent power at sample points, an axis translation may be made to minimize the average or peak power without changing the probability of error of any symbol. Specifically this means we should arrange the α and β axes so that

$$\sum_{i=1}^{n} \alpha_i = \sum_{i=1}^{n} \beta_i = 0 \qquad (29)$$

in an optimum placement of the symbol phasors since this arrangement minimizes both peak and average power. In general the optimum placement for any given state is as far away from all other states as possible consistent with the peak or average power constraint. Since a peak power constraint is represented by a circle on the α, β plane, it may be seen that the optimum distribution of a small number of states, ($n \leq 4$), is obtained by pushing all states out to the peak power constraining circle. This optimum system is a phase modulated system.

Several other general results may be obtained by using the approximate expressions for probability of error derived in the preceding sections In the low signal-to-noise ratio region the probability of error, Equation (20), is a minimum when the perimeter L around the outside states is maximized. For peak power constraint this perimeter is maximized by a regular n-sided polygon inscribed in the peak power constraining circle. Thus, phase modulation minimizes the probability of error for low signal-to-noise ratios and peak power constraints.

For high signal-to-noise ratios optimum arrays of states for large values of n may be obtained by differentiating the probability of error, Equation (28), with respect to the 2n variables r_1, $r_2 \cdots$, r_n, ϕ_1, $\phi_2 \cdots \phi_n$ and considering the resulting set of equations to establish equilbrium positions for the symbol phasors. The results of this procedure for peak and average power constraints agree with optimum continuous ($n \to \infty$) signal distributions derived by Shannon[4] That is, for a peak power constraint a uniform array of states is optimum while for an average power constraint a Rayleigh (Gaussian on each axis) distribution of states is optimum. (The properties of continuous phase and amplitude modulated systems have been treated from a channel capacity standpoint by Blachman.[1] For digital systems with high alphabet sizes see Cahn[5] and Hancock and Lucky.[6]

Specific optimum systems for alphabet sizes $n \leq 16$ have been derived with the aid of the approximate expressions for probability of error. These results are summarized in a table of optimum channel structures, Figure 7.

System Performance

In this section we wish to compare the performance of AM, PM, and optimum AM-PM systems. This comparison will be made on the basis of probability of error and channel capacity for both peak and average signal-to-noise ratios.

---

*This may be stated explicitly using vector terminology. That is, A_j belongs to Ω_1 if for every integer $k \neq 1$,

$$(\overline{r_j} - \overline{r_i}) \cdot (\overline{r_k} - \overline{r_i}) \leq (\overline{r_k} - \overline{r_i}) \cdot (\overline{r_k} - \overline{r_i}) \ .$$

From the table in Figure 7 it can be seen that phase modulation is optimum for low signal-to-noise ratios and for alphabet sizes $n \leq 8$ in the high signal-to-noise ratio region; nowhere is amplitude modulation alone optimum. It is interesting to compare the AM system with PM for the low alphabet sizes. In the case of AM the probabilities of error are easily found using tables of the error function. For the transitional probabilities in a PM system a numerical integration of the marginal phase density, Equation 13, is needed. Cahn[3] has tabulated the integral of this function from which probabilities of error are easily calculated. Figures 8 and 9 show this probability of error together with the probability of error for AM systems. Asymptotes for these families of curves may be obtained by series expansion of the probability integrals. The following results are obtained:

$$p_e(\mathrm{AM}) \cong \frac{2e^{\frac{-r_m^2}{2\sigma^2(n-1)^2}}}{\sqrt{2\pi}\,\frac{r_m}{\sigma(n-1)}} \qquad (30)$$

$$p_e(\mathrm{PM}) \simeq \frac{2e^{\frac{-r_m^2}{2\sigma^2}\sin^2\frac{\pi}{n}}}{\sqrt{2\pi}\,\frac{r_m}{\sigma}\sin\frac{\pi}{n}} \qquad (31)$$

The channel capacities of these low alphabet size systems are shown in Figures 10 and 11. A digital computer was used to calculate the PM curves shown. While the curves shown are exact, two asymptotes may be given for each curve. Below the knee, the capacities are approximately given by the channel capacities of the continuous systems. The capacities of the continuous systems are well known to be[7]

$$C_{(\mathrm{AM})} = \frac{1}{2}\log\frac{2r_m^2}{\pi e\sigma^2} \text{ nits/symbol}^* \qquad (32)$$

$$C_{(\mathrm{PM})} = \frac{1}{2}\log\frac{2\pi}{e}\frac{r_m^2}{\sigma^2} \text{ nits/symbol}^* \qquad (33)$$

Above the knee the curves are obviously asymptotic to log n.

- -

*These capacities are for high signal-to-noise ratios.

For alphabet sizes $n \geq 8$ for average power and $n \geq 16$ for peak power constraints the optimum system differs from PM. The probabilities of error calculated for the optimum systems for $n = 8$ and $n = 16$ (designated A and B in Figure 7) are shown compared with corresponding PM systems in Figures 12 and 13. For higher alphabet sizes the advantage of optimum AM-PM over PM increases.

The final set of curves, Figures 14 and 15, show approximate channel capacities of all three systems for high signal-to-noise ratios. The lower asymptotes for the AM-PM systems are the continuous capacities.

$$C_{\mathrm{AM\text{-}PM}} \cong \log\frac{P_{\mathrm{peak}}}{2e\sigma^2} \text{ nits/symbol}^* \qquad (34)$$

(for peak power)

$$C_{\mathrm{AM\text{-}PM}} \cong \log\frac{P_{\mathrm{ave}}}{\sigma^2} \text{ nits/symbol}^* \qquad (35)$$

(for average power)

Summary

The objectives of this study have been twofold - first to obtain optimum N-ary AM-PM systems and second to compare these optimum systems with AM-only and PM-only systems. The derivation of optimum AM-PM systems involves a general two-dimensional decision process. Approximate expressions for the average probability of error for such a process have been derived for high and low signal-to-noise ratios. Using these expressions optimal channel structures may be derived. These optimal structures were summarized in Figure 7. In general PM-only is optimum for low signal-to-noise ratios and for small alphabet sizes. The AM-PM system becomes optimum only for $n \geq 8$ and average power constraints or for $n \geq 16$ and peak power constraints in the high signal-to-noise ratio region.

The three systems: AM-only, PM-only, and AM-PM, were compared in terms of channel capacity and average probability of error in a series of curves, Figures 8 to 15. From a channel capacity standpoint the AM-PM system is more efficient above a peak signal-to-noise ratio of 16 db or above an average signal-to-noise ratio of approximately 11 db.

- -

*These capacities are for high signal-to-noise ratios.

It is of course impossible to compare the utility of the three systems for a given set of operating conditions without considering the relative complexities, costs, weights, etc. All the instrumentation problems inherent in a coherent system are present in the AM-PM system to a greater degree than in AM-only or in PM-only. In any event it is hoped that the curves given here present a complete comparison of the theoretical performances of the three systems.

References

1. Blachman, N. P., "A Comparison of the Informational Capacities of Amplitude and Phase Modulated Systems," Proc. IRE, vol. 14, p. 748, 1953.

2. Kotelnikov, V. A., Theory of Optimum Noise Immunity, Government Power Engineering Press, Moscow-Leningrad, 1956; translated by R. A. Silverman, McGraw-Hill, New York, 1960.

3. Cahn, C. R., "Performance of Digital Phase Modulation Communication Systems", Ramo-Wooldridge Corp., Los Angeles, Calif., Tech. Report No. M110-9U5, April, 1959.

4. Shannon, C. E., "A Mathematical Theory of Communication", Bell System Technical Journal, vol. 27, pp. 379-423, 623-656, July-October, 1948.

5. Cahn, C. R., "Combined Digital Phase and Amplitude Modulation Communication Systems", IRE Transactions on Communication Systems, Vol. CS-8, pp. 150-154, September, 1960.

6. Hancock, J. C. and R. W. Lucky, "Performance of Combined Amplitude and Phase Modulated Communication Systems," IRE Transactions of Communication Systems, vol. CS-8, pp. 232-237, December 1960.

7. Jelonek, Z., "A Comparison of Transmission Systems," in W. Jackson (ed.), Communication Theory, p. 44, Academic Press, New York, 1953.

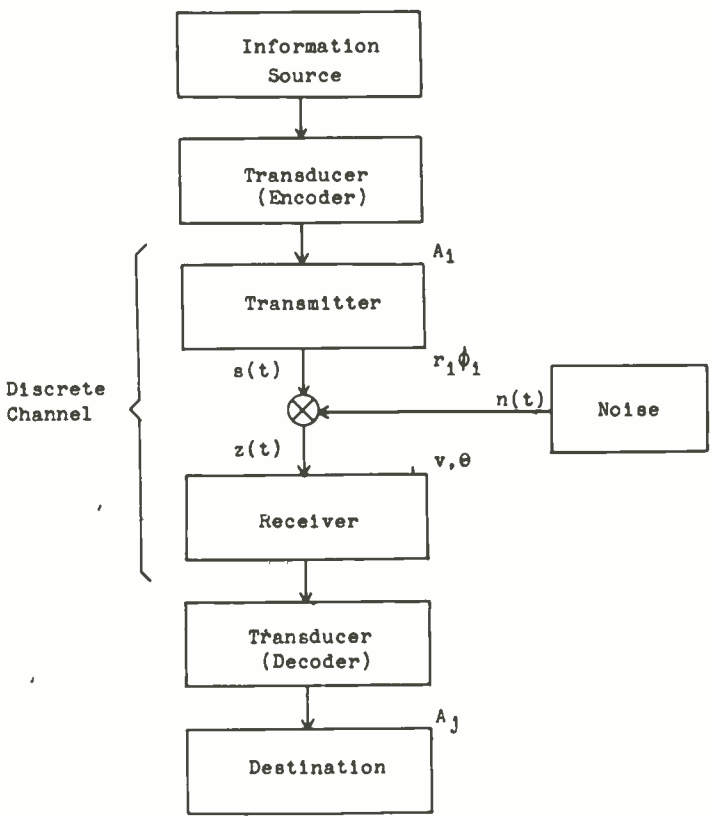


Fig. 1. A general communication system.

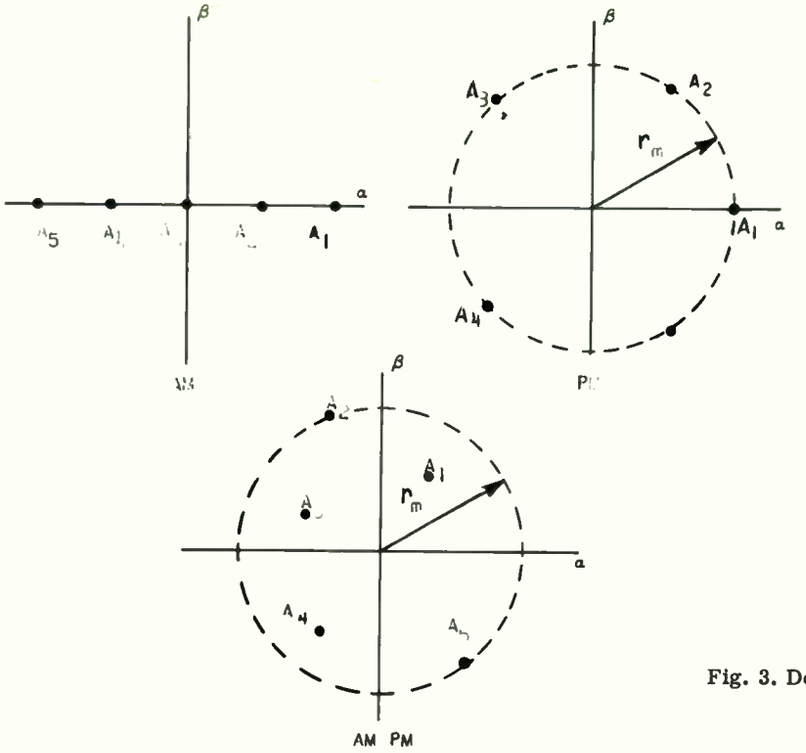


Fig. 2. The three systems of phasor placement.

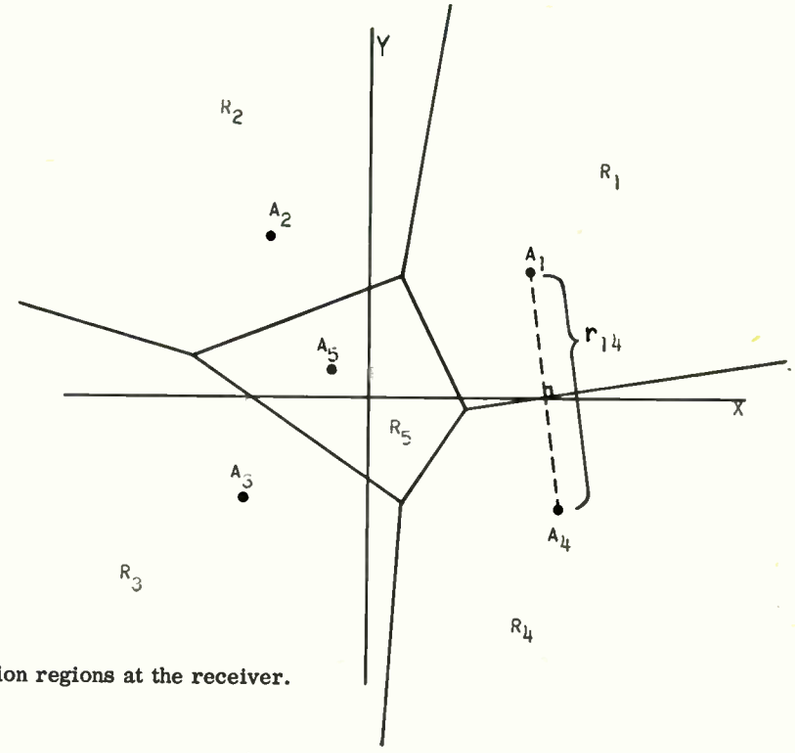


Fig. 3. Decision regions at the receiver.

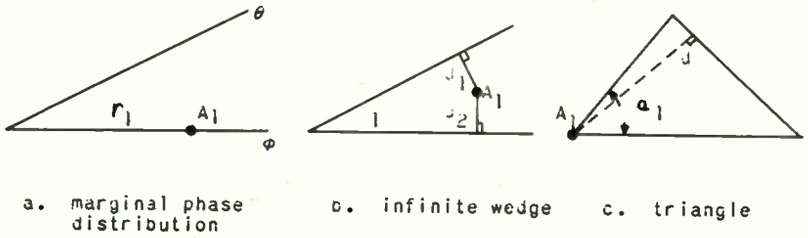


Fig. 4. "Building block" decision regions.

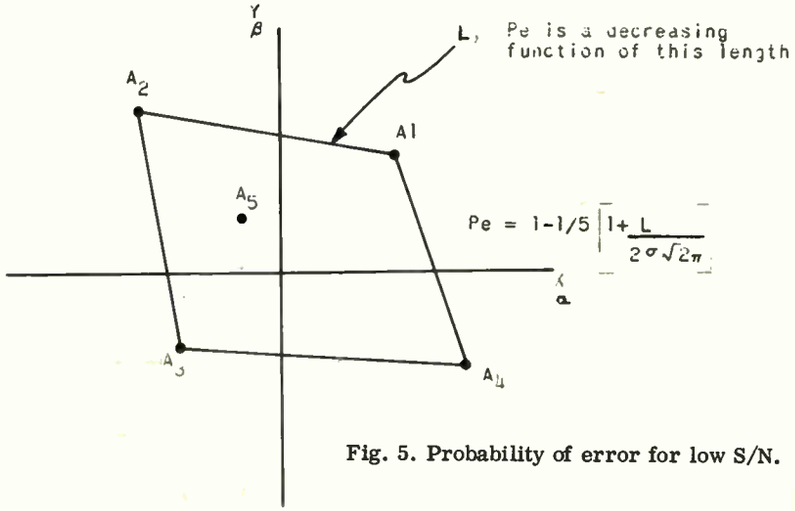
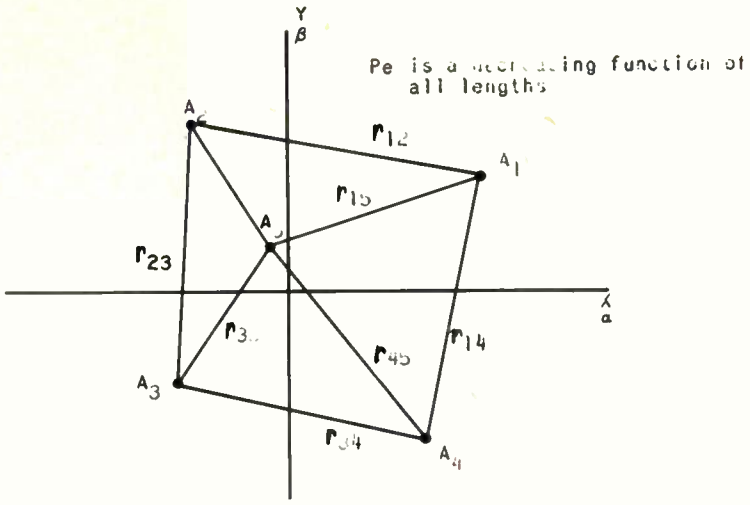


Fig. 5. Probability of error for low S/N.



$$Pe = 1 - 1/5 \left[2V\left(\frac{r_{23}}{2\sigma}\right) + 2V\left(\frac{r_{12}}{2\sigma}\right) + 2V\left(\frac{r_{14}}{2\sigma}\right) + \cdots \quad -11 \right]$$

Fig. 6. Probability of error for high S/N.

CHANNEL STRUCTURE

		n = 2	n = 4	n = 8	n = 16	n = ∞
Peak Power	High S/N	Phase Only	Phase Only	Phase Only	Phase + Amp.-B	Uniform
	Low S/N	Phase Only →				→

Average Power	High S/N	Phase Only	Phase Only	Phase + Amp.-A	Phase + Amp.-B	Rayleigh
	Low S/N	Phase Only →			→	Rayleigh

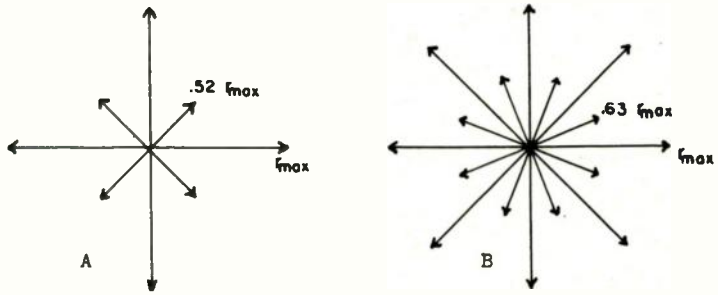
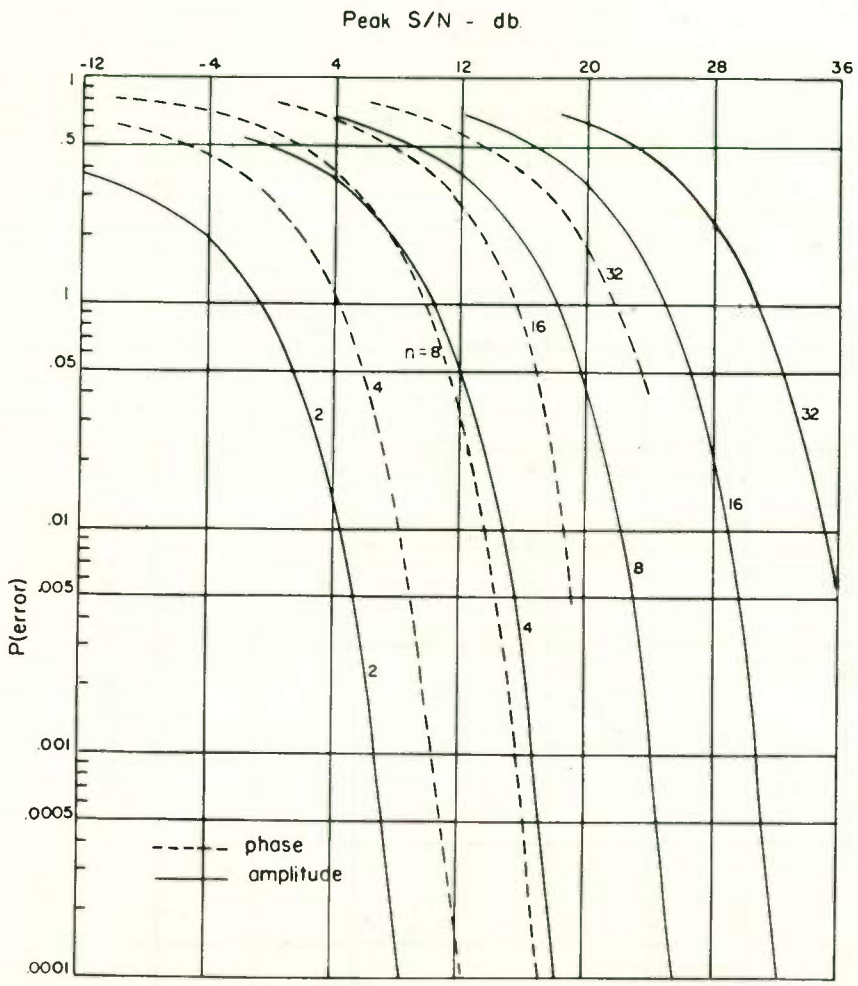
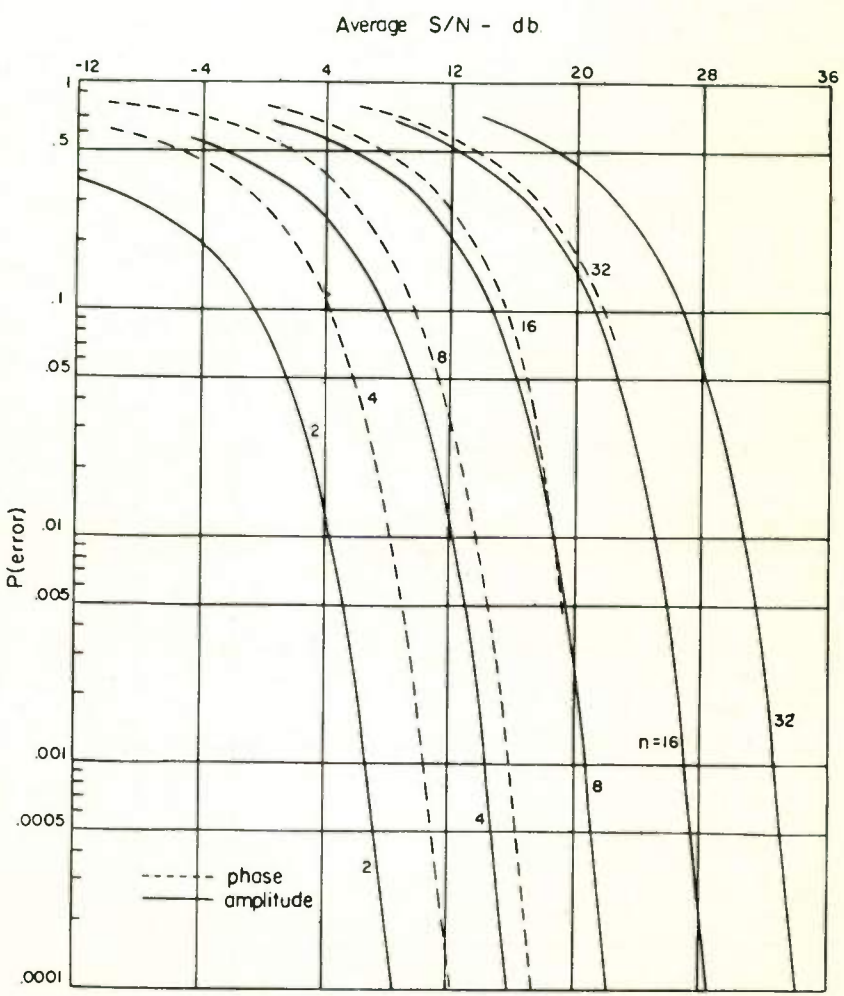


Fig. 7. Channel structure.



PROBABILITY OF ERROR VERSUS PEAK SIGNAL-TO-NOISE RATIO FOR AMPLITUDE MODULATED AND FOR PHASE MODULATED SYSTEMS

Fig. 8.



PROBABILITY OF ERROR VERSUS AVERAGE SIGNAL-TO-NOISE RATIO FOR AMPLITUDE MODULATED AND FOR PHASE MODULATED SYSTEMS

Fig. 9.

CHANNEL CAPACITY VERSUS PEAK SIGNAL-TO-NOISE RATIO
FOR AMPLITUDE MODULATED AND FOR PHASE MODULATED SYSTEMS

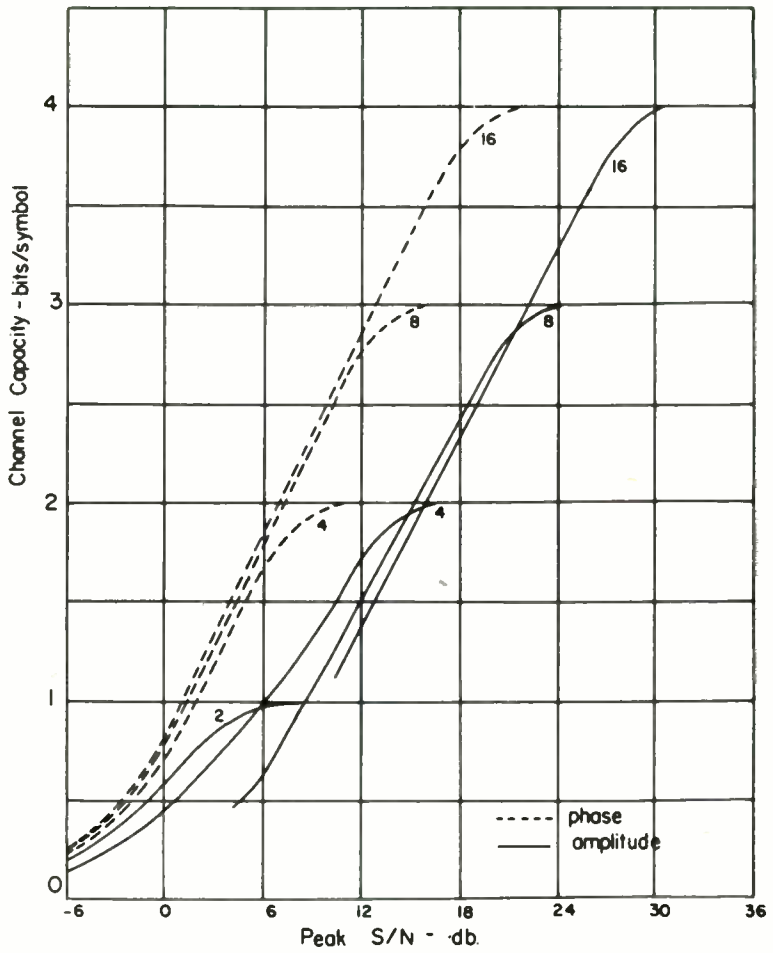


Fig. 10.

CHANNEL CAPACITY VERSUS AVERAGE SIGNAL-TO-NOISE RATIO
FOR AMPLITUDE MODULATED AND FOR PHASE MODULATED SYSTEMS

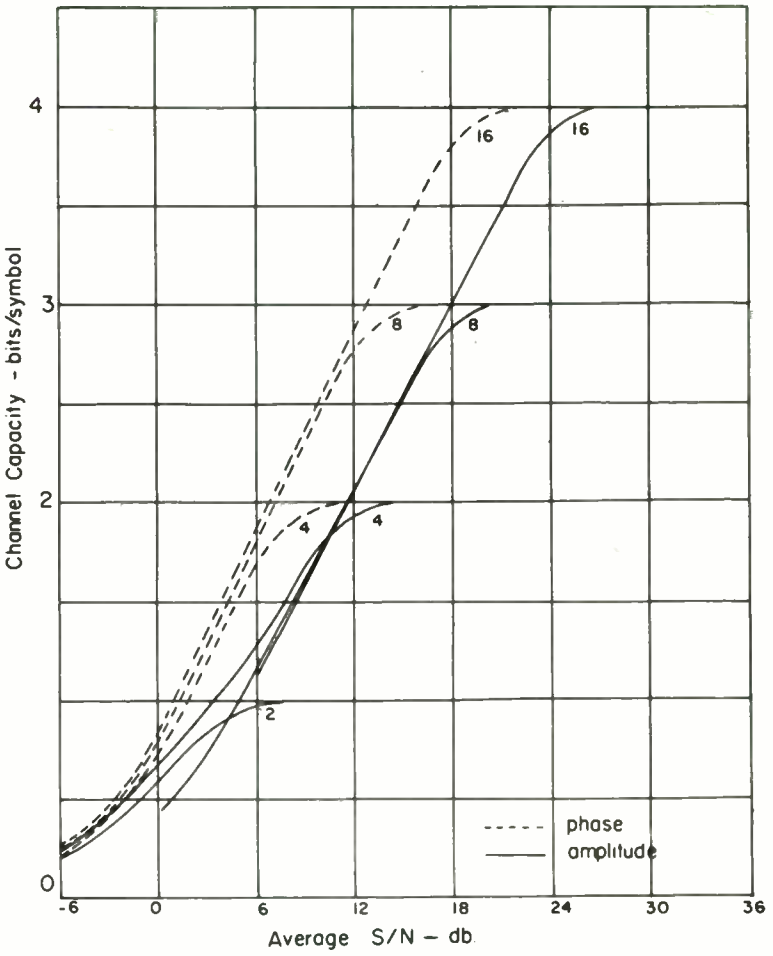
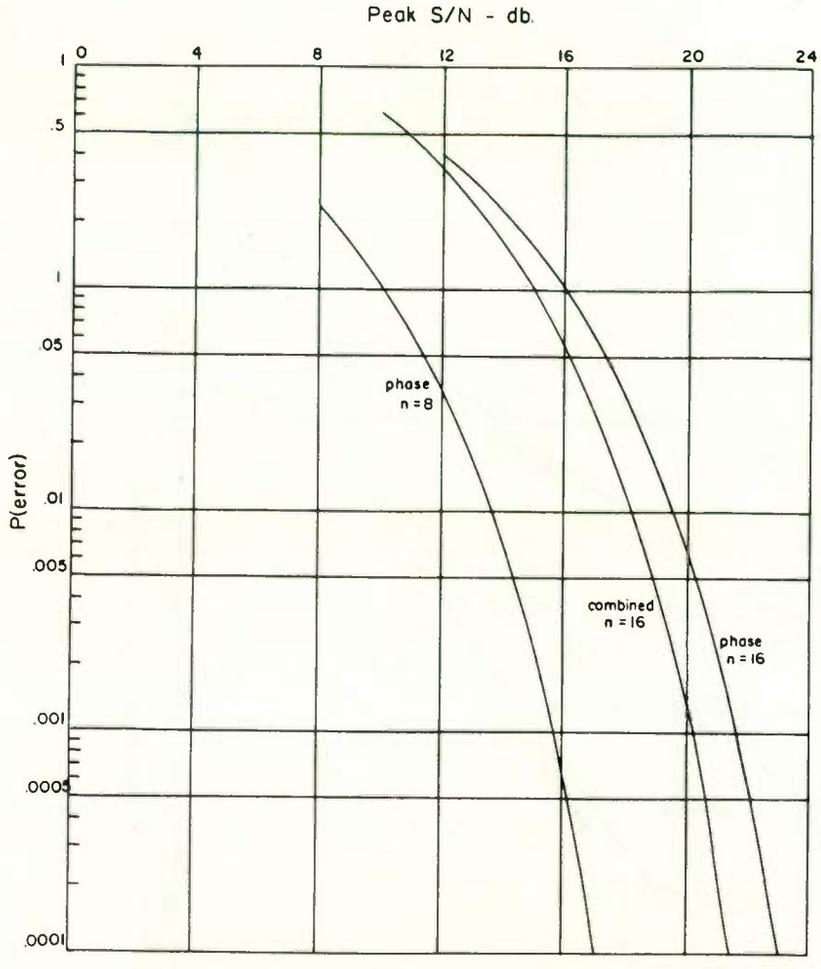
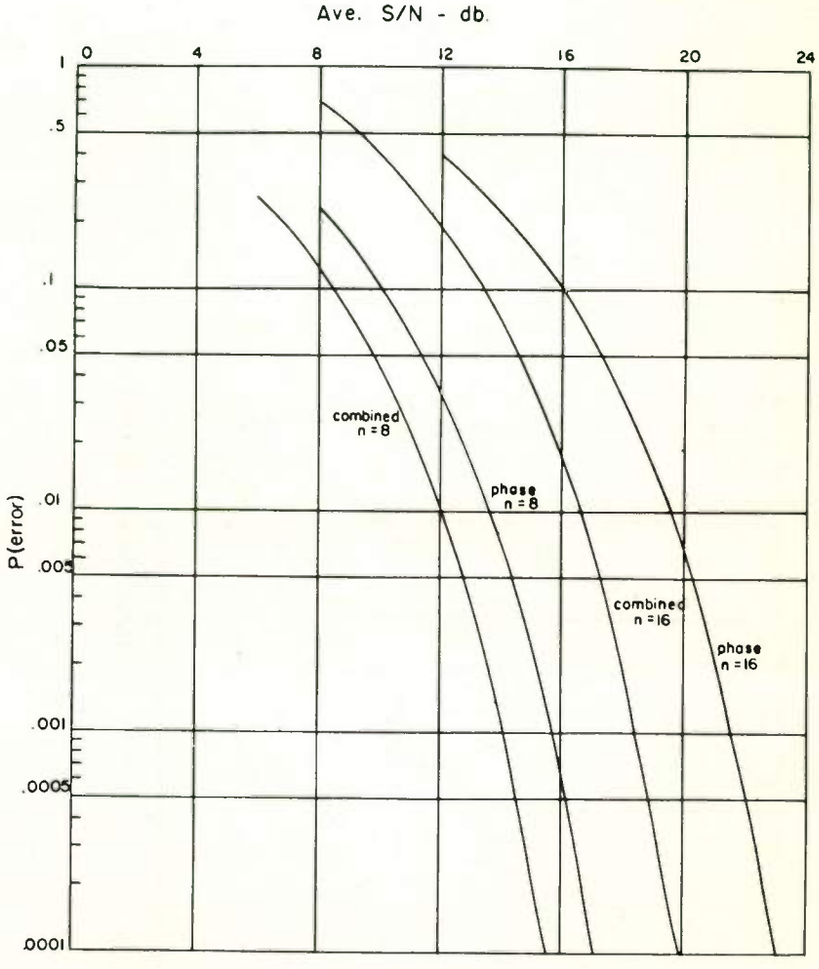


Fig. 11.



PROBABILITY OF ERROR VERSUS PEAK SIGNAL-TO-NOISE RATIO FOR N = 8 AND N = 16 CHANNELS

Fig. 12.



PROBABILITY OF ERROR VERSUS AVERAGE SIGNAL-TO-NOISE RATIO FOR N = 8 AND N = 16 CHANNELS

Fig. 13.

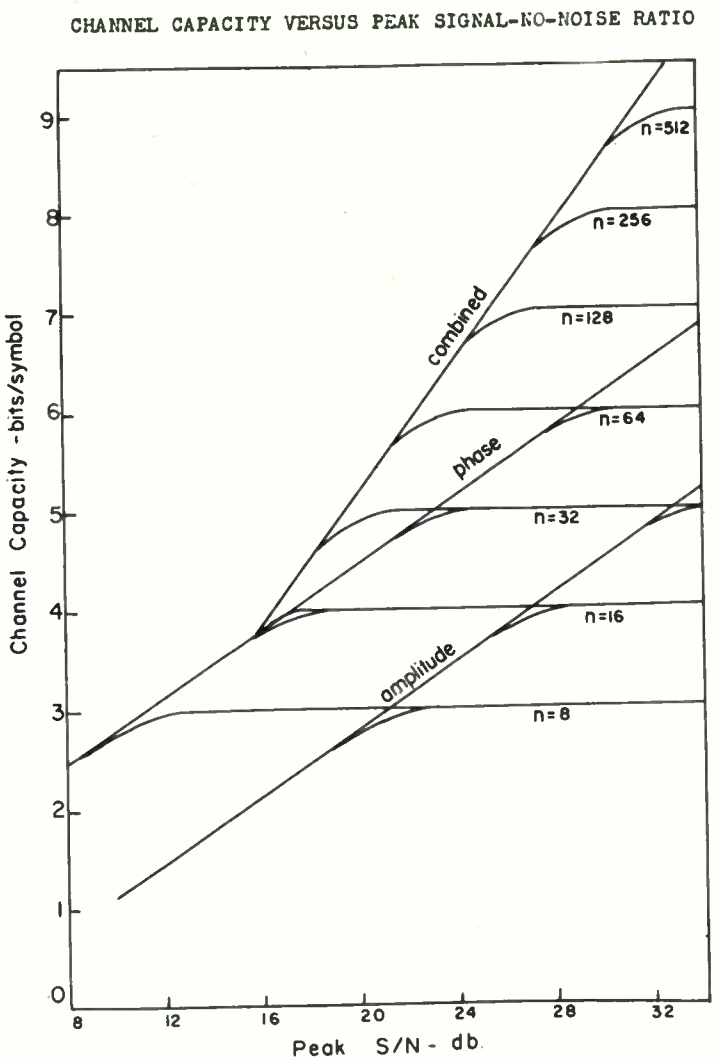

CHANNEL CAPACITY VERSUS PEAK SIGNAL-NO-NOISE RATIO
Channel Capacity - bits/symbol
Peak S/N - db
combined
phase
amplitude
n=512
n=256
n=128
n=64
n=32
n=16
n=8
0
1
2
3
4
5
6
7
8
9
8
12
16
20
24
28
32

Fig. 14.

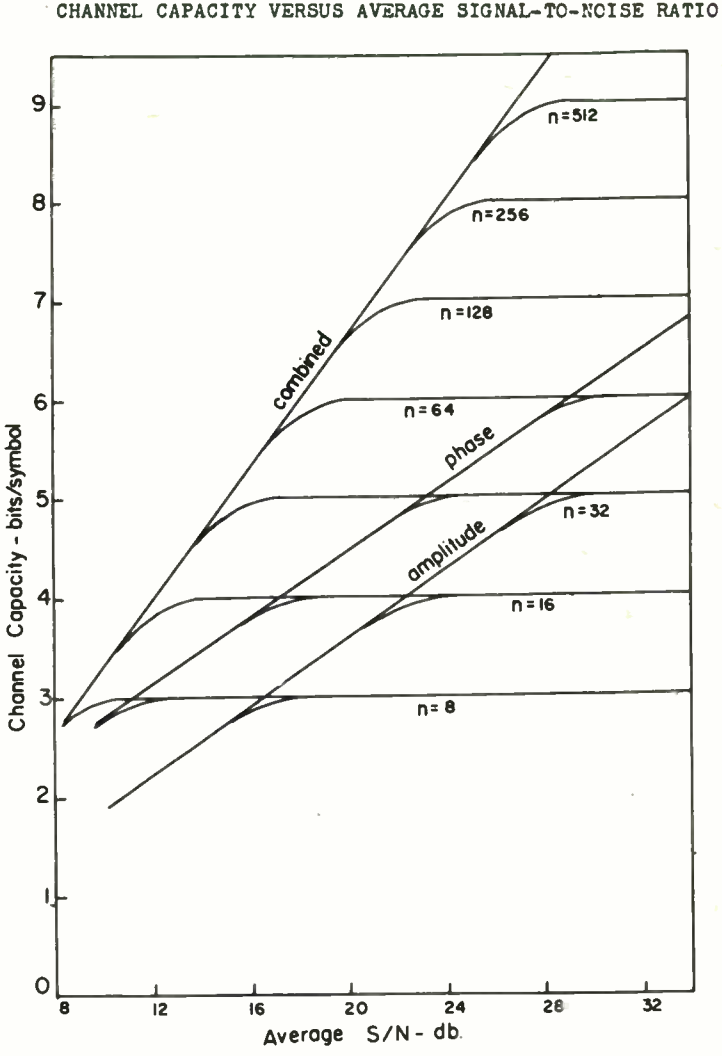

CHANNEL CAPACITY VERSUS AVERAGE SIGNAL-TO-NOISE RATIO
Channel Capacity - bits/symbol
Average S/N - db
combined
phase
amplitude
n=512
n=256
n=128
n=64
n=32
n=16
n=8
0
1
2
3
4
5
6
7
8
9
8
12
16
20
24
28
32

Fig. 15.

VIDEO BANDWIDTH COMPRESSION

William F. Haagen
Antenna Laboratory
Dept. of Electrical Engineering
The Ohio State University
Columbus, Ohio

Abstract

Several techniques are examined for encoding the information in a spatial field with the intention of minimizing redundancy. It was found that areas of the field which convey the most information could be divided into two classes—areas of low spatial frequencies or equal light intensity and areas of high spatial frequencies or equal statistical properties. An analysis is given of the processing required to extract non-redundant elements from the spatial field prior to transmission.

Theoretical verification using specific representative pictures together with results from optical spatial filtering indicate that recognizable images may be transmitted over a much compressed bandwidth. The compression factor is a function of the class of pictures to be transmitted.

TV BROADCAST FROM AN EARTH SATELLITE

RICHARD G. GOULD
STANFORD RESEARCH INSTITUTE
MENLO PARK, CALIFORNIA

Summary—This paper considers the need for and the feasibility of direct continental TV broadcast to conventional home receivers from an orbiting stationary satellite. Among the factors considered are: time and language differences of the potential audience, the lack of suitable programming material, and frequency allocation problems. Required transmitter and primary power have been calculated for several coverage situations on both a VHF and a UHF channel. These powers are significantly above previous estimates published elsewhere, and above the power capability of even the proposed SNAP 8, 60 kw, nuclear reactor except for coverage of limited areas on the ground.

I INTRODUCTION

In a previous article, Pierce[1] described several of what he termed "Spacious Fantasies" now prevalent concerning space travel and the uses of space. This paper treats another use of space which has been proposed recently in several quarters and which we would like to designate also as a *Spacious Fantasy*.

The proposals are for direct TV broadcast service to home receivers from an orbiting satellite. Most of them imply that no significant changes will be required in transmission standards, frequencies, or home receiving equipment. All are based on the use of a stationary (24-hour) satellite (of necessity at an altitude of 22,300 statute miles over the equator) so that receiving antennas can be fixed, obviating the need for tracking a moving satellite. Several proposals suggest that the transmitter will be powered by one of the proposed SNAP nuclear power sources.

This paper will consider several questions: What advantages does a satellite system offer over conventional TV broadcasts? What programming will be transmitted? What audience will receive it? What receiving equipment will be used? And what transmitting equipment and standards will be employed?

The transmitting and receiving equipment necessary to implement this system will be treated in Sec. V. We will consider first the need for this system, its advantages compared with alternative means, its audience, and the programming.

II AUDIENCE AND RECEIVING EQUIPMENT[2]

Assuming that coverage of North and South America is provided (the "continent" coverage of one proposal), what is the make up of the potential audience? Table I lists existing television sets in the Western Hemisphere, transmission standards, and language, by country. If the U.S. broadcast standard is to be used, existing and future viewers in Argentina and Venezuela (primarily) will be excluded, since the transmission standard in these countries is the Western European standard of 625 lines, 50 fields per second, rather than the U.S. standard of 525 lines 60 fields/sec. This leaves a total of about 55 million sets owned by English-speaking peoples in the United States, Canada, and the hemispheric possessions of Great Britain, plus 1.4 million sets owned by Spanish-speaking peoples, 1.1 million by Portugese, and 1 million (estimated) by those speaking French (see Table II).

Those speaking Portugese are distributed over about 3 time zones, French-speaking over about 2 time zones, Spanish-speaking over about 5 time zones, and English-speaking over 4 time zones. Just how the transmitting time of a satellite would be apportioned to serve an audience so divided by time and language is not apparent.

One solution would be to transmit programs in each language roughly in proportion to the number of viewers who speak that language. This implies 22.4 hours out of 24 in English, 35 minutes in Spanish, 27 minutes in Portugese and 24 minutes in French; a plan of dubious merit. Another solution would be to supply several audio channels and transmit in all four languages simultaneously. Although such a modification to the existing TV broadcast system might be accomplished, it requires a change in broadcast standards, as well as the purchase and installation of additional equipment. Such a modification might employ converters now being manufactured for FM multiplex systems, although this possibility is not considered in detail in this paper.

If direct transmission to Europe and Africa (or even to either area by itself) is considered, the problems of time and language difference, and transmission standards become even more acute (Table III), while differences in channel frequencies compound the difficulty even further (Table IV and Fig. 1).

III PROGRAMMING

The only programming advantage of a hemispheric system over a network of ground stations is that it allows the instantaneous transmission of programs over its wide area of coverage without the need for a network of microwave or cable circuits on the ground between stations. The United States transcontinental microwave network now makes possible such broadcasts within this country, but little use is made of this capability. According to information supplied by the Columbia Broadcasting System, 147 hours of TV were made available for instantaneous viewing throughout the country during 1961. Of this total, 131 hours were devoted to professional baseball games. The remainder, 16 hours, included newsworthy events and news commentary. It is reasonable to suppose that records of other networks would be similar in terms of the total number of hours, the portion devoted to sports and news, and even to the actual events and games covered. It is doubtful if foreign viewers would be interested in such a heavy dose of American sports events, and the figure of 16 hours gives, in our opinion, a good indication of the amount of programming that might justify instantaneous, live transmission abroad.

Consider what events that take place elsewhere in the world might be candidates for transmission to the United States on an instantaneous, live basis.

For example, it might be pointed out that the duty cycle on Olympic Games is 0.25 per year and that coronations of rulers of Great Britain from 1837 to the present have occurred at an average rate of 0.04 per year, there having been five coronations between 1837 and 1961.* With advances in geriatrics, it can be assumed that the coronation rate will fall even lower over the last half of this century, resulting in fewer such events which might be suitable for instantaneous transmission to the United States.

Other material has been suggested as being suitable for immediate, hemispheric TV viewing. One recent advertisement depicting a satellite TV broadcast system in operation shows natives watching a program on agricultural techniques on a communal TV set. Presumably, the program was filmed, so that the time from inception to production of the finished film was probably on the order of several weeks or even months. It is not clear why the instantaneous broadcast of such material is necessary. It would appear that furnishing prints of the film with sound tracks in appropriate languages would be a better solution. The same remarks apply to educational material of almost any nature: there is no demonstrated need for instantaneous transmission. Even the experimental airborne educational TV system over Indiana relies almost exclusively on recorded material.

Alternatively, the use of satellite TV broadcast might be justified on the basis that *regional* coverage essentially similar to that supplied at present, may be less expensive than a network of several ground stations. This means that one satellite may replace the network of ground stations necessary to serve just Argentina (625-line system, Spanish language) or just Brazil (525-line system, Portugese language) with one or several channels depending on the available transmitter power. Another such regional service might be Mexico and Central America (525-line system, Spanish language), still another might be just the United States and Canada (525 lines, English language). These possibilities are treated from a technical standpoint in Sec. V. The question of relative costs still must be determined and it is not certain that a single channel satellite system will prove less costly, per viewer, than a network of ground stations installed where required by population figures. During the next 20 years, perhaps, several ground stations might adequately serve the great majority of viewers in such a region as Brazil, or Argentina for example.

IV TECHNICAL CONSIDERATIONS

Regardless of the size of the area to be served, two cases must be distinguished: one or more satellites supplying the sole TV coverage in an area, or satellites supplementing conventional

* Although there were 6 rulers during that period, one (Edward VIII) reigned but was never crowned. If abdication speeches are included, then the United Kingdom has supplied 0.0484 events of the first magnitude per year.

ground station coverage. In the first case, one antenna installation will suffice at each receiving site. In the second case, not only must a receiving antenna of the requisite gain be supplied, but it must be capable of being rotated and elevated either towards the satellite, which will be at an elevation of from 40 degrees to 90 degrees above the horizon in the case of "continent" coverage, or towards existing ground stations which are essentially on the horizon.

In the case of "continent" coverage an additional problem exists. Although the satellite "sees" an area a little larger than one third of the earth's surface, the transmitting antenna pattern may be shaped to cover a smaller region to reduce the required transmitter power to a realizable amount, or to supply service to only a specific region.

Ideally, the pattern would be shaped to cover only the desired service area. In a proposal for "continent" coverage of North and South America,[3] it was planned to shape the transmitting antenna pattern to cover an area 3000 miles wide by 6000 miles long. While this area of 18 million square miles is close to the actual land area of the continents (16.1 million square miles), the continents of North and South America are neither roughly regular, nor are they symmetrical about the same, or any, longitude line. This implies that complex antenna pattern shaping would be required to reduce the coverage to just the land areas of the Western hemisphere. Without this exact pattern shaping, the transmitter power requirements increase above the amounts given in Sec. V. The requirements for exact and complicated patterns increase the stabilization requirements of the satellite and the complexity of the antenna structure.

Still another problem presents itself in the selection of operating frequencies for this system. If one channel is to be allocated in the present VHF or UHF TV bands, existing transmitters must be moved in frequency. In a typical case (Channel 13), this would require the displacement of 38 existing continental U.S. TV stations now operating, if reception of the satellite signal is desired in areas served by those stations. However, the satellite might not interfere with existing service on a channel in those same areas due to the low gain of conventional TV antennas at high elevation angles. This latter question has not been considered in detail in this paper.

Five Canadian TV stations use Channel 13, as do two in Brazil, one in Columbia, one in Peru, and one in Venezuela (Venezuela's Channel 10 has essentially the same frequency assignment as Channel 13). Essentially the same channel usage exists throughout the VHF TV band.

If Channel 83, at the high end of the UHF band is selected, no existing U.S. or foreign stations are affected. While it would be easier to assign a high UHF channel to this service, it must be emphasized that there are no UHF TV receivers outside the United States and that use of UHF within the U.S. is extremely limited. This is not to say that such a UHF service might not be offered and operated; however, it should be realized that we are then talking about a new TV service, not just the addition of one more channel to the existing system. Instead of several million sets in South America capable of receiving this channel, there are none. The higher transmitter power requirement for UHF broadcast is discussed in Sec. V.

V CALCULATIONS FOR REQUIRED TRANSMITTER POWER

To determine the required satellite transmitter power, the beacon equation will be applied to several coverage situations: the "continent"; the U.S. and Canada; the U.S. alone; Mexico and Central America; Argentina; and Brazil.

In general, P_t, the required transmitter power in db above 1 watt, is given by

$$P_t = G - A_t - A_r - K + S/N + F + N + B + P_r \qquad (1)$$

where G is free-space attenuation between isotropic antennas in db. The value for G is given by

$$G = 36.6 + 20 \log_{10} D + 20 \log_{10}(f_{mc}) \qquad (2)$$

where

D is the distance in miles

f_{mc} is the operating frequency in megacycles

A_t is the transmitting antenna gain in db above an isotropic antenna

A_r is receiving antenna gain in db above an isotropic antenna

K is the modulation improvement factor in db

S/N is the required signal-to-noise ratio in db

F is an allowance for fading in db

N is the noise figure of the receiver in db

B is an allowance for transmission line loss in db, and

P_r is the theoretical sensitivity of a receiver in db.

Several factors are the same for all coverage situations:

(1) D will be taken as 22,300 statute miles, although the actual distance to some receivers will be greater at the edge of the covered area.

(2) f_{mc} will be taken at the high end of the VHF band (Channel 13, 210 Mc) and at the high end of the UHF band (Channel 83, 884 Mc).

(3) A_r is determined by the receiving antenna to be required of the set owner. We assume a parabolic antenna 10 feet in diameter which results in a gain of 14 db at 210 Mc and 27 db at 884 Mc. It should be realized that a 10-foot-diameter antenna, or its equivalent, is a significantly more expensive antenna than is now in use in practically all home TV receiver installations.

(4) K is 0 db for television.

(5) S/N has been picked arbitrarily as 40 db. This choice will be justified in some detail since it appears that this is the most significant point of difference between this analysis and other estimates. Anderson,[3] using what appears to be a S/N ratio of 18 db, estimates that about 5 kw will be sufficient for coverage of North and South America on a UHF channel using present converters and conventional home receivers and a receiving antenna of 20 db gain. However, we feel that 40 db is the minimum value even in rural areas with small amounts of impulsive noise. In urban areas where coherent and impulsive noise are present, the requirement may be as high as 55 db.[4] Even a signal-to-noise ratio of 40 db is significantly below the value set by the FCC for both "Local Community" and "Grade A" service. ("Grade A" service is defined as the quality acceptable to a median observer which is available 90 percent of the time at the best 70 percent of receiver locations at the outer limits of the service area. Grade B service provides the same quality for the same time but at the best 50 percent of locations.) In terms of field strength, these grades of service are defined in Table VI.

These field strengths can be converted to signal-to-noise ratios once the gain of the receiving antenna is specified:

$$S/N = P_s/P_n \qquad (3)$$

where P_s is available signal power and P_n is available noise power. Noise power is

$$P_n = KTB \qquad (4)$$

where

K is Boltzmann's constant $= 1.374\ 10^{-23}$ in Joule/°K

T is temperature in degrees Kelvin

B is receiver bandwidth in cycles/sec.

Then, available signal power is

$$P_s = WA_e \qquad (5)$$

where W is the power density incident on the antenna in watts/m^2 and A_e is the maximum effective aperture of the antenna (effective area) in meters^2.

The power density is

$$W = E^2/Z_0 \qquad (6)$$

where E is the field intensity in volts/m, and Z_0 is the intrinsic impedance of the medium in ohms = 377Ω.

The maximum effective aperture is

$$A_e = \lambda^2 G/4\pi \qquad (7)$$

where λ is the wavelength in meters and G is the gain of the antenna.

Thus, finally,

$$S/N = \frac{E^2\lambda^2 G}{Z_0 4\pi KTB} \qquad (8)$$

The S/N ratio will be calculated for Channel 13, using the FCC standards for "Local Community" service, 7,080 μv/m. Assuming a simple dipole antenna with a gain of 1.5 (1.8 db) over an isotropic radiator,

$$S/N = \frac{(7080\ 10^{-6})^2 \left(\frac{300}{218}\right)^2 1.5}{(377)4\pi(1.374 \cdot 10^{-23})(300)(5 \cdot 10^6)}$$

$$= 1.455\ 10^6$$

$$= 61.4 \text{ db} \quad . \tag{9}$$

For Grade A service, 3550 μv/m, and assuming an antenna with the same gain:

$$S/N = \frac{(3550 \cdot 10^{-6}) \left(\frac{300}{218}\right)^2 (1.5)}{(377)4\pi(1.374 \cdot 10^{-23})(300)(5 \cdot 10^6)}$$

$$= 3.63\ 10^5$$

$$= 55.6 \text{ db} \quad . \tag{10}$$

It would appear that the FCC standards result in S/N ratios of at least 55 db. In "Grade B" areas, a higher-gain antenna would be expected and the S/N ratio would still be in the vicinity of 55 db.

The results of subjective tests of picture quality *vs.* signal-to-noise ratio are shown in Fig. 2. If a reasonable design criterion for a broadcast system is that at least 95 percent of the viewers must receive a "fine" picture, then these tests indicate that a S/N ratio on the order of 44 db is necessary. Even if the design standard is set so that 90 percent of the viewers receive a "passable" picture, the required S/N ratio is 34 db.

(6) F has been picked arbitrarily as 5 db. This is a low estimate of expected fading: Haviland assumes 10 db in a similar situation.[6]

(7) B has been picked arbitrarily as 0.5 db for Channel 13 and 1.3 db for Channel 83. These are averages based on measurements on new transmission lines,[7] a line length of 30 feet having been assumed. This value rises rapidly with the age of the line and with moisture on the line. Haviland assumed 3 db in a similar situation.[6] It should be remembered that, in distributions that are not highly skewed, "average" implies that about half of the installations will have greater loss than this value. This means that this number of sets will receive a worse than satisfactory picture.

(8) N has been picked arbitrarily as 9 db, an average for operating TV receivers at the high end of the VHF band, and 14 db at the upper end of the UHF band.[7] These figures likewise are averages, about half the sets having worse noise figures.

(9) P_r is −144 db below 1 watt/Mc. This means −136 db for 6 Mc bandwidth TV, and −137 for 5 Mc.

All the parameters are now determined with the exception of A_t, the transmitting antenna gain, which is a function of area on the ground to be covered. It will be computed separately for each coverage case.

Note that the transmitter antenna is gain-limited and that the receiving antenna is area-limited. Transmitter power would thus be independent of frequency, were it not for the different transmission-line loss and receiver noise figure at the two operating frequencies.

It should also be noted that calculations for required transmitter power do not include aural power. FCC regulations specify a minimum aural transmitter power of 50 percent of the peak visual transmitter power, but aural transmitters of significantly smaller power would be sufficient in these applications.

A. "CONTINENT COVERAGE"

This area of 3000 miles by 6000 miles subtends angles of about 7.4° by 13.2° at a distance of 22,300 miles. Thus

$$A_t = 10 \log \frac{30{,}000}{(13.2)(7.4)} \text{ db} = 24.6 \text{ db} \quad . \tag{11}$$

Case I (Channel 13)

$$P_t = 36.6 + 20 \log D + 20 \log f_{mc} - A_t - A_r$$

$$- K + S/N + F + N + B - P_r$$

$$P_t = 36.6 + 86.96 + 46.46 - 24.6 - 14 - 0$$

$$+ 40 + 5 + 9 + 0.5 - 137$$

$$= 48.92 \text{ dbw}$$

$$= 100 \text{ kw}^{*} \quad . \qquad (12)$$

Case II (*Channel 83*)

$$P_t = 36.6 + 20 \log D + 20 \log f_{mc} - A_c - A_r - K + S/N + F + N + B - P_r$$

$$= 36.6 + 86.96 + 58.94 - 24.6 - 27 - 0 + 40 + 5 + 14 + 1.3 - 137$$

$$= 54.2 \text{ dbw}$$

$$= 265 \text{ kw} \quad . \qquad (13)$$

B. U.S. AND CANADA

This area of 3000 miles subtends a "square" angle of about 7.4° at a distance of 22,300 miles. Thus

$$A_t = 10 \log \frac{30,000}{56.2} = 27.2 \text{ db} \qquad (14)$$

a difference of 2.6 db from "continent" coverage.

Case I (*Channel 13*)

$$P_t = 48.92 - 2.6 = 46.32 \text{ dbw}$$

$$= 44 \text{ kw} \quad . \qquad (15)$$

Case II (*Channel 83*)

$$P_t = 54.2 - 2.6 = 51.6 \text{ dbw}$$

$$= 145 \text{ kw} \quad . \qquad (16)$$

C. UNITED STATES

This area of 3000 miles by 1500 miles subtends angles of about 7.4° by 3.5° at a distance of 22,300 miles. Thus

$$A_t = 10 \log \frac{30,000}{25.9} = 30.65 \text{ db} \quad . \qquad (17)$$

a difference of 6 db from the "continent" coverage.

Case I (*Channel 13*)

$$P_t = 48.92 - 6 = 42.92 \text{ dbw}$$

$$= 20 \text{ kw} \quad . \qquad (18)$$

Case II (*Channel 83*)

$$P_t = 54.2 - 6 = 48.2 \text{ dbw}$$

$$= 66 \text{ kw} \quad . \qquad (19)$$

D. MEXICO AND CENTRAL AMERICA

This area of 3000 miles by 2000 miles subtends angles of about 7.5° by 5° at 22,300 miles. Thus

$$A_t = 10 \log \frac{30,000}{37.4} = 29 \text{ db} \qquad (20)$$

a difference of 4.4 db from "continent" coverage.

Case I (*Channel 13*)

$$P_t = 48.92 - 4.4 = 44.52 \text{ dbw}$$

$$= 28.5 \text{ kw} \quad . \qquad (21)$$

Case II (*Channel 83*)

$$P_t = 54.2 - 4.4 = 49.8 \text{ dbw}$$

$$= 96 \text{ kw} \quad . \qquad (22)$$

E. ARGENTINA

This 2000 mile by 1000 mile area subtends angles of about 5° by 2.5° at 22,300 miles. Thus

$$A_t = 10 \log \frac{30,000}{12.5} = 33.8 \text{ db} \qquad (23)$$

a difference of 9.2 db from "continent" coverage. The bandwidth of the Argentinian system is 6 Mc.

Case I (*Channel 13*)

$$P_t = 48.92 - 9.2 + 1 = 40.72 \text{ dbw}$$

$$= 11.8 \text{ kw} \quad . \qquad (24)$$

* For comparison, this is also the rated power of KOOK-TV, Channel 2, Billings, Montana, with studio and transmitter located 2 mi ESE of Billings atop Westergard Hill.

Case II (Channel 83)

$$P_t = 54.2 - 9.2 + 1 = 46.0 \text{ dbw}$$

$$= 40 \text{ kw} \qquad (25)$$

F. BRAZIL

This 1300 mile by 2000 mile area subtends angles of about 3.25° by 5° at 22,300 miles. Thus

$$A_t = 10 \log \frac{30{,}000}{16.3} = 32.6 \text{ db} \qquad (26)$$

a difference of 8 db from "continent" coverage.

Case I (Channel 13)

$$P_t = 48.92 - 8 = 40.92 \text{ dbw}$$

$$= 12.4 \text{ kw} \qquad (27)$$

Case II (Channel 83)

$$P_t = 54.2 - 8 = 46.2 \text{ dbw}$$

$$= 42 \text{ kw} \qquad (28)$$

To produce inputs to the transmitting antenna of these amounts, about 3 times this much primary power is required. These results are summarized below.

COVERAGE	CHANNEL	TRANSMITTER POWER (kw)	PRIMARY POWER* (kw)
"Continent"	13	100	300
	83	265	795
U.S. and Canada	13	44	132
	83	145	435
U.S.	13	20	60
	83	66	198
Mexico and Central America	13	28.5	85.6
	83	96	288
Argentina	13	11.8	35.5
	83	40	120
Brazil	13	12.4	37.2
	83	42	126

* 3 times transmitted power.

VI CONCLUSIONS

It seems clear that hemispheric TV broadcast from a satellite cannot be justified solely on the basis that it will provide instantaneous transmission of programs. Moreover, the problems of different transmission standards, language, and wide differences in local time make a hemispheric system, from the standpoint of programming, extremely inefficient and uneconomical. Even if the transmitting antenna pattern could be shaped to cover only the desired service area, the required primary video transmitter power would be 300 kw on a VHF channel and 795 kw on a UHF channel. Both are greater than even the 60 kw available from the proposed SNAP 8 nuclear power source, and do not include power requirements of the aural transmitter, the aural and video receiver and any control units.

Video power requirements for coverage of smaller areas are proportionately less, and might conceivably be supplied by one of several proposed SNAP units. Such service might be justified on the basis of lower cost than a network of ground TV transmitters connected by microwave or coaxial cable networks. However, the calculations for required transmitter power are extremely generous in all of the arbitrary assumptions: average rather than minimum receiver noise figures were used; line losses of new, dry, transmission line were used; a minimum signal-to-noise ratio was assumed, and so on. For coverage of the United States, the SNAP 8 power source might conceivably supply sufficient power for a video transmitter on a VHF channel, but many existing stations would have to be displaced to make one channel available. With a source of power of about 40 kw, video coverage of either Brazil or Argentina is possible on a VHF channel. An additional antenna, or an antenna rotator/elevator, must be furnished by the set owner if both the satellite and existing ground transmitters are to be received. On any of the UHF channels, the set owner must purchase, install, and operate a UHF converter.

If multiple audio channels are to be supplied, a modification to existing broadcast standards must

be made and equipment designed, manufactured, sold, and installed to implement this new service.

It should be emphasized that it is the *direct* hemispheric TV broadcast aspect which is considered fantastic. The use of communications satellites to *relay* television programs between and across continents is a desirable and achieveable goal. In this case, programs received at the satellite system terminal would be fed into the TV networks within each continent or area for immediate or delayed broadcast by local TV stations.

The space age appears to offer promise of providing new services and performing present tasks in novel ways. One criterion of this age should be that the new services must be necessary and the novel methods of providing an existing service should produce results that are at least as satisfactory as the method being replaced. Thus a TV broadcast system providing less than 40 db or so of signal-to-noise ratio, and based on averages rather than minimums (thus providing marginal service) is no real advance or improvement. Moreover, the installation of a system for which there is no need is an extravagant and unnecessary demonstration of technical competence.

REFERENCES

1. J. R. Pierce, "Spacious Fantasies," Trans. IRE PGMIL-4, 1, p. 3 (January 1960).
2. Two reports have just been brought to the author's attention. One by the Brookings Institution considers the questions discussed in Secs. II and III of this report in considerable detail and comes to many of the same conclusions. The report is entitled "Proposed Studies on the Implications of Peaceful Space Activities for Human Affairs," House of Representatives, 87th Cong. 1st Session, Report No. 242 (U.S. Government Printing Office, Washington, D.C., April 18, 1961). The second, "Satellite Broadcasting?" by S. G. Lutz of Hughes Research Laboratories also considers much of the same material and comes to many of the same conclusions.
3. Lt. Col. G. M. Anderson, USAF, "More Electrical Power from Space Vehicles," *Aircraft and Missiles*, pp. 23-27.
4. Fink, D. G., *Television Engineering Handbook*, 1st ed., p. 318 (McGraw Hill Book Co., Inc., New York, 1957).
5. R. P. Haviland, "The Communication Satellite," Astronautica Acta Vol. IV, 1958, pp. 70-89.
6. VHF and UHF TV Receiving Equipment, W. O. Swinyard, Proc. IRE, Vol. 48, No. 6, 1960, pp. 1060-1080.

Table I

TV BROADCAST STANDARDS
WESTERN HEMISPHERE

COUNTRY	NO. OF SETS	POPULATION (in millions)	LANGUAGE	SET TOTALS
U.S. STANDARD				
United States	52,600,000	180.0	Eng.	52,600,000
Canada	3,940,000	16.0	Eng. & Fr.	3,940,000
Bermuda	8,500	0.042	Eng.	
Bolivia		3.3	Sp.	
Brazil	1,100,000	53.0	Port.	
Chile	*	7.0	Sp.	
Columbia	150,000	13.0	Sp.	
Costa Rica	3,000	5.4	Sp.	
Cuba	500,000	5.4	Sp.	
Dominican Republic	15,000	2.7	Sp.	
Ecuador		3.9	Sp.	
El Salvador	30,000	2.4	Sp.	2,546,400
Guatemala	25,000	3.4	Sp.	
Haiti	1,200	3.4	Creole & Eng.	
Honduras	2,000	1.8	Sp.	
Mexico	660,000	31.4	Sp.	
Nicaragua	4,000	1.3	Sp.	
Panama	11,000	0.96	Sp.	
Peru	33,200	9.9	Sp. plus Quecha & Aymara	
Uruguay	24,000	2.7	Sp.	
WESTERN EUROPEAN STANDARD				
Argentina	450,000	19.9	Sp.	703,024
Trinidad	24	--	Sp.	
Venezuela	250,000	5.1	Sp.	
Netherlands Antilles (West Indies)	3,000	0.43	Various	

* Negligible.

SOURCES: Television Factbook, Fall-Winter Edition 1960, No. 31, pp. 261-276. Published by Television Digest, Triangle Publications, Inc., Radnor, Pennsylvania, and World Almanac, 1959.

Table II

NUMBER OF SETS IN WESTERN HEMISPHERE

LANGUAGE	U.S. STANDARD* 525 LINES 60 FIELDS/SEC	WESTERN EUROPEAN STANDARD* 625 LINES 50 FIELDS/SEC
English (U.S., Canada, Bermuda)	55,449,000	--
Spanish	1,438,000	700,000
Portugese (Brazil)	1,100,000	--
French (Canada)	1,100,000 (estimated)	--

* See Table V for description of TV broadcast standards.

SOURCE: Television Factbook, loc. cit.

Table III

TV BROADCAST STANDARDS

COUNTRY*	POPULATION (in millions)	SETS	LANGUAGE	STANDARD§ W.E. = Western European E.E. = Eastern European
EUROPE				
Austria	7.0	150,000	German	W.E.
Belgium	9.0	400,000	French Flemish	French in French districts W.E. in Flemish districts
Bulgaria	7.7	3,000	Bulgarian	E.E.
Cyprus	0.5	1,000	Greek Turkish	W.E.
Czechoslovakia	13.3	600,000	Czech, Moravian, Slovak	E.E.
Denmark	4.5	400,000	Danish	W.E.
Finland	4.3	60,000	Finnish	W.E.
France	4.4	1,500,000	French	French
Germany (East)	17.8	503,000	German	W.E.
Germany (West)	53.3	4,250,000	German	W.E.
Greece	8.0	--†	Greek	E.E.
Hungary	9.8	75,000	Hungarian	E.E.
Ireland	2.8	70,000	English Irish	British
Italy	48.3	1,800,000	Italian	W.E.
Luxembourg	0.3	6,000	French	French
Malta	0.3	8,500	Italian	W.E. (viewers watch Italian stations)
Monaco	0.02	11,000	French	French
Netherlands	11.0	700,000	Dutch	W.E.
Norway	3.5	17,500	Norwegian	W.E.
Poland	28.5	300,000	Polish	E.E.
Portugal	8.9	40,000	Portugese	W.E.
Rumania	17.8	22,000	Rumanian	E.E.
Spain	29.4	240,000	Spanish	W.E.
Sweden	7.4	810,000	Swedish	W.E.
Switzerland	5.1	105,000	German French Italian	W.E.
Turkey	24.8	1,000	Turkish	W.E.
U.K.	51.0	10,900,000	English	British
USSR	200.0	4,000,000	Russian, etc., etc., etc., etc.	E.E.
Yugoslavia	18.2	15,000	Serbo-Croatian	W.E.

Table III *Concluded*

COUNTRY*	POPULATION (in millions)	SETS	LANGUAGE	STANDARD$ W.E. = Western European E.E. = Eastern European
AFRICA				
Algeria	9.5	47,700	French & Arabic	French (simultaneous French and Arabic sound channels)
Ghana	4.7	--		W.E.
Morocco	9.8	5,000	French, Moorish Arabic, and Berber Dialects	French
Nigeria	31.8	6,000		W.E.
Rhodesia		--		W.E.
Tunisia	3.8		Arabic	French
ASIA				
China	621.0	Unknown	Chinese plus dialects, etc., etc., etc.	Unknown
India	393.0	300	Hindi, Urdu English, Tamil, etc., etc., etc.	W.E.
Iran	18.9	36,000	Persian	U.S.
Iraq	6.5	70,000	Arabic plus other	W.E.
Japan	90.0	5,000,000	Japanese	U.S.
Korea	22.0	7,000	Korean	U.S.
Kuwait		600	Arabic	U.S.
Lebanon	1.5	12,000	Arabic	W.E.
Philippines	23.0	30,000	Tagalog English Spanish	U.S.
Taiwan	9.5	100	Chinese	U.S.
Thailand	21.0	50,000	Thai	U.S.
UAR	28.0	30,000	Arabic	W.E.

* Countries not listed here have none or few operating TV stations.

† -- indicates a few TV sets.

$ See Table V for description of TV broadcast standards.

SOURCE: Television Factbook, op. cit.

Table IV

FOREIGN CHANNEL FREQUENCIES

WESTERN EUROPEAN STANDARDS (except Italy)			FRENCH STANDARDS			ITALIAN CHANNELS		
Channel	Vision Carrier (mc)	Sound Carrier (mc)	Channel	Vision Carrier (mc)	Sound Carrier (mc)	Channel	Vision Carrier (mc)	Sound Carrier (mc)
E-1	41.25	46.75	F-1	46.00	42.00	A	52.75	59.25
E-1A	42.25	46.75	F-2	52.40	41.25	B	62.25	67.75
E-2	48.25	53.75	F-3	56.15	67.75	C	82.25	87.75
E-2A	49.75	55.75	F-4	65.55	54.40	D	175.25	180.75
E-3	55.25	60.75	F-5	164.00	175.15	E	183.75	189.25
E-4	62.25	67.75	F-6	173.40	162.25	F	192.25	197.75
E-4A	82.25	87.75	F-7	177.15	188.30	G	201.25	206.75
E-5	175.25	180.75	F-8A	185.25	174.10	H	210.25	215.75
E-6	182.25	187.25	F-8	186.55	175.40			
E-7	189.25	194.75	F-9	190.30	201.45			
E-7A	192.25	197.75	F-10	199.70	188.55			
E-8	196.25	201.75	F-11	203.45	214.60			
E-8A	201.25	206.75	F-12	212.85	201.70			
E-9	203.25	208.75						
E-10	210.25	215.75						
E-11	217.25	222.75						
EASTERN EUROPEAN STANDARDS			BRITISH STANDARDS			JAPANESE CHANNELS		
R-1	49.75	56.25	B-1	45.00	41.50	J-1	91.25	95.75
R-2	59.25	65.75	B-2	51.75	48.25	J-2	97.25	101.75
R-3	77.25	83.75	B-3	56.75	53.25	J-3	103.25	107.75
R-4	85.25	91.75	B-4	61.75	58.25	J-4	171.25	175.75
R-5	93.25	99.75	B-5	66.75	63.25	J-5	177.25	181.75
R-6	175.25	181.75	B-6	179.75	176.25	J-6	183.25	187.75
R-7	183.25	189.75	B-7	184.75	181.25	J-7	189.25	193.75
R-8	191.25	197.75	B-8	189.75	186.25	J-8	193.25	197.75
R-9	199.25	205.75	B-9	194.75	191.25	J-9	199.25	203.75
R-10	207.25	213.75	B-10	199.75	196.25	J-10	205.25	209.75
R-11	215.25	221.75	B-11	204.75	201.25	J-11	211.25	215.75
R-12	223.25	229.75	B-12	209.75	206.25			
			B-13	214.75	211.25			
			U.S. (WORLD WIDE) STANDARD (except Japan)					
			2	54	58.5			
			3	60	64.5			
			4	60	70.5			
			5	70	80.5			
			6	82	86.5			
			7	174	178.5			
			8	180	184.5			
			9	180	190.5			
			10	192	196.5			
			11	198	202.5			
			12	204	208.5			
			13	210	214.5			

SOURCE: Television Factbook, op cit., p. 269.

Table V

PRINCIPAL TELEVISION SYSTEMS OF THE WORLD

	BRITISH	U.S.	WESTERN EUROPE	EASTERN EUROPE	FRENCH
Location of principal use	United Kingdom	Western Hemisphere	Western Europe	Soviet orbit	France and possessions
Number of lines per picture	405	525	625	625	819
Video bandwidth (mc)	3	4	5	6	10.4
Channel width (mc)	5	6	7	8	14
Sound carrier relative to vision carrier (mc)	-3.5	+4.5	+5.5	+6.5	11.15*
Sound carrier relative to edge of channel (mc)	+0.25	-0.25	-0.25	-0.25	0.10*
Interlace	2/1	2/1	2/1	2/1	2/1
Line frequency (cycles per sec)	10,125	15,750	15,625	15,625	20,475
Field frequency (cycles per sec)	50	60	50	50	50
Picture frequency (cycles per sec)	25	30	25	25	25
Vision modulation	+	-	-	-	+
Level of black as percentage of peak carrier	30	75	75	75	25
Sound modulation	AM	FM	FM	FM	AM

* French standards invert the video and audio frequencies for certain channels.

SOURCE: Television Factbook, op cit., p. 268.

Table VI

FCC GRADES OF SERVICE

GRADE OF SERVICE	CHANNELS 2-6		CHANNELS 7-13		CHANNELS 14-83	
	dbu*	μ volts/m	dbu*	μ volts/m	dbu*	μ volts/m
Local community minimum	74	5010	77	7080	80	10,000
Grade A	68	2510	71	3550	74	5,010
Grade B	47	224	56	631	64	1,580

* dbu = db above 1 microvolt per meter.

SOURCE: Fink, op cit., p. 3-18.

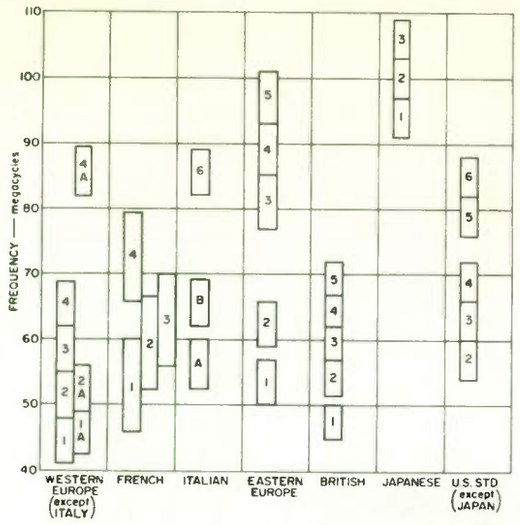


Fig. 1 (a). VHF television channel frequencies.

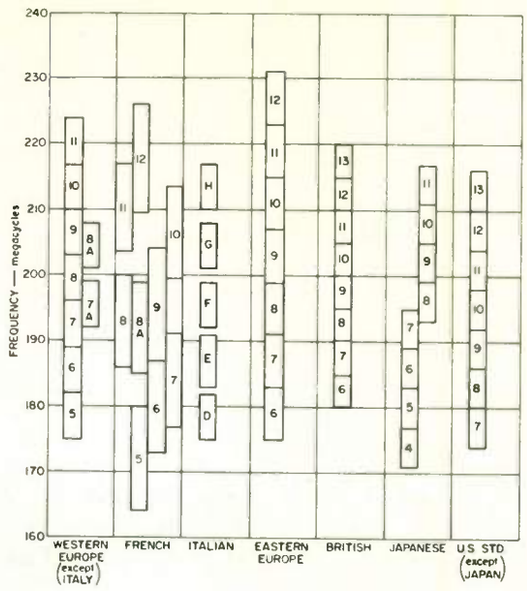


Fig. 1 (b). UHF television channel frequencies.

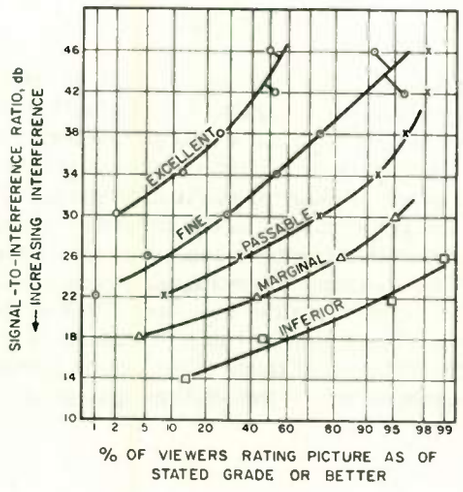


Fig. 2. Picture quality vs SNR. (From: C. E. Dean, "Measurement of the Subjective Effects of Interference in Television Reception," Proc. IRE, vol. 48, pp. 1035-1049; June, 1960.)

THE EFFECT OF FREQUENCY-COMPRESSIVE FEEDBACK UPON ADDITIVE DISTURBANCES IN WIDEBAND-FM SYSTEMS

Elie J. Baghdady
ADCOM, Inc.
Cambridge, Mass.
M. I. T.
Cambridge, Mass.

Abstract

A theoretical analysis is presented of the effect of frequency-compressive feedback upon various types of additive disturbances in a wideband-FM receiver. Of particular relevance to communications via earth satellites are the various aspects of the performance in the presence of random-fluctuation noise. It is shown, for example, that the performance in the presence of noise of frequency-compressive feedback loop depends primarily upon the convolution of the envelope of the impulse response of the IF filter and the impulse response of the low-pass filter within the loop. This dependence allows considerable exchanges in the frequency-response characteristics of these filters to be introduced without affecting the performance of the system against noise.

THE RESPONSE OF AN AUTOMATIC PHASE CONTROL SYSTEM TO FM SIGNALS AND NOISE

Donald L. Schilling
Electronics Research Laboratories
School of Engineering
Columbia University
New York 27, N. Y.

and

Mischa Schwartz
Department of Electrical Engineering
Polytechnic Institute of Brooklyn
Brooklyn 1, N. Y.

Summary

An Automatic Phase Control (APC) System is analyzed in order to determine its response to frequency modulated signals, and to colored Gaussian noise. Emphasis is placed on the system's response to signals having "ramp" type FM.

The APC System is the basic element in many communication and radar systems. The resulting second order nonlinear differential equation also represents the motion of a pendulum when acted on by an applied force and a nonlinear friction force. The device can be used to synchronize two oscillators, to track an accelerating satellite or other target, or as an FM receiver.

In contradistinction to the analyses performed by previous investigators, the assumptions of linearizing the system, and of assuming a signal-to-noise (S/N) ratio greater than unity, will not be made.[1]

I. Theoretical Results

The APC System analyzed in this paper is shown in Fig. 1-a. The signal voltage,

$$e_c(t) = S \sin \phi_1(t)$$

$$= S \sin\left(\omega_1 t + \alpha \int_0^t e_m(\lambda)d\lambda\right), \quad (1a)$$

and the output voltage of the voltage controlled oscillator (VCO)

$$e_v(t) = \cos \phi_2(t)$$

$$= \cos\left(\omega_2 t + \int_0^t G_2 e_o(\lambda)d\lambda\right) \quad (1b)$$

are multiplied in the phase detector. The resulting difference frequency signal,

$$e_d(t) = SG_1 \sin(\phi_2(t) - \phi_1(t)) \quad (2a)$$

$$= SG_1 \sin\left(\Omega t + \int_0^t (G_2 e_o(\lambda) - \alpha e_m(\lambda))d\lambda\right)$$

represents the input to the filter. In this equation

$$\Omega = \omega_2 - \omega_1$$

G_1 represents the gain of the phase detector, and

G_2 represents the sensitivity of the voltage controlled oscillator (rad/sec/volt).

The voltage output of the filter

$$e_o(t) = \int_0^t h(t - \lambda)e_d(\lambda)d\lambda \quad (2b)$$

corrects the frequency of the VCO. Thus the frequency of the VCO attempts to follow the instantaneous frequency of the input signal.

A. The Response of an APC System to an FM Signal

The type of filter used greatly influences the performance of an APC System. It limits the frequency range of operation of the system and restricts the type of input signals that it can synchronize to.

To enable the system to synchronize to a signal whose frequency varies linearly with time, when the signal is embedded in noise, an active phase lag filter, as shown in Fig. 1-b, is employed. Using this filter in the system shown in Fig. 1-a the second order non-linear differential equation describing the performance of the system becomes

$$\frac{d^2\phi}{dt^2} + \omega_n \zeta \cos\phi \frac{d\phi}{dt} + \omega_n \omega_a \sin\phi = -\alpha \frac{de_m(t)}{dt} \quad (3)$$

where $\phi = \phi_2 - \phi_1$, $\omega_n = G_1 G_2 S$ rad/sec, and $\alpha e_m(t)$ represents the frequency modulation of the input signal. It should be noted that the loop gain (or sensitivity) of the system, ω_n, can be considered equal to unity, and its actual magnitude considered as part of the magnitude of the parameters of the phase lag filter.

Equation (3) can be rewritten in normalized form

$$\frac{d^2\phi}{d\tau^2} + \epsilon \cos\phi \frac{d\phi}{d\tau} + \sin\phi = -a \frac{de_m}{d\tau} \quad (4)$$

where $\tau = \sqrt{\omega_a \omega_n}\, t$

$\epsilon = \sqrt{\omega_n/\omega_a}\, \zeta$

and $a = \alpha/\omega_n \omega_a$.

If $\epsilon = a = 0$, this equation becomes the well-known pendulum equation. If 'a' is not zero, the equation represents the motion of a pendulum with an applied force. The term, $\epsilon \cos\phi\, d\phi/d\tau$ represents a nonlinear friction force, which damps the motion of the pendulum as long as $|\phi| < \pi/2$. When $|\phi| > \pi/2$, the coefficient of friction, $\epsilon \cos\phi$, becomes negative. If one assumes that the phase error, ϕ, is always much less than 1 radian, Eq. (4) can be linearized and a linear analysis performed.

It is of interest to approximately determine the time required for the system to complete a cycle of its motion, given a set of initial conditions. The time, τ, can be written as,

$$\tau_{ab} = \int_a^b \frac{d\phi}{\phi'} \quad (5)$$

where $\phi' = \frac{d\phi}{d\tau}$.

As previously noted, if $a = \epsilon = 0$, Eq. (4) reduces to the pendulum equation. A first integral can be obtained, and ϕ' found. If when $\phi = 0$, $\phi'(o) = \Omega_n$, ϕ' becomes

$$\phi' = \sqrt{\Omega_n^2 - 4 \sin^2 \frac{\phi}{2}} \quad (6)$$

and

$$\tau_{ab} = \int_a^b \frac{d\left(\frac{\phi}{2}\right)}{\sqrt{\left(\frac{\Omega_n}{2}\right)^2 - \sin^2 \frac{\phi}{2}}} \quad . \quad (7a)$$

The time, τ, that it takes a pendulum to move 1/4 of a cycle is[2]

$$\tau = K(k^2) \quad (7b)$$

where $K(k^2)$ is the complete elliptic integral of the first kind, and when $\Omega_n < 2$,

$$k^2 = \left(\frac{\Omega_n}{2}\right)^2 = \sin^2\left(\frac{\phi_{max}}{2}\right) . \quad (7c)$$

If $\phi_{max} \leq 128^o$, $|\Omega_n| \leq 1.8$. The time, T, (in seconds) required by the pendulum to complete a full period is bounded by,

$$2\pi \leq (\omega_a \omega_n)^{\frac{1}{2}} T \leq 9 \quad . \quad (7d)$$

When $\Omega_n > 2$, the motion of the pendulum is no longer bounded. The pendulum now continually rotates in one direction. If some damping were present, the path of the pendulum would be perturbed, and the pendulum would rotate until, when $\phi = 2n\pi$, the velocity of the pendulum, Ω_n, became less than 2.

The concept of perturbing the motion of the pendulum by a friction force, and an applied force, will be discussed in detail in the following section.

1. A Perturbation Solution[3]

The equation describing the APC System can be rewritten in the form

$$\phi'' + \sin\phi = -\epsilon \cos\phi\, \phi' - ae'_m \quad . \quad (8)$$

If ϵ and 'a' are each less than unity, a solution using the perturbation technique can be found. It has been shown, using a phase plane analysis[4] that 'a' should be less than one-half, for locking to occur. No such fundamental restriction on ϵ exists. Although choosing ϵ less than unity restricts the range of usefulness of the solution, the solution obtained using the perturbation technique is valid even when the phase error, $\phi(\tau)$, and the frequency error, $\phi'(\tau)$, are large. This should be contrasted to the results that can be obtained using a linear analysis, where although ϵ can take on any value,

the solution is restricted to small values of $\phi(\tau)$ and small values of $\phi'(\tau)$.

A solution to Eq. (8), can be written in the form

$$\phi(\tau)=\Phi_o(\tau)+\epsilon\Phi_1(\tau)+a\Phi_2(\tau)+O(\epsilon^2,a^2,\epsilon a) \quad . \quad (9)$$

Substituting Eq. (9) into Eq. (8), and neglecting second order terms, one obtains,

$$\Phi_o'' + \sin \Phi_o = 0 \ , \qquad (10a)$$

$$\Phi_1' + \frac{\sin \Phi_o}{\Phi_o'} \Phi_1 = -\frac{1}{\Phi_o'} \int_o^\tau \Phi_o'^2 \cos \Phi_o d\tau \ , \quad (10b)$$

and

$$\Phi_2' + \frac{\sin \Phi_o}{\Phi_o'} \Phi_2 = -\frac{1}{\Phi_o'} \int_o^\tau e_m'(\tau)\Phi_o'(\tau)d\tau \qquad (10c)$$

Equation (10a), is the equation of the pendulum. Equations (10b), (10c), and the differential equations resulting from higher order terms, are all first order linear differential equations, with time-varying coefficients. The solution to this type of equation is well known.

These equations can be solved in terms of elliptic functions and integrals.[5] The solution for the case of a frequency ramp modulated input signal is shown in Fig. 2. In this figure the parameters chosen were

$$a = 0.157 \ , \ \epsilon = 0.25 \ ,$$

and $\left.\frac{d\phi}{d\tau}\right|_{\tau=o} = \Omega_{n_o} = 1.6$, when $\phi(o) = 0$.

An interesting characteristic of the perturbation solution, as seen in the figure, is that not only does the period decrease with time, τ, but the negative half cycle takes longer to complete than does the preceding positive half cycle. The reason for this is that during the positive half cycle $\phi_{max} < \pi/2$. The coefficient of friction, $\epsilon \cos \phi$, although small, is still positive. During the negative half cycle, due to the offset caused by the frequency ramp, the friction term has less effect on the result. In this example, $|\phi_{max}| > \pi/2$, and the friction term is actually negative over part of the negative half cycle. Thus it takes more time to complete this half cycle.

The perturbation solution approaches the solution obtained using the simple linear analysis, as the phase error decreases. Although the perturbation technique yields more accurate results than the linear analysis, it is more difficult to use. A simple technique, which takes into account the occurrence of "negative damping," has also been investigated. This is a piecewise-linear analysis, and is here described.

2. A Piecewise-Linear Analysis

A piecewise-linear solution was obtained using Fig. 3 as a model. Using this solution the effects of various types of frequency modulation were investigated. In addition, the maximum initial frequency error, to achieve locking, was determined. This result is most important when $a \neq 0$. When $a = 0$ (i.e., the frequency of the input signal and the initial frequency of the VCO differ by Ω), the APC System will always eventually lock (synchronize) to the input signal.[6]

Using Eq. (4) and Fig. 3, the piecewise-linear equations become

$$\frac{d^2\phi}{d\nu^2} + \epsilon_1 \frac{d\phi}{d\nu} + \phi = -a_1 \frac{de_m}{d\nu} \quad , \quad -\frac{\pi}{2} \leq \phi \leq \frac{\pi}{2} \quad (11a)$$

and

$$\frac{d^2\phi}{d\nu^2} - \epsilon_1 \frac{d\phi}{d\nu} - \phi = \pi - a_1 \frac{de_m}{d\nu} \quad , \quad -\frac{3\pi}{2} \leq \phi \leq -\frac{\pi}{2} \quad (11b)$$

where $$\nu = \sqrt{\frac{2}{\pi}}\,\tau = \sqrt{\frac{2}{\pi}\,\omega_a\omega_n}\,t$$

$$\epsilon_1 = \sqrt{\frac{2}{\pi}}\,\epsilon = \sqrt{\frac{2}{\pi}\left(\frac{\omega_n}{\omega_a}\right)}\,\zeta$$

and $$a_1 = \sqrt{\frac{\pi}{2}}\,a = \sqrt{\frac{\pi}{2}\,\frac{\alpha}{\omega_a\omega_n}}$$

Using the case of ramp modulation, a comparison between the results obtained using the perturbation, piecewise linear, and simple linear techniques was made. These results are shown in Fig. 2. In this example the parameters chosen were

$$a = 0.157, \ \epsilon = 0.25 \ ,$$

and $\left.\frac{d\phi}{d\tau}\right|_{\tau=o} = \Omega_{n_o} = 1.6$, when $\phi(o) = 0$.

The simple linear solution damps out most quickly. The frequency of the damped oscillations obtained using the piecewise

linear analysis is smaller than that obtained using the simple linear analysis. It is clear from the figure that the piecewise linear analysis approximates the solution obtained using the perturbation technique more closely than does the simple linear solution.

Using Eq. (11a) and (11b) the set of necessary and sufficient conditions for the APC system to synchronize (lock) to a frequency ramp modulated signal can be shown to be

$$a < 1$$

and

$$\left|\frac{d\phi}{d\tau}\right| < \begin{cases} \sqrt{\frac{\pi}{2}}\left(1-\frac{\epsilon_1}{2}\right)(1-a) \ , & \epsilon_1 \ll 2 \ \text{(underdamped)} \\ \sqrt{\frac{\pi}{2}}(\sqrt{2}-1)(1-a) \ , & \epsilon_1 = 2 \ \text{(critically damped)} \\ \sqrt{\frac{\pi}{2}}\frac{(1-a)}{\epsilon_1} \ , & \epsilon_1 \gg 2 \ \text{(overdamped)} \end{cases}$$

when $\phi = -\frac{\pi}{2}$ radians . (12)

These results are plotted in Fig. 4. It is seen from this figure that when the initial phase error is $-\pi/2$, the maximum frequency error, for locking to occur, decreases as the normalized damping coefficient, ϵ_1, increases. This is due to the fact that when the phase error, ϕ, becomes more negative than $-\pi/2$ radians, the damping force becomes negative. However, locking is still possible when the phase error exceeds $-\pi/2$ radians.

As shown in Fig. 4, similar results can be obtained using a phase plane analysis. The approximation of a constant coefficient of damping, in the piecewise linear analysis, results in the differences between the two curves. These results should be contrasted to the results obtained using a simple linear analysis. Using the simple linear analysis it can be shown that locking will always occur; a result which is obviously incorrect when $a \neq 0$.

The piecewise-linear solution can also be used to obtain the response of the APC System to an arbitrary frequency modulated signal. Equations (11a) and (11b) represent linear equations and superposition can be used to determine the response of the system to any type of periodic modulation.

To illustrate this consider a phase modulated signal,

$$e_c(t) = \sin(\omega_1 t - A \sin \omega_m t) \ . \quad (13)$$

To avoid distortion the phase error, ϕ, should be less than $\pi/2$ radians. Using Eq. (11a), the steady-state phase error becomes

$$\phi(t)_{\text{steady state}} = A \frac{\delta^2(\delta^2-1)}{(\delta^2-1)^2+\epsilon_1\delta^2}\left[\sin \omega_m t + \frac{\epsilon_1 \delta}{\delta^2-1}\cos \omega_m t\right] \quad (14)$$

where

$$\delta^2 = \frac{\pi}{2}\left(\frac{\omega_m^2}{\omega_a \omega_n}\right)$$

and

$$\epsilon_1 = \sqrt{\frac{2}{\pi}\left(\frac{\omega_n}{\omega_a}\right)}\,\zeta \ .$$

It is seen that if

$$\delta^2 \gg 1 \quad (15a)$$

and

$$\delta \gg \epsilon_1 \ , \quad (15b)$$

the steady-state phase error, $\phi(t)$ will faithfully reproduce the input phase modulation. In this case

$$\phi(t)_{\text{steady state}} = A \sin \omega_m t \ . \quad (16)$$

Using a simple linear analysis similar results can be obtained. Equation (15a) requires that the fundamental frequency of the modulating signal, ω_m, be much greater than $\sqrt{\omega_a \omega_n}$, which is the geometric mean of the loop gain of the APC System and the coefficient of integration of the phase lag filter. Eq. (15b) requires that $\omega_m \gg \zeta\omega_n$; i.e., the fundamental frequency of the modulating signal, ω_m, be much greater than the coefficient of friction, $\zeta\omega_n$.

If in steady-state operation $\phi(t)$ exceeds $\pi/2$ radians (this occurs if $A > \pi/2$), the phase error $\phi(t)$ becomes distorted and is no longer a faithful reproduction of the input modulation. Although Eq. (16) shows that it is possible for $\phi(t)$ to be proportional to the input phase modulation, the voltage e_d, is proportional to the sin ϕ not to ϕ.

Then

$$e_d(t) \simeq G_1 \sin(A \sin \omega_m t) \qquad (17a)$$

Expanding in a Fourier series, Eq. (17a) becomes

$$e_d(t) \simeq 2G_1 J_1(A) \sin \omega_m t + 2G_1 J_3(A) \sin 3\omega_m t + \ldots . \qquad (17b)$$

If $A \geq 1$ radian, the error in making the approximation

$$e_d(t) \simeq 2G_1 J_1(A) \sin \omega_m t \qquad (17c)$$

is greater than ten per cent.

It is seen, from the above results, that the APC System studied, can be used as an FM receiver if the modulation index is small. An advantage of this type of receiver is that it can be used to communicate with an accelerating target.

B. The Response of an APC System to Colored Gaussion Noise

Since the APC System is a nonlinear closed loop device, its response to noise cannot simply be determined by using ordinary linear feedback theory, and a Spectral Density analysis. To determine the statistics of the noise at e_d and at e_o, an iteration technique was utilized.

1. The Iteration Technique

The iteration technique employed in this analysis can be explained using Fig. 5. Consider that initially one assumes that $e_{o_o} = 0$. Then,

$$\theta_{n_o} = \int_o^t G_2 e_{o_o} dt = 0 .$$

Using Fig. 5-a, the statistics of e_d, can be determined once the statistics of the noise are specified. The statistics of e_{o_1} can be calculated once e_{d_1} is determined, since e_{o_1} and e_{d_1} are related by the transfer function of the active phase lag filter. It should be noted that since the phase lag filter involves an integration, the statistics of e_o will not be stationary. Using successive iterations, as illustrated in Fig. 5-b, the statistics of e_{o_i} can be determined.

2. The Statistics of e_d and e_o Due to Colored Gaussion Noise

The results obtained below refer to an input consisting of narrow-band noise. The noise was derived by passing white Gaussion noise through a single tuned IF filter, having a center frequency, ω_2, and a bandwidth, α. For such a circuit the correlation function of the noise is given by,

$$R_n(\tau) = \sigma_n^2 e^{-\alpha|\tau|} \cos \omega_2 \tau . \qquad (18)$$

Using the iterative procedure outlined above the statistics of e_{d_i} and e_{o_i} can be found.[7] e_d and e_o were found to be Gaussian distributions. The expected value of e_{d_i} is

$$E(e_{d_i}) = 0 . \qquad (19a)$$

The correlation function of e_{d_i} is

$$R_{e_{d_i}}(\tau) = \frac{\omega_o}{\alpha} \sigma_n^2 G_1^2 e^{-\omega_o|\tau|} \qquad (19b)$$

where ω_o is the bandwidth of the RC low pass filter shown in Fig. 5, and G_1 is the gain of the phase detector. As e_{d_i} is stationary the spectral density can be easily determined.

$$S_{e_{d_i}}(\omega) = \left(\frac{\omega_o \sigma_n^2}{\alpha}\right) \frac{\omega_o G_1^2}{\omega^2 + \omega_o^2} . \qquad (19c)$$

The corresponding statistics for e_{o_i} are given below. The expected value of e_{o_i} is zero. Its correlation function, $R_{e_{o_i}}(t,\tau)$ is

$$R_{e_{o_i}}(t,\tau) \simeq \frac{\omega_o \sigma_n^2 G_1^2}{\alpha} \left[\zeta^2 e^{-\omega_o|\tau|} + 2\zeta\left(\frac{\omega_a}{\omega_o}\right) e^{-\frac{\omega_a|\tau|}{A}} + \left(\frac{\omega_a}{\omega_o}\right)\frac{A}{2}\left(1-e^{-\frac{2\omega_a t}{A}}\right)\left(1+e^{-\frac{\omega_a|\tau|}{A}}\right)\right] \qquad (20a)$$

where A is the gain of the operational amplifier, and

$$\omega_o t \gg 1 .$$

It is observed that the correlation function of e_{o_i} is nonstationary as previously noted.

To compare the theoretical results to the experimental results which are obtained by time averaging, the "time dependent" autocorrelation function[8] was calculated. The time dependent autocorrelation function, $\mathcal{R}_{e_{o_i}}(\tau)$, is

$$\mathcal{R}_{e_{o_i}}(\tau) = \frac{\omega_o \sigma_n^2}{\alpha} G_1^2 \zeta^2 e^{-\omega_o |\tau|} \quad . \quad (20b)$$

Since e_{d_i} is stationary, the time averaged and ensemble averaged correlation functions of e_{d_i} are the same.

The results obtained for e_d (Eq. (19a), (19b), and (19c)) could be obtained if the system operated as an "open loop" system, rather than as a closed loop system. The open loop system would be obtained by disconnecting e_o from the VCO in Fig. 1. The VCO then would transmit a constant frequency signal.

From the preceding analyses it is seen that the APC System analyzed operates as a closed loop system with respect to an input signal, but as an open loop system with respect to the noise.

II. Experimental Results

An experimental model of the APC System was constructed and tested. The gain, G_1, of the phase detector was found to be

$$G_1 = 0.25 \quad . \quad (21a)$$

The sensitivity of the voltage controlled oscillator was measured and is

$$G_2 = 1{,}700\ \pi \text{ rad/sec/volt} \quad (21b)$$

at a center frequency of 35 kcps. The loop gain (sensitivity) of the system, ω_n, is then

$$\omega_n = G_1 G_2 S = 425\pi S \text{ rad/sec} \quad (21c)$$

where $e_c = S \sin \phi_1$.

The equation describing the operation of the system (Eq. (3)) then becomes

$$\frac{d^2\phi}{dt^2} + 425\pi S\zeta \cos\phi \frac{d\phi}{dt} + 425\pi S\omega_a \sin\phi = -\alpha \frac{de_m(t)}{dt} \quad (21d)$$

A. A Comparison of Experimental Results With Theoretical Results When the Input Signal Is Modulated by a Frequency Ramp

To determine the transient response of the system to an input signal, the difference frequency voltage, $e_d(t)$, was monitored using a recorder. It should be noted from Eq. (2a) that $e_d(t)$ is proportional to $\sin\phi$, not to the phase error, ϕ.

To compare the experimental results with the theoretical results obtained using the perturbation, piecewise linear, and simple linear techniques shown in Fig. 3, the system coefficients were chosen to be,

$$\epsilon = 0.25$$

and

$$a = 0.157 \text{ radians.} \quad (22a)$$

Choosing a signal strength $S = 1$ volt and $\omega_a = 1$ rad/sec. α, and ζ were calculated using Eq. (4).

$$\alpha \simeq 33 \text{ cps/sec}$$

$$\zeta = 6.94 \times 10^{-3} \quad (22b)$$

The difference frequency signal, e_d , was recorded for two cases: that of a frequency ramp with $\alpha = 0$, and a frequency ramp with an $\alpha = 33$ cps/sec. The result when the system is unmodulated ($\alpha = 0$) is shown in Fig. 6. The response shown in the figure is under damped as expected. When the phase error exceeded $\pi/2$ radians, the time to complete a period of the motion was greater than when the phase error was much less than $\pi/2$ radians. The time per cycle was approximately 0.2 sec when the phase error exceeded $\pi/2$ radians and was approximately 0.11 sec when the phase error was less than $\pi/2$ radians. When the phase error became small, the system seemed to stop locking, as a continual phase jitter was present. This phase jitter was inherent in the input signal source.

It is when one observes the locking phenomenon as shown in Fig. 6, that it becomes apparent that the simple linear solution is extremely restrictive. Locking begins when the phase error is much greater than $\pi/2$ radians. This is seen by the dip in the peak of the curve. The dip indicates that $e_d(= \sin\phi)$ becomes less than unity, and therefore the phase error, ϕ , becomes greater than $\pi/2$. During the time that it takes the maximum phase error to decrease from π radians to 1 radian, the simple linear solution is of little value.

Using the results obtained when the input signal was ramp modulated (α=33 cps/sec),

a comparison was made between the theoretical and experimental results. The time per cycle and the attenuation (damping) per period for each type of solution is given in Table I. The average time per cycle as shown in Table I is broken into two parts. This was done since the period of the damped oscillations of the system varied as the system locked. This occurred in the experimental and perturbation results, but did not occur in the results obtained using the piecewise linear and the simple linear solution. The "average damping factor" was obtained by considering the phase error, ϕ, to be approximated by a damped sinusoid of the form

$$e^{-m\tau} \cos \beta\tau \quad . \tag{23}$$

The "average damping factor," m, can then be determined.

The results given in Table I clearly indicate that the perturbation technique approximates the experimental results more closely than the piecewise linear or the solution obtained using the simple linear approximation. The experimental and perturbation results would have checked even more precisely if the initial conditions of each would have been closer. The maximum phase error obtained with the perturbation technique was 1.6 radians, while the maximum phase error obtained using the experimental results was 2 radians. This resulted in the discrepancies that occurred in the "time per cycle" readings. The damping per period occurring experimentally is seen to be the same as the damping per period obtained using the perturbation technique. This indicates that while the time per cycle is a function of the maximum phase error, the damping per period is not.

B. The APC System as a Phase Demodulator

As mentioned previously the APC System can be used as an FM demodulator, where the demodulated output is proportional to the phase modulation.

Photographs were taken to illustrate the operation of the APC System as a phase demodulator when a square wave frequency modulated input signal was applied. Sketches of these photographs are shown in Figs. 7-a and 7-b. The modulating frequency used was 110 cps, and the carrier frequency was 35 kcps. The sensitivity of the system, $\omega_n = 1{,}275\ \pi$ rad/sec. Figure 7a represents the input modulating signal. the demodulated waveform is the integral of the square wave frequency modulating signal, as the demodulated output is proportional to the phase modulation. This is seen in Fig. 7-b. The thickness of the saw tooth waveform is due to inadequate filtering of the second, third and higher harmonics of the 35 kcps carrier frequency. At lower modulating frequencies the saw tooth became distorted as the necessary conditions outlined in Eq. (15) were not satisfied.

As can be seen from this figure the APC System represents an excellent demodulation device.

C. The Experimental Response of an APC System to Colored Gaussian Noise

Before determining the response of the APC System to a signal embedded in noise, the behavior of the system to colored Gaussian noise (when no signal was present) was investigated. The theoretical aspect of this investigation was presented above. That discussion and the discussion of the signal plus noise[9] rely on the fact that the response of the system to noise can be analyzed using the iteration technique. One of the more important results obtained using this technique was that the system acted as an open loop system to the noise. The following discussion describes the experiments performed and shows that the theoretical model is correct.

The colored Gaussian noise used in this experiment had a center frequency of 35 kcps and a bandwidth of 6 kcps. Using a true rms meter, the relation between the rms input noise, $e_{d_{rms}}$, and $e_{o_{rms}}$, were determined using different parameters of the active phase lag filter. The results of this measurement are shown in Table II.

It should be observed that the readings are independent of ω_a. The measured results shown in the Table are compared to the results calculated using Eq. (19b) and (20b). The rms values of e_d and e_o are therefore given by

$$e_{d_{rms}} = G_1 \sqrt{\frac{\omega_o}{\alpha}}\, \sigma_n \tag{24a}$$

and

$$e_{o_{rms}} = G_1 \zeta \sqrt{\frac{\omega_o}{\alpha}}\, \sigma_n \tag{24b}$$

where $\omega_o = 2\pi(1{,}000)$rad/sec, and $\alpha = 2\pi(6{,}000)$rad/sec. The measured values definitely verify the calculated values.

The measurements were made when the system operated normally in the closed loop position, and again, with the system operating as an open loop system with e_o disconnected from the VCO. The same values were obtained in each case for $e_{d_{rms}}$ and $e_{o_{rms}}$. This clearly indicates that e_v and e_c (Fig. 5) are uncorrelated, and as a result of this the noise sees an open loop system.

To determine the probability distribution of e_d the voltage was recorded, and 24 amplitude intervals and 1,700 samples taken from the recording. The probability density was determined and plotted on probability paper as shown in Fig. 8. The resulting curve approximates a straight line and therefore approximates the Gaussian distribution found theoretically.

A χ^2 "goodness of fit" test[10] was performed. The results obtained indicate that by chance alone, if the probability distribution is Gaussian, then with a probability of 0.99, the χ^2 test would indicate a worse fit than the one actually observed.

III. Conclusions

The comparison made between the theoretical and the experimental results indicate that the perturbation technique and the piecewise linear technique yield excellent approximations to the second order nonlinear differential equation studied.

The APC System is capable of locking to a signal having frequency ramp modulation with a slope, α, of less than $1/2\ (\omega_a\omega_n)$. The device is also able to be used as a phase demodulator, if the phase modulation index is less than $\pi/2$ radians (i.e., narrow-band FM).

The use of an iteration technique allows the response of the APC System to noise to be obtained without "linearizing" the system equation. Using this technique it was shown that the system investigated responds to noise as an open loop system, and therefore the attenuation of the noise is due to the open loop bandwidth of the APC System.

IV. Acknowledgment

The work done on this project was performed at the Electronics Research Laboratories of Columbia University.

The authors express their appreciation to Dr. R. I. Bernstein of Columbia University Electronics Research Laboratories for suggesting the problem as a thesis topic to Mr. Schilling. This paper represents a portion of the dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Electrical Engineering at the Polytechnic Institute of Brooklyn.

The authors acknowledge the assistance and encouragement given by Professor L. H. O'Neill, Director, and Mr. L. B. Lambert, Laboratory Supervisor, Columbia University Electronics Research Labortories.

V. References

1. Kreindler, E., "The Theory of Phase Synchronization of Oscillators with Application to the DOPLOC Tracking Filter," Technical Report T-1/157, Electronics Research Laboratories, School of Engineering, Columbia University, New York 27, N. Y., August 1, 1959.

2. Milne-Thomson, L. M., Jacobian Elliptic Function Tables, Dover Publications, Inc., New York, N. Y., 1950.

3. Stoker, J. J., Nonlinear Vibrations, Interscience Publishers, Inc., New York, N. Y., 1950.

4. Viterbi, A. J., "Acquisition and Tracking Behavior of Phase-Locked Loops," Jet Propulsion Laboratory, California Institute of Technology, External Publication No. 673, July 14, 1959.

5. Schilling, D. L., "The Response of an Automatic Phase Control System to Frequency Modulated Signals and Noise," Doctoral Dissertation, Polytechnic Institute of Brooklyn, Brooklyn, N.Y., June 1962.

6. Viterbi, A. J., op. cit.

7. Schilling, D. L., op. cit.

8. Davenport and Root, Random Signals and Noise, McGraw-Hill Publishing Co., Inc., New York, N. Y., 1958.

9. Schilling, D. L., op. cit.

10. Cramer, H., Mathematical Methods of Statistics, Princeton University Press, 1946.

TABLE I

	Averaged Normalized Time (τ) Per Cycle		m Averaged Attenuation Per Normalized Period $(Rad)^{-1}$
	Phase Error $>$ 1 Rad	Phase Error $<$ 1 Rad	
Experimental	10.4	8	0.0625
Perturbation	7.75	7.4	0.0635
Piecewise Linear	7.85	7.85	0.1
Simple Linear	6.3	6.3	0.125

$$\phi(\tau) \sim e^{-m|\tau|} \cos \beta\tau$$

TABLE II

ζ	σ_n (volts)	Measured (volts)		Calculated (volts)	
		$e_{d_{rms}}$	$e_{o_{rms}}$	$e_{d_{rms}}$	$e_{o_{rms}}$
6.8×10^{-3}	3.5	0.38	2.8×10^{-3}	0.36	2.4×10^{-3}
47×10^{-3}	3.5	0.38	17.5×10^{-3}	0.36	17.9×10^{-3}
100×10^{-3}	3.5	0.38	40×10^{-3}	0.36	38×10^{-3}
820×10^{-3}	3.5	0.38	330×10^{-3}	0.36	312×10^{-3}

$$e_{d_{rms}}(\text{calc}) = G_1 \sqrt{\frac{1000}{6000}}\, \sigma_n$$

$$e_{o_{rms}}(\text{calc}) = \zeta\, e_{d_{rms}}$$

$$\frac{1}{11} \leq \omega_a \leq 1$$

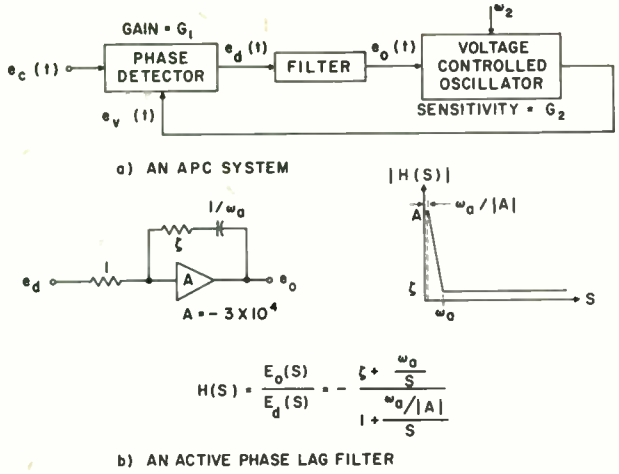


FIG. 1 AN AUTOMATIC PHASE CONTROL SYSTEM

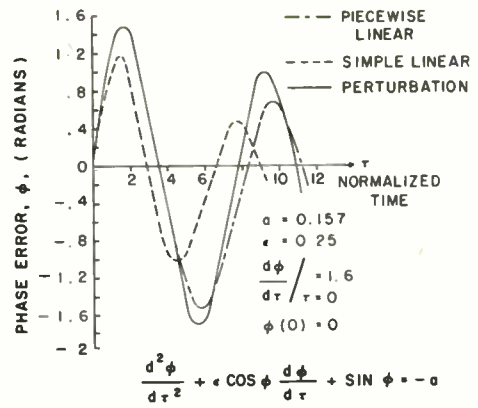


FIG. 2 A COMPARISON OF THE APPROXIMATE SOLUTIONS USED TO OBTAIN THE RESPONSE OF THE APC SYSTEM TO A FREQUENCY RAMP MODULATED SIGNAL

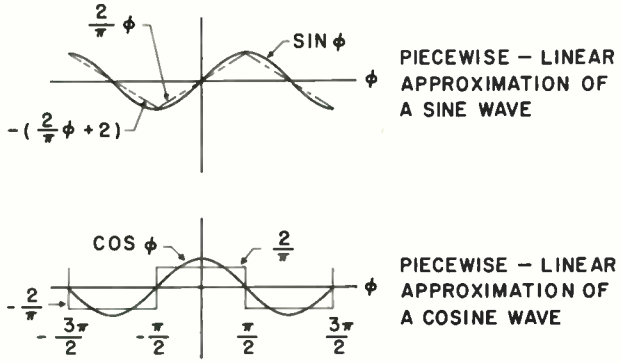


FIG. 3 A PIECEWISE LINEAR APPROXIMATION

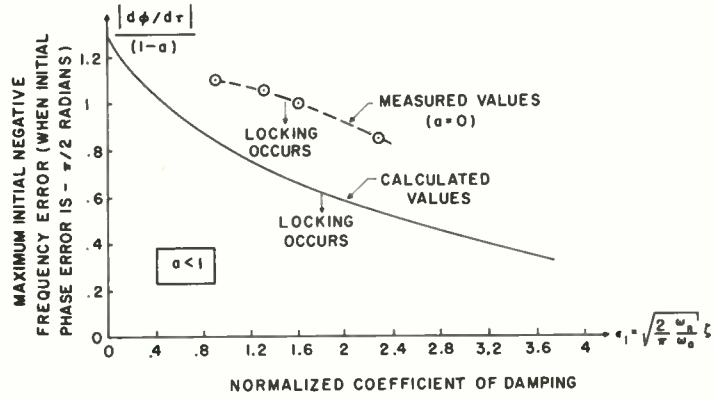


FIG. 4 THE NECESSARY AND SUFFICIENT CONDITIONS FOR THE APC SYSTEM TO LOCK TO A FREQUENCY RAMP MODULATED SIGNAL

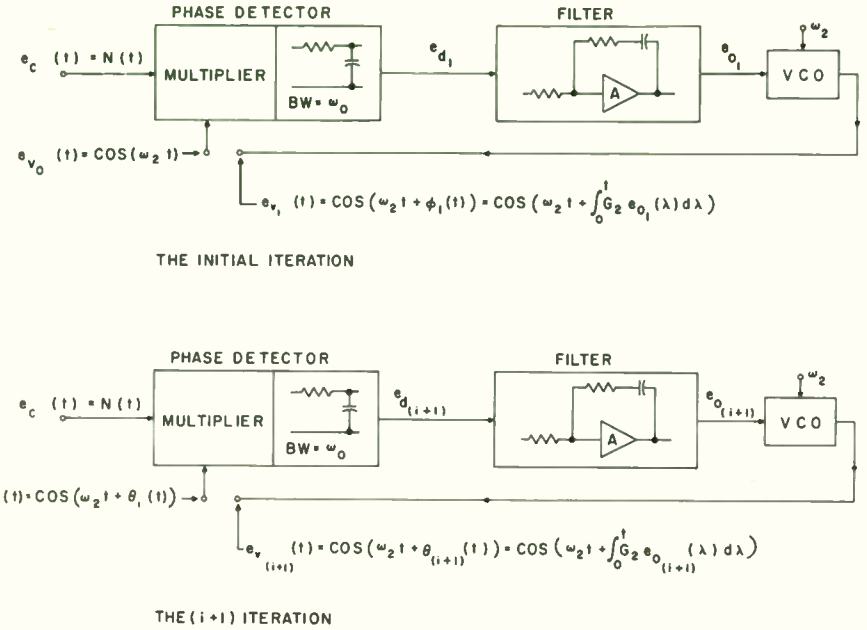


FIG. 5 THE ITERATION TECHNIQUE

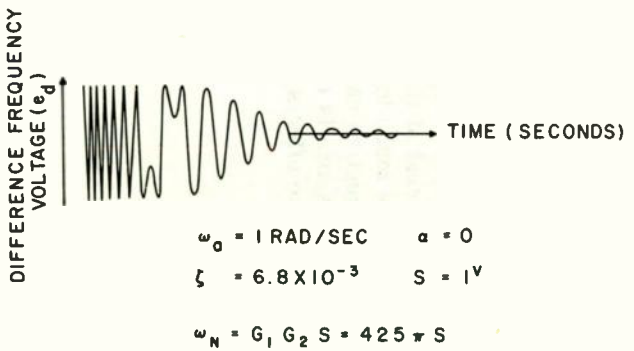


FIG. 6 THE RESPONSE OF AN APC SYSTEM TO AN UNMODULATED INPUT SIGNAL

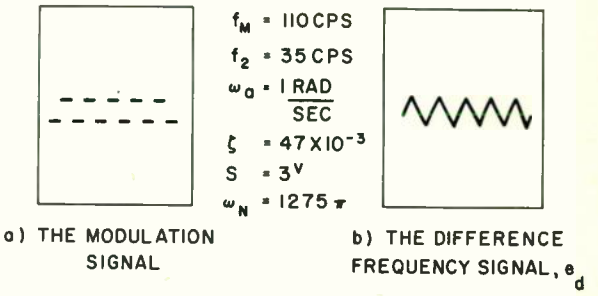


FIG. 7 THE RESPONSE OF THE APC SYSTEM TO A FREQUENCY MODULATED SQUARE WAVE

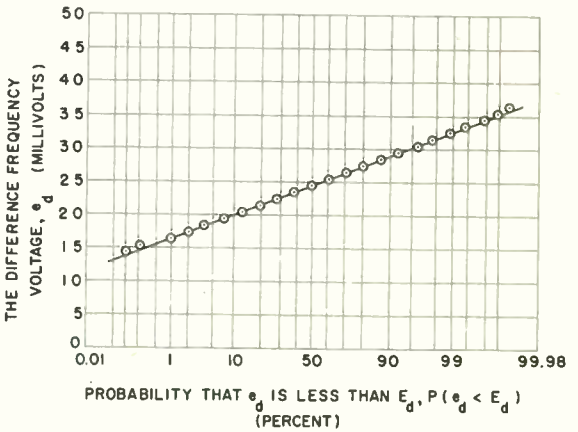


FIG. 8 EXPERIMENTAL VERIFICATION OF THE GAUSSIAN RESPONSE OF THE DIFFERENCE FREQUENCY VOLTAGE, e_d, TO GAUSSIAN INPUT NOISE

EFFECTS OF TIME DELAY AND ECHOES ON TELEPHONE CONVERSATIONS

J. W. Emling and D. Mitchell
Satellite Sys. Studies Dept.
Bell Telephone Labs., Inc.
Murray Hill, N. J.

Abstract

This paper presents a summary of our present knowledge on the effects of time delay, echoes and echo suppressors on telephone conversations. It is essentially tutorial and covers the historical background of this subject. This seems of particular interest at this time because circuits using satellite links would have much longer delay than most present-day toll circuits. The subjective nature of impairment due to time delay and echo suppressors is emphasized, along with the need for more data on these effects from actual telephone conversations.

TRANSMISSION PROGRAMMING FOR ENERGY CONSERVATION IN SPACE COMMUNICATIONS

Norman S. Potter
Manager, Systems Engineering
Missiles & Surface Radar Division
Radio Corporation of America
Moorestown, New Jersey

Abstract

An analytic and quantitative study is performed of the programming of transmitter power output to minimize the energy consumption required to maintain a specified minimal information rate from an astronautical vehicle to a terrestrial receiving system over the trajectory of the former. The representative trajectories used as a basis of discussion include an idealized deep-space probe with variable terminal ranges, and solar circular orbits at variable radii in the plane of the ecliptic. For the first, consideration of the integral of the admissible lower bound to the radiated power shows that the minimal energy consumption is approximately half of that resulting from a constant power output program. An optimal two-stage program is shown to be an efficient approximation to the idealized one of continuous power variation. For the orbital case, the economies resulting from an optimal program are seen to be moderately functionally dependent upon the orbital radius, but independent of the sense of rotation of the vehicle. As before, a satisfactory two-stage fixed transmitter program may be defined, in which the optimal time of transition between stages is independent of the orbital radius and sense of rotation.

Variational methods are used to determine a lower bound to the energy consumption if the totality of information output is specified and the system is assumed to be operating at the maximum channel capacity corresponding to the instantaneous power radiation level. It is shown that if the mission duration and the totality of information communicated are both large, independently of the trajectory the resultant minimal energy consumption corresponds to the difference of those of a constant power output system and one programmed to yield a fixed signal power density at the receiver. Lastly, generally applicable contours are provided which display the relative energy consumptions involved in the optimal and approximating power radiation programs, and the transmitter-receiver separations at the time that switching of the transmitter power level is to be accomplished for both categories of trajectories.

1. Introduction

The energy starvation which characterizes astronautical missions leads to a requirement that communication transmissions from the vehicle be accomplished on a minimal energy consumption basis. No less pressing are the associated problems of system reliability which, in a related sense, suggest the desirability of effecting the necessary information transmission with minimal communication system on-time. The investigation reported herein concerns itself in part with the optimal programming in the time domain of the mean power output of the transmitter as a function of the vehicle trajectory in a central force field, if the totality or rate of information to be transmitted is used as a constraint.

To facilitate the discussion, two representative categories of astronautical probes are considered. In the first, the motion of the deep-space probe vehicle is taken as anti-parallel to the gravitational vector of the earth which, for simplicity, is regarded as the only contributor to the idealized force field. In the second, a family of circular orbital trajectories, concentric with the sun and in the plane of the ecliptic, is investigated.

An important example of the latter would be the delivery of a transmitter in a soft-landing planetary probe. The orbit of the planet would then correspond to that of the astronautical vehicle discussed herein if long-term telemetering of locally derived data is required.

2. Elementary System Optimization

It is reasonable to specify the minimal admissible signal to noise ratio σ at the terrestrial receiver as a measure of system performance, and to assure some designated communication channel capacity. Assuming the availability of sufficient peak and/or average power, the most inefficient program which may be pursued would consist of transmission over the entire trajectory at the power output P required at the maximum receiver-

transmitter separation. If the density of the information output is uniform in time, the required energy $\mathcal{E}$ would then be proportional to PT, where T denotes the duration of the mission. Alternately, if the power output $p(t)$ were continuously varied in accordance with the changing receiver-transmitter relative geometry so as to yield a fixed σ, the integral with respect to time of $p(t)$ would then provide an estimate of the minimal energy requirement. Plainly, the computations hypothesize such simplifying conditions as appropriate vehicle-antenna stabilization and/or the uniformity of the gain of the latter, as well as the fact that, in the circumstances that are of interest, the transmitter-receiver separation is large relative to the radius of the earth.

2.1 Mission Duration

The parameter T is readily determined for the two categories of trajectories. For the first, if ρ denotes the separation of the astronautical vehicle and the center of the earth whose radius is R, and k^2 the product of the mass of the latter and the gravitational constant, one has immediately

$$\frac{d^2\rho}{dt^2} = -\left(\frac{k}{\rho}\right)^2$$

or

$$\left(\frac{d\rho}{dt}\right)^2 = 2k^2\left(\frac{1}{\rho} - \frac{1}{\tilde{\rho}}\right) \tag{1}$$

where $\tilde{\rho}$ is the upper bound of ρ if, to characterize the end of the mission, the velocity of the vehicle vanishes for $\rho = \tilde{\rho}$. The return of the space probe if escape velocities are not reached is, of course, symmetric in time. Accordingly,

$$T = \frac{1}{k}\sqrt{\frac{\tilde{\rho}}{2}}\int_R^{\tilde{\rho}} \frac{\sqrt{\rho}}{\sqrt{(\tilde{\rho}-\rho)}}\,d\rho = \frac{\tilde{\rho}^{3/2}}{k\sqrt{2}}\int_{\sin^{-1}\sqrt{(R/\tilde{\rho})}}^{\pi/2} (1-\cos 2\theta)\,d\theta$$

$$= \frac{\tilde{\rho}^{3/2}}{k\sqrt{2}}\left\{\sqrt{\left(\frac{R}{\tilde{\rho}}\left(1-\frac{R}{\tilde{\rho}}\right)\right)} + \cos^{-1}\sqrt{\frac{R}{\tilde{\rho}}}\right\} \tag{2}$$

$$= r^{3/2}\sqrt{\frac{R}{2g}}\left\{\frac{\sqrt{(r-1)}}{r} + \cos^{-1}1/\sqrt{r}\right\}$$

where the right-hand integral in the first line of (2) follows from the substitution $\rho = \tilde{\rho}\sin^2\theta$ and, in the final form, r denotes $\tilde{\rho}$ in astronomical units and $k^2 = gR^2$.

To facilitate the subsequent analysis, in the orbital case, the mission duration T is defined as the interval required to pass from the minimum to maximum transmitter-receiver separation, or, alternately stated, from inferior to superior conjunction with respect to the sun. If K^2 denotes the product of the latter's mass and the gravitational constant, and R_e and R_a are, respectively, the orbital radii of the earth and the astronautical vehicle, their angular rates are, from Kepler's law,

$$K/R_e^{3/2} \quad \text{and} \quad K/R_a^{3/2},$$

subject to the assumption that the sun's attraction completely defines the force field. Clearly, one then has for

$$R_e \geq R_a, \quad T\left(\frac{K}{R_a^{3/2}}\right) = T\left(\frac{K}{R_e^{3/2}}\right) + \pi$$

or

$$T = \frac{\pi}{K}\Big/\left(\frac{1}{R_a^{3/2}} - \frac{1}{R_e^{3/2}}\right)$$

and, for (4)

$$R_a \geq R_e, \quad T\left(\frac{K}{R_e^{3/2}}\right) = T\left(\frac{K}{R_a^{3/2}}\right) + \pi$$

or

$$T = \frac{\pi}{K}\Big/\left(\frac{1}{R_e^{3/2}} - \frac{1}{R_a^{3/2}}\right)$$

if the sense of rotation is the same for both. Identical equations result if one considers the time required for passage from superior to inferior conjunction. The preceding observations and a similar analysis for the case in which the astronautical vehicle's rotation is retrograde leads to the conclusion that the time interval required for passage from inferior to superior conjunction and conversely are equal, both being given by

$$T = R_e^{3/2}\,\pi\, r_a^{3/2}/K\left|r_a^{3/2} \pm 1\right|,$$

where the negative or positive sign is selected according as the rotations are similarly or oppositely sensed, and $r_a = R_a/R_e$ is the orbital radius of the probe in astronomical units. Lastly, from

$$K = 2\pi R_e^{3/2}/T_e,$$

where T_e denotes the earth's period,

$$T = r_a^{3/2}T_e/2\left|r_a^{3/2} \pm 1\right|.$$

The resultant mission durations are displayed in Figure 1 to provide some insight into the magnitude of the problem. The solid contour, which relates to the rectilinear trajectory, displays the required time in earth rotational periods as a function of the terminal altitude r_h, in multiples of earth radii. Replacing the irrational and trigonometric terms by their Maclaurin developments, and r by $r_h + 1$, one obtains

$$T = (r_h+1)^{3/2}\sqrt{\frac{R}{2g}}\left\{\frac{\sqrt{r_h}}{r_h+1} + \cos^{-1}\frac{1}{\sqrt{(r_h+1)}}\right\}$$

$$= \sqrt{\frac{R}{2g}}\left\{\left[r_h - \frac{1}{2} + O\left(\frac{1}{r_h}\right)\right] + r_h^{3/2}\left[\frac{\pi}{2} - \frac{1}{\sqrt{r_h}} - \frac{1}{6r_h^{3/2}} + O\left(\frac{1}{r_h^{5/2}}\right)\right]\right\}$$

whence

$$T = \sqrt{\frac{R}{2g}}\left\{\frac{\pi}{2}\rho_h^{3/2} - \frac{1}{3}\rho_h + O\left(\frac{1}{\rho_h}\right)\right\} \tag{5}$$

$$\rightarrow \pi\left(\frac{\rho_h}{2}\right)^{3/2}\sqrt{\frac{R}{g}} - \frac{1}{3}\rho_h\sqrt{\frac{R}{2g}}$$

where the resultant asymptotic estimate applies when ρ_h is large, a condition that defines the circumstances under which energy conservation is of interest. The resultant expression in (5) is dominated by its first term, so that one has $T \propto \rho_h^{3/2}$. The contour demonstrates the rapid convergence to this form which appears linear in the logarithmic scale that is used.

With respect to the orbital trajectories, excepting the region in the immediate vicinity of the terrestrial orbit the mission durations, displayed as dashed contours, are seen to be essentially same. In particular, both rapidly converge to a half-year which, from examination of the indicated planetary radii, defines the required intervals for the Jovian planets.

2.2 Energy Considerations

2.2.1 The Idealized Rectilinear Trajectory.

The ratio of the minimal to maximal energy requirements $\mathcal{E}$ is, from

$$p(t) \propto (\rho(t) - R)^2$$

$$\mathcal{E} = \frac{1}{T[\tilde{\rho}-R]^2}\int_0^T \frac{(\rho-R)^2}{(d\rho/dt)}\,d\rho = \frac{1}{RT[\tilde{\rho}-R]^2}\sqrt{\frac{\tilde{\rho}}{2g}}\int_R^{\tilde{\rho}} (\rho-R)^2\sqrt{\frac{\rho}{\tilde{\rho}-\rho}}\,d\rho \tag{6}$$

or, with the substitution

$$\rho = \tilde{\rho}\sin^2\theta$$

used earlier and replacing T by (2),

$$\mathcal{E} = \frac{2}{\left(1-\frac{1}{\rho}\right)^2\left(\frac{\sqrt{(\rho-1)}}{\rho} + \cos^{-1}\frac{1}{\sqrt{\rho}}\right)}\int_{\sin^{-1}\frac{1}{\sqrt{\rho}}}^{\pi/2}\left(\sin^2\theta - \frac{1}{\rho}\right)^2\sin^2\theta\,d\theta$$

$$\sim \frac{4}{\pi}\int_0^{\pi/2}\sin^6\theta\,d\theta = 5/8. \tag{7}$$

The limiting value indicated above for ρ infinitely large provides an upper bound to the economies that may be effected by these means in this instance. The reason for its modesty is, of course, the rapid decline in opening rate due to the gravitational deceleration, which causes the vehicle to spend a comparatively large portion of its mission at the longer range, high power output end of its trajectory.

The preceding computations assumed continuous variation of the power output over an enormous dynamic range which, in practice, is obtainable only at a pronounced degradation of transmitter efficiency if, indeed, it is implementable. Alternately, and more realistically, a multiple stage system, involving discrete increases in power output between fixed stages, could be utilized. Plainly, the system efficiency must increase with the number of optimally chosen increments in power since it then approximates more closely to the continuous variation discussed earlier. Accordingly, if one determines the optimal switching point for a two stage system, a lower bound to the efficiency of multiple stage approximations may be obtained.

Let $\tilde{\mathcal{E}}$ denote, as before, the ratio of the energies required by the (sub)optimized optimized transmitter programs. Since the signal to noise ratio at the terrestrial receiver must equal or exceed some specified figure, if the switching point is located at separation $\tilde{\rho}$ which is reached in time $\tilde{T}$, one must have

$$\tilde{\mathcal{E}} = \frac{(\tilde{\rho}-1)^2\tilde{T} + (T-\tilde{T})(\rho-1)^2}{T(\rho-1)^2}$$

$$= \left(\frac{\frac{\tilde{\rho}}{\rho} - \frac{1}{\rho}}{1-\frac{1}{\rho}}\right)^2\frac{\tilde{T}}{T} + \left(1-\frac{\tilde{T}}{T}\right) \tag{8}$$

$$\sim 1 - \frac{\tilde{T}}{T}\left(1 - \left(\frac{\tilde{\rho}}{\rho}\right)^2\right)$$

where $\tilde{\rho}$, as ρ, is in multiples of earth radii and the asymptotic estimate obtains for ρ large.

Paralleling the earlier analysis, $\tilde{T}$ must be given by

$$\tilde{T} = \rho^{3/2}\sqrt{\frac{R}{2g}}\int_{\sin^{-1}\sqrt{\frac{1}{\rho}}}^{\sin^{-1}\sqrt{\frac{\tilde{\rho}}{\rho}}}(1-\cos 2\theta)\,d\theta$$

$$= \rho^{3/2}\sqrt{\frac{R}{2g}}\left\{\sin^{-1}\sqrt{\frac{\tilde{\rho}}{\rho}} - \sin^{-1}\sqrt{\frac{1}{\rho}}\right.$$

$$\left. + \frac{1}{\sqrt{\rho}}\sqrt{\left(1-\frac{1}{\rho}\right)} - \sqrt{\frac{\tilde{\rho}}{\rho}}\sqrt{\left(1-\frac{\tilde{\rho}}{\rho}\right)}\right\}$$

whence, if $\tilde{r}/r$ approaches a non-zero limit λ for r large, one must have

$$\frac{\tilde{T}}{T} \sim \frac{2}{\pi}\left(\sin^{-1}\sqrt{\lambda} - \sqrt{\lambda}\sqrt{(1-\lambda)}\right)$$

and (9)

$$\tilde{\mathcal{E}} \sim 1 - \frac{2}{\pi}(1-\lambda^2)\left(\sin^{-1}\sqrt{\lambda} - \sqrt{\lambda}\sqrt{(1-\lambda)}\right)$$

$$= 1 - \frac{2}{\pi}\varphi(\lambda).$$

The problem is then one of maximizing the positive function $\varphi(\lambda)$ which vanishes for λ unity or zero, yielding $\tilde{\mathcal{E}} = 1$, since both values define a single stage system. Numeric analysis establishes the location of the desired extremal in the vicinity of $\lambda = 0.7\,(-)$, a result that is in accord with the prior observations upon the time history of the position of the vehicle. The corresponding minimum of $\tilde{\mathcal{E}}$ is found to be 0.69, in the vicinity of which its first derivative is not rapidly changing so the extremal is not critical. Though the improvement is modest, it should be noted that virtually the entirety of the attainable economy is achieved through the addition of only one supplementary transmitter stage.

It is clear that essentially the above conclusion holds uniformly over the domain of r of interest. Recalling the earlier demonstration that for all r sufficiently large to be of interest one has, to a good order of approximation, $T \propto r^{3/2}$, in terms of the present notation,

$$\tilde{\mathcal{E}} \sim 1 - \left(\frac{\tilde{r}}{r}\right)^{3/2}\left(1 - \left(\frac{\tilde{r}}{r}\right)^{2}\right)$$

$$= 1 + \alpha^{7/2} - \alpha^{3/2}, \quad \alpha = \left(\frac{\tilde{r}}{r}\right). \tag{10}$$

(10) is readily shown to have a minimum at $\alpha = \sqrt{(3/7)} = 0.65$, which corresponds to $\tilde{\mathcal{E}} = 1 - (4/7)(3/7)^{3/4} = 0.7$. The preceding figures are seen to correspond favorably with those derived earlier by more precise methods subject to the hypothesis that r is extremely large.

2.2.2 The Orbital Case. The square of the instantaneous transmitter-receiver separation is

$$\rho^2 = \left(R_e \cos\frac{2\pi}{T_e}t - R_a \cos\left(\frac{2\pi k}{T_e}\left(\frac{R_e}{R_a}\right)^{3/2}t + \phi\right)\right)^2$$

$$+ \left(R_e \sin\frac{2\pi}{T_e}t - R_a \cos\left(\frac{2\pi k}{T_e}\left(\frac{R_e}{R_a}\right)^{3/2}t + \phi\right)\right)^2$$

$$= R_e^2 + R_a^2 - 2R_eR_a\cos\left\{\frac{2\pi}{T_e}\left[\left(\frac{R_e}{R_a}\right)^{3/2}k - 1\right]t + \phi\right\} \tag{11}$$

$$= R_e^2\left(1 + r_a^2 - 2r_a\cos\left\{\frac{2\pi}{T_e}\left[\frac{k}{r_a^{3/2}} - 1\right]t + \phi\right\}\right),$$

where k is plus or minus unity according as the senses of rotation are the same or different. ϕ is zero if they are initially at inferior conjunction or π if at superior conjunction. It follows then that

$$\mathcal{E} = \frac{R_e^2}{T(R_e+R_a)^2}\int_0^T\left(1 + r_a^2 - 2r_a\cos\left\{\frac{2\pi}{T_e}\left[\frac{k}{r_a^{3/2}} - 1\right]t + \phi\right\}\right)dt$$

$$= \frac{1 + r_a^2}{(1 + r_a)^2} - \frac{T_e\, r_a}{\pi T\left[\frac{k}{r_a^{3/2}} - 1\right]}\,\sigma \tag{12}$$

where

$$\sigma = \sin\left\{\frac{2\pi T}{T_e}\left[\frac{k}{r_a^{3/2}} - 1\right] + \phi\right\}.$$

Replacing T by $T(R_a)$, the trigonometric term σ assumes the form $\sin(\pi\,\mathrm{signum}(k - r_a^{3/2}) + \phi)$ which vanishes for all combinations of k and ϕ, whence $\mathcal{E}$ is given by $(1 + r_a^2)/(1 + r_a)^2$. The minimal energy consumption program then leads to economies which are functions only of the ratio of orbital radii and are independent of the initial relative orientations and senses of rotation.

As before, the ideally continuous variation introduces excessive design demands, so that a multiple, fixed-stage system may be contemplated. In particular, a large dynamic range is involved when the ratio of orbital radii is near unity, a condition which from the preceding analysis, also characterizes the circumstances under which the most substantial economies are effected. For a single stage system, one may write

$$\tilde{\mathcal{E}} = \left\{\tilde{T}R_e^2\left[1 + r_a^2 - 2r_a\cos\frac{2\pi\tilde{T}}{T_e}\left[\frac{k}{r_a^{3/2}} - 1\right]\right]\right.$$

$$\left. + (T - \tilde{T})[R_e + R_a]^2\right\} / T[R_e + R_a]^2 \tag{13}$$

$$= 1 - \alpha + \frac{\alpha}{(1 + r_a)^2}(1 + r_a^2 - 2r_a\cos\pi\alpha), \quad \alpha = \frac{\tilde{T}}{T},$$

for the passage from inferior to superior conjunction if, as in the earlier analysis, $\tilde{T}$ denotes the switching time to higher power transmission, since

$$\frac{2\pi T}{T_e}\left[\frac{k}{r_a^{3/2}} - 1\right] = \pi \operatorname{signum}(k - r_a^{3/2}). \tag{14}$$

Similarly, for a passage from superior to inferior conjunction

$$\tilde{\mathcal{E}} = \left\{(T-\tilde{T})R_e^2\left(1 + r_a^2 - 2r_a\cos\left(\frac{2\pi\tilde{T}}{T_e}\left[\frac{k}{r_a^{3/2}} - 1\right] + \pi\right)\right)\right.$$

$$\left. + \tilde{T}\,[R_e + R_a]^2\right\} \Big/ T\,[R_e + R_a]^2 \tag{15}$$

$$= \alpha + \frac{(1-\alpha)}{(1+r_a)^2}\left(1 + r_a^2 + 2r_a\cos\pi\alpha\right), \quad \alpha = \frac{\tilde{T}}{T}$$

where $\tilde{T}$ now denotes the switching time from high power to lower power transmission. In both instances $\tilde{\mathcal{E}}$ is independent of the relative senses of rotation.

A choice of zero or unity for α corresponds to the designation of a single stage system, for which $\tilde{\mathcal{E}} = 1$. Differentiation with respect to α leads to

$$1 = \pi\alpha \sin\pi\alpha - \cos\pi\alpha \tag{16}$$

for the first, and

$$1 = \cos\pi\alpha + (1-\alpha)\pi\sin\pi\alpha \tag{17}$$

for the second case in the vicinity of the desired extremals, which are thus seen to be independent of the orbital radii or rotational senses of the bodies involved. Considerations of symmetry suggest that the sum of the zeros of the preceding equations, $0 < \alpha < 1$, must be unity, a fact that is readily demonstrated by replacing α by $1-\alpha$ in either, a substitution that converts the equation in question into the form of the other. This symmetry is, of course, characteristic of equations (13) and (15) as well.

Numeric analysis then shows that if the transmitter and receiver are initially at inferior conjunction, the transition from low to high power output must occur at the end of $0.42\,T$ whereas, if they are initially at superior conjunction, the transition among transmitter stages occurs at $0.58\,T$.

Figure (2) contains contours which quantitatively summarize the results of the analysis of optimal switching points in two-stage systems. $\tilde{r}$, the separations at transfer from lower to higher power transmission levels, and conversely, are displayed as multiples of the maximum separation $r_M = 1 + r_a$ in astronomical units. The computations underlying the contour for the rectangular trajectory are founded upon the asymptotic estimate $T \propto r_M^{3/2}$, for which $\tilde{T} \sim \sqrt{(3/7)}\,r_M$. In the case of orbital trajectories, writing $\tilde{T} = cT$, from (11) one has immediately

$$\tilde{r} = \frac{\sqrt{\left(1 + r_a^2 - 2r_a\cos\left(C\operatorname{signum}(k - r_a^{3/2}) + \phi\right)\right)}}{1 + r_a}. \tag{18}$$

If the transmitter-receiver are initially at inferior conjunction, the preceding becomes

$$\sqrt{\left\{1 - 3.8\,\frac{(r_M - 1)}{r_M^2}\right\}}$$

and (19)

$$\sqrt{\left\{1 - 0.33\,\frac{(r_M - 1)}{r_M^2}\right\}}$$

otherwise. It is seen that, in both, as the orbital radii increase, the transfer point between transmitter stages approaches that of maximal separation. Though both contours exhibit minima when the trajectory of the astronautical vehicle is coincident with the terrestrial orbit, if the transmitter and receiver are initially at inferior conjunction the variation in the optimal transfer point is large to either side of the minimum. By way of contrast, if they are initially at superior conjunction, the optimum is essentially independent of the orbital radius and in the vicinity of the greatest separation.

Figure (3) displays the results of an evaluation of $\mathcal{E}$ and $\tilde{\mathcal{E}}$, where the latter is computed for an optimal choice of the parameter $\alpha = \tilde{T}/T$. The reduction in the required energy consumption is seen to be modest, even with an ideal, continuously variable transmission program. The addition of a single stage is seen to uniformly achieve approximately half the attainable economies. It should be noted that, for both, the energy consumption reduction is greatest for r_a unity, to either side of which it declines.

A more complete measure of the absolute energy savings is provided by the product $(1-\mathcal{E})\,T\,(r_a+1)^2$ which, normalized to the unity orbital radius situation in which the transmitter orbit is retrograde, is

$$\hat{P} = \left[1 - \frac{1 + r_a^2}{(1+r_a)^2}\right]\left[\frac{r_a^{3/2}\,T_e}{2\,|r_a^{3/2} - k|}\right](r_a+1)^2 \Big/ \frac{T_e}{2} \tag{20}$$

or

$$\hat{P} = 2\mathcal{R}_a^{5/2} / |\mathcal{R}_a^{3/2} - k| .$$

The earlier estimates of energy requirements were made relative to what would have been necessitated at the given $\mathcal{R}_a$ had not optimal transmission programming been introduced. By way of contrast, $\hat{P}$ may assume large values even when the percentile economies are modest provided that the energy needs are large, as in long-duration missions. In particular, $\hat{P}$ becomes indefinitely large with $\mathcal{R}_a$ and therefore T, as well as when $\mathcal{R}_a$ is in the vicinity of unity and the senses of rotation of the transmitter and receiver are the same. The preceding characteristics are apparent in the normalized energy consumption reduction contours of Figure (3).

3. Generalized Optimal Transmission Programming

One may approach the problem of minimizing the energy consumption, subject to constraints upon the totality of transmitted information, by the methods of the variational calculus. Evoking the Shannon-Hartley law, the minimization of

$$\mathcal{E} = \int_0^T p(t)\,dt \tag{21}$$

with

$$I = \gamma \int_0^T \ln\left(1 + \beta \frac{p(t)}{\mathcal{R}^2(t)}\right) dt$$

is required, where γ is a positive constant having the dimensions of power and β and I are system parameters, the latter providing a measure of the information communicated over the trajectory The instantaneous transmitter-receiver separation, denoted by $\mathcal{R}(t)$ is, to within the reasonable restrictions upon its mathematical behavior which are always satisfied under the circumstances of interest, entirely general. If the magnitude of the intelligence to be acquired by the vehicle is anticipated or, alternately stated, readout proceeds independently of the acquisition of data, a physically-realizable, conventional system, involving storage and programmed output, is described by the preceding equations.

Employing the method of Lagrange, the extremals of

$$\int_0^T \left\{ p + \lambda\gamma \ln\left(1 + \beta \frac{p}{\mathcal{R}^2}\right) \right\} dt \tag{22}$$

are desired, λ being an undetermined multiplier. Forming the first variation, one is led to the equation

$$1 + \gamma \frac{\lambda\beta/\mathcal{R}^2}{1 + \beta p/\mathcal{R}^2} = 0 \tag{23}$$

or

$$p = -\left(\lambda\gamma + \frac{\mathcal{R}^2}{\beta}\right)$$

whence, from the constraining integral,

$$\lambda = -e^{\frac{I}{\gamma T} - \frac{1}{T}\int_0^T \ln\left(\gamma\beta/\mathcal{R}^2\right) dt} \tag{24}$$

and

$$\mathcal{E} = \gamma T\, e^{\frac{I}{\gamma T} - \frac{1}{T}\int_0^T \ln\left(\gamma\beta/\mathcal{R}^2\right) dt} - \frac{1}{\beta}\int_0^T \mathcal{R}^2 dt . \tag{25}$$

Since λ is negative, if it exists the second variation of (22),

$$-\lambda\beta^2\gamma \int_0^T \left(\mathcal{R}^2 + \beta p\right)^{-2} dt \tag{26}$$

is positive, so that the extremal is a minimum. If, in particular, the information received is essentially proportional to the mission duration, a condition which generally obtains, and, as with the relative trajectories discussed herein, for large T the ratio of T and the integral of $\ln \mathcal{R}(t)$ vanishes or oscillates within narrow bounds, the exponential in (25) largely behaves like some constant μ. One then has

$$\mathcal{E} \sim \mu\gamma T - \frac{1}{\beta}\int_0^T \mathcal{R}^2(t) \tag{27}$$

or, the minimal energy consumption requirement of a system employing storage and optimal readout of the latter, corresponds to the difference of that of a constant power output system and a constant signal power density system.

4. Principal Conclusions

a. Let a vehicle be projected outward from the earth on an idealized rectilinear path to a designated maximum separation point. Then, if the power transmission program is such as to maintain a fixed, minimal signal power density at the terrestrial receiver, rather than maintaining a constant power

output out to maximum separation that is consistent with the minimal signal amplitude requirement, the energy consumed by the continuously variable transmitter program rapidly converges to 5/8 of that which is required by the fixed power program as the maximum separation tends to become indefinitely large.

b. If the continuously variable power transmission program is replaced by a two-stage system utilizing fixed levels of power output in each stage, to a good order of approximation the optimum transition point from low to high power outputs is at $\sqrt{(3/7)}$ of the maximum transmitter-receiver separation. The energy consumption is then $1-(4/7)(3/7)^{3/4}$ that of a single stage or completely fixed power output system, and provides a close approximation to the economies yielded by the idealized continuously variable, or infinite-multiple-stage system.

c. Let the astronautical vehicle be assumed to pursue an orbit concentric with the sun and in the plane of the ecliptic, with orbital radius r_a in astronomical units. Then, the energy consumption required by the transmitter if continuous variation of power output is used to maintain a minimal admissible signal amplitude at the terrestrial receiver is $(1+r_a^2)/(1+r_a)^2$ of that which is needed if the power output is fixed at the level indicated at maximum separation at superior conjunction.

d. Replacing the continuously variable system in the above with two fixed transmitter stages, the resultant economies are seen to be essentially half of those yielded by the idealized, infinitely multiple stage system. It is shown that in passage from inferior to superior conjunction, the optimal transition between the two stages is at CT and, in passage from superior to inferior conjunction, at $(1-C)T$, where $C=0.48$ and T denotes the time interval involved in the change in their relative configuration with respect to the sun. For the former, the variation of the optimum separation at transition is a strong function of the orbital radius, and largely independent of it for the latter case.

e. Let the relative transmitter-receiver motion be unrestricted in an astronautical vehicle which has storage and optimal readout available to its communication system. Then, if the totality of information to be communicated to the terrestrial receiver is essentially proportional to the mission duration, which is assumed to be long, the minimal energy consumption corresponds to the difference of those of a constant power output system and one programmed to yield a fixed signal power density at the receiver.

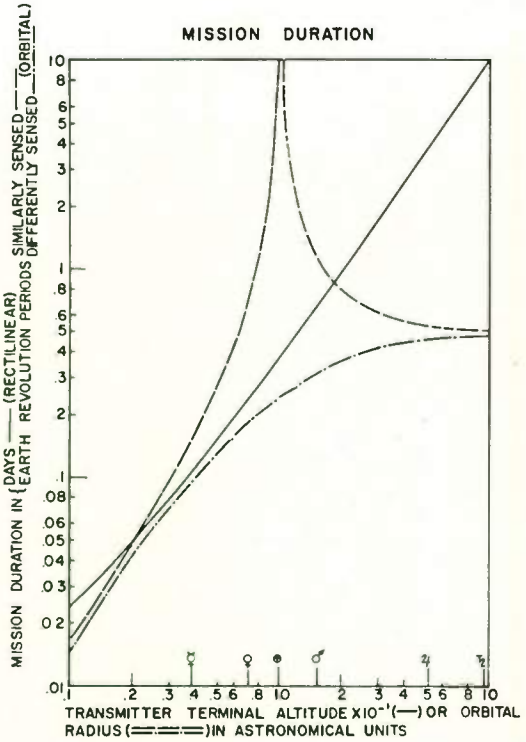


Fig. 1.

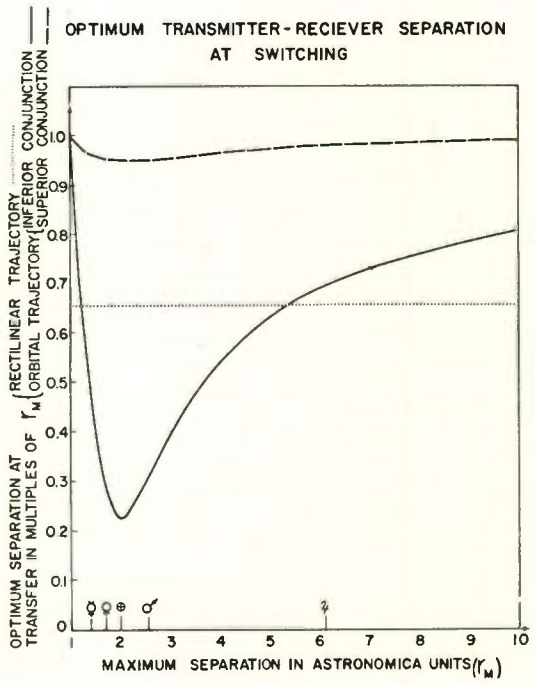


Fig. 2.

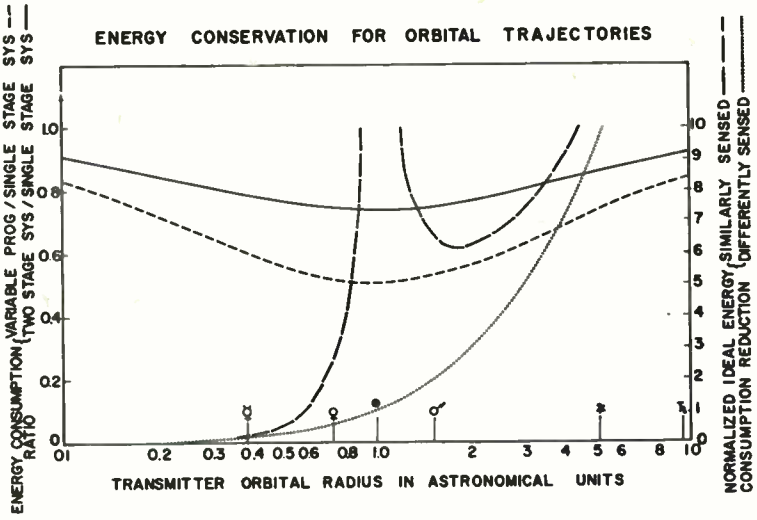

ENERGY CONSERVATION FOR ORBITAL TRAJECTORIES
ENERGY CONSUMPTION RATIO (VARIABLE PROG / SINGLE STAGE SYS — (TWO STAGE SYS / SINGLE STAGE SYS ——
NORMALIZED IDEAL ENERGY CONSUMPTION REDUCTION (SIMILARLY SENSED —— —— (DIFFERENTLY SENSED ——————
TRANSMITTER ORBITAL RADIUS IN ASTRONOMICAL UNITS
0.1
0.2
0.3
0.4
0.5
0.6
0.8
1.0
2
3
4
5
6
8
10

Fig. 3.

FEASIBILITY STUDY OF CHAFF COMMUNICATION*

L. H. Bauer
Radiation Incorporated
Melbourne, Florida

C. E. Sharp
U. S. Army Signal Research and Development Laboratory
Fort Monmouth, N. J.

R. Herring
U. S. Army Signal Research and Development Laboratory
Fort Monmouth, N. J.

Summary

There has been recent interest in the use of re-radiation from chaff as a means of communication over non-line-of-sight paths at UHF. Recent experimental investigations have indicated the feasibility of the use of chaff for this purpose. These investigations were undertaken by the U. S. Army Signal Research and Development Laboratory, Fort Monmouth, New Jersey, to determine the characteristics of chaff clouds which might influence the design of equipment for tactical communications systems.

Tests were made using military radio relay equipment. In addition, radar equipments were utilized to determine chaff cloud position and size. The chaff was dispensed at an altitude of 10,000 feet by aircraft. Various chaff payloads were used. Instrumentation included chart recorders and radar displays. Information on chaff cloud characteristics included bistatic cross-section, disperson, full rate, and drift, polarization characteristics and fading rates.

Chaff fall rates varied from 100 to 200 feet per minute and increased with increasing altitude. Horizontal and vertical, cloud dimensions in the range of 2000 to 3000 yards were measured. Drift rates were those of the horizontal component of wind velocity.

* Research sponsored by U. S. Army Signal Research and Development Laboratory, Fort Monmouth, N. J. under contract DA36-039 SC-88897

The median bistatic cross section (σ_{1-2}) was determined as a function of time and was 2000 to 4000 square meters for a chaff unit containing 3.75 (10^6) dipoles. The bistatic cross section was proportional to the number of dipoles. The chaff cloud was found to be sensitive to the polarization of the transmitting and receiving antennas. The spectrum of fading frequencies varied from 0 to 10 cps. The probability distribution of the received signal power was exponential.

INTRODUCTION

Chaff has been used extensively as a countermeasure against both tracking and search radars. Recently, there has arisen much interest in the use of chaff as a communication media. The use of chaff appears particularly attractive over obstructed transmission paths, when the use of natural propagation modes such as troposcatter or line-of-sight are severely limited.

Much is known about the monostatic or back scatter characteristics of chaff. In communication links, the bistatic scattering cross section is the parameter of interest, but it has not been measured extensively. Some analytical work on scattering from chaff dipoles has been done by Hessemer[1]. His results predict the bistatic cross section to be a function of antenna elevation angles and a statistical description of the orientation of the dipoles in the chaff cloud. Blom[2], using Hessemer's results, has discussed the technique of chaff communications

and compared it with tropospheric scatter communication. In 1961 preliminary experiments on the use of chaff for communication were performed by the U. S. Army Signal Research and Development Laboratory at Fort Monmouth, N. J.

The purpose of the present effort is to determine the feasibility of chaff communications in the 1000 to 3000 mcps portion of the frequency spectrum. Measured data are being obtained on the following:

1. Bistatic cross section (σ_{1-2})
2. Directivity
3. Polarization behavior
4. Information bandwidth
5. Effects of meteorological environment.

Phase I effort was concerned with short range (4 to 6 mile site separation) tests with chaff dispensed from aircraft at an altitude of 10,000 feet. Phase II tests will utilize site separations of 50 to 200 miles. The results reported in this paper are those of Phase I.

The remainder of this report presents a discussion of chaff scatter geometry and the equipment used to collect data. This is followed by results obtained from the radar and radio data. Next chaff communication and tropospheric scatter are compared on the basis of median basic transmission loss. The paper is concluded with a statement of significant results obtained to date and the objectives of future effort.

Chaff Communication Geometry

A chaff communication system consists of two terminals and a cloud of a large number of chaff dipoles located near either terminal as shown in Figure 1. The chaff is dispensed at an altitude which will insure a clear line-of-sight from the cloud to both sites for the desired period of communication. The chaff cloud is illuminated by the transmitter, energy is extracted from the incident wave and re-radiated in all directions. Re-radiated energy from dipoles in the receiving antenna beam is collected and amplified in the receiver. Although the contribution from a single chaff element is small, the resultant signal from a large number is significant.

The slant ranges from the transmitter and receiver to the cloud are designated as R_1 and R_2 respectively. Using the geometry shown, the bistatic cross section σ_{1-2} is expressed by Equation (1) of Figure (1) in terms of known or measurable quantities. Since the product $(R_1 R_2)^2$ is a minimum when the cloud is over either terminal, it is desirable to dispense chaff as near the sites as possible.

Equipment

Radar

An M-33C radar system consisting of an x-band conical scan tracking radar and an s-band search radar provided cloud position, pointing angles for the transmitting antenna, and information on cloud size and shape. Cloud position was obtained from the tracking radar in the form of ground track and altitude plots. For the ground track, ground range was computed from the slant range and angle data and plotted against azimuth. Altitude was computed and plotted as a function of slant range. These plots provided the data necessary to compute R_1 and R_2 of Equation (1).

Information on cloud size and shape was supplied by both the track and search radars. Azimuth and range information were taken from photographs of an expanded portion of the search radar plan position indicator, 27 degrees in azimuth and 5000 yards in range. Vertical dimensions were obtained by scanning the tracking radar antenna in elevation and determining the upper and lower limits of the cloud.

Radio

A military radio relay system, the AN/GRC-50, was used to supply data on the bistatic properties of the chaff cloud. Parameters of the system are:

Transmitted power = 10 watts

Operating frequency = 1498 mcps

Antenna gain (both transmitter and receiver) = 16 db

Antenna beamwidth (3 db points) = 25 degrees

Predetection bandwidth	= 1.5 mcps
Postdetection bandwidth	= 60 cps
Minimum detectable signal	= -140 dbw

The transmitter was located at the radar site with antenna pointing information supplied by the tracking radar. Transmissions were c-w with output power of 10 watts ± 1 db. The fm receiver, for these experiments, was modified slightly. An amplitude detector was used to provide an output voltage proportional to receiver input power. This voltage was recorded on a paper chart recorder whose frequency response was 0 to 60 cps. The receiver-recorder system was calibrated using a precision attenuator and signal generator. The calibration was accurate to within ± 2 db.

Chaff

Foil chaff having the following specifications was used:

Material	0.00045" HA Ridaram Coated
Dipole dimensions	0.016" x 3.7"
Frequency	1498 mcps
Unit weight	10.0 pounds
Dipoles per unit	3.75 (10^6)

Except for special experiments 10 pound chaff units were dispensed. Drop altitude for all tests was 10,000 feet. An aircraft was used to transport the chaff.

Experimental Results

Phase I experiments were conducted in the Melbourne, Florida area during October 1961. A total of 15 chaff drops were made during this phase. Two drops have been selected to show bistatic cross section behavior, polarization effects, dispersion and the relationship between $\sigma_{1\text{-}2}$ and the number of dipoles. The geometry for Phase I is shown in Figure 2. The radar and transmitter were located at the same site with the receiver approximately 4.4 miles to the west. Also shown on the figure are the cloud altitude and ground tracks scaled from the radar plots. These data are typical of cloud trajectories for wind speeds of 4 to 8 miles per hour.

Radar Results

Fall rate, cloud size and drift rates are important factors in the design of a chaff communications system. All three affect the length of time communications can be maintained. Chaff which has a high fall rate will reach the ground soon after being dispensed and will result in a short communication time or require dispensing at high altitudes to insure adequate time for transmission of information. Cloud size can limit the useful transmission time if the cloud becomes so large that all dipoles are not encompassed by the transmitting and receiving antenna beams thereby reducing the bistatic cross section and the received signal. Horizontal drift of the cloud could result in a decrease in received signal due to an increase in the $(R_1R_2)^2$ also limiting communication time.

Fall Rate

The fall rate data of seven drops have been plotted in Figure 3 as a function of altitude. The fall rate for the foil chaff described above decreased from approximately 200 to 100 feet per minute for a decrease in altitude from 10,000 to 5,000 feet. This behavior is explained by the increased air density at lower altitudes and thermal updrafts near the earth's surface.

Cloud Size

Cloud size and shape are determined by wind shear, turbulence and gravity. Typical cloud dimensions for 3.75 (10^6) dipoles are shown in the sequence of sketches drawn from photographs of the expanded scope of the s-band search radar. For the 14 minutes covered by these pictures the cloud dimensions increased by a factor of 1.25 and 2.4 in range and azimuth respectively. Vertical dimensions were found to be of like magnitude.

Drift Rate

Horizontal movement of the chaff cloud is that of prevailing winds. The ground track shown in Figure 2 depicts cloud drift and for this drop was approximately 4 miles per hour which agrees well with wind speeds obtained from balloon sounding. As an example of the change which may be encountered in signal level, $(R_1R_2)^2$ decreased 7 db during the thirty minutes this cloud was tracked.

Radio Results

The 15 second sample recording of Figure 5

is typical of signals received from chaff clouds. The signals exhibit the rapid fluctuation in signal level, characteristic of scatter propagation. The median level was used as a measure of the received signal. The median level was determined as a function of time for the drop. (The intervals for which the median signal was determined varied from 15 seconds to one minute.) Next the median signal was corrected for transmitting and receiving antenna pointing errors. The median bistatic cross section ($\overline{\sigma_{1-2}}$) was then computed using Equation (1); where values of R_1 and R_2 were obtained from radar data.

Bistatic Cross Section

Figure 6 is a graph of the median bistatic cross section $\overline{\sigma_{1-2}}$, as a function of time for a chaff drop of 3.75 (10^6) dipoles. The curve for 0 degrees polarization is typical for all drops. $\overline{\sigma_{1-2}}$ rapidly increased during the first minute and then increased at a slower rate until four minutes when it levels off for the remainder of the time. The initial rapid increase in $\overline{\sigma_{1-2}}$ is caused by blooming of the chaff. The more gradual increase is believed due to dipoles which break loose from clumps of dipoles as they fall toward the ground. Once the dipoles have dispersed, $\overline{\sigma_{1-2}}$ remains essentially constant. The range of median $\overline{\sigma_{1-2}}$ for 3.75 (10^6) dipoles was approximately 2000 to 4000 square meters.

Polarization Effects

If chaff dipoles are randomly oriented, then the bistatic cross section would be insensitive to variation in antenna polarization. However, should there be a preferred dipole orientation, $\overline{\sigma_{1-2}}$ would vary with changes in polarization. Polarization measurements were made on one drop to determine if the cloud was polarization sensitive. The transmitting antenna was aligned with the E-vector of the radiated field parallel to the ground (designated as 0 degrees polarization) and remain fixed during the drop. At the receiver site the antenna was alternately operated with the same alignment as the transmitter or rotated 90 degrees from this alignment (designated as 90 degrees polarization). The results are shown in Figure 6.

The upper curve is for 90 degrees polarization and the lower for 0 degrees. The dashed portion of the 90 degree curve is an estimate of the growth of $\overline{\sigma_{1-2}}$ since no data were collected at this polarization prior to 4 minutes. Maximum median $\overline{\sigma_{1-2}}$ was 1800 and 3500 square meters for 0 and 90 degree polarization, a difference of 2.5 db. It is obvious that the dipoles of this cloud were not completely randomly oriented.

Number of Dipoles

The received signal from a chaff cloud is a function of the number of dipoles (N) contained in the cloud. That is:

$$S \propto \overline{\sigma_{1-2}} \propto N$$

Where: S = median received signal

$\overline{\sigma_{1-2}}$ = median bistatic cross section

Thus, if the number of dipoles is varied, $\overline{\sigma_{1-2}}$ and S should change by the same ratio. This relationship was tested by dispensing 3.75 (10^6) and 0.625 (10^6) dipoles on two drops. The ratio of the number of dipoles for the two chaff units is 6 or 7.8 db. Figure 7 shows the measured median $\overline{\sigma_{1-2}}$ for the two drops. Since only 90 degrees polarization was employed for the 0.625 (10^6) dipoles drop, the comparison is made for this polarization. The ratio of the maximum $\overline{\sigma_{1-2}}$ is 5.8 or 7.7 db. Good agreement between the ratios of measured $\overline{\sigma_{1-2}}$ and the number of dipoles verifies the relationship of Equation (2).

Distribution of Signal Fluctuations

According to theory[3] the probability distribution in power of a signal scattered from a large number of randomly moving dipoles should be exponential. The distribution of signal fluctuations for the bistatic case were determined for a number of samples of which the 25 and 15 second samples shown in Figure 8 are typical. The average power (P_o) of the histograms were computed. Then, using the measured average power an exponential curve was fitted to the histograms as shown by the smooth curve.

As shown in the figure, the measured data is quite close to the theoretical distribution. To further test the measured data, the median signal level was computed from the measured average power assuming an exponential distribution. This value was then compared to the measured median signal level. For the sample of 8A the computed and measured median signal levels were the same; 1.46 $\mu\mu$ watts. The sample of 8B yielded computed and measured medians of 0.7 and 0.8 $\mu\mu$ watts. This good

agreement between calculated and measured median signal level was characteristic of all samples studied.

Spectrum of Fading Frequencies

The rate of fluctuation (fading frequencies) of a scattered signal depends upon the relative motion of the scatterers. These fluctuations arise from Doppler beats of the moving dipoles and should be a function of the velocity of the dipoles. Figure 9 shows two typical spectra for samples of 15 and 25 seconds. Measured wind speeds for the two cases differed by a factor of approximately 2 and were 8 to 12 mph and 4 to 8 mph respectively for the samples. The curves of Figure 9 show that the width of the spectrum for the lower speeds is narrower than that of the higher wind speeds. The power spectrum for Curve A is one twentieth of its zero frequency value at approximately 9 cps whereas for Curve B the frequency is 6 cps at one twentieth of the zero frequency value. For all chaff drops made to date wind speeds have been between 5 and 20 mph and the spectra of fading frequencies from 0 to 15 cps.

Bandwidth

An important characteristic of a chaff transmission link is the information bandwidth which the cloud will support. Because of the time delay (path length difference) between different paths through the cloud, multipath interference will limit the coherent bandwidth. Bandwidth measurements were not included in the first series of tests; however, the transmitter was voice modulated and transmission made to the receiving site. Intelligibility was good for median signal-to-noise ratio of 10 db and was excellent for ratios of 15 to 20 db.

Chaff Compared to Tropospheric Scatter

Over distances of 50 to 200 miles, it is of interest to compare chaff communications and tropospheric scatter. To compare the two modes the median basic transmission loss of both were computed at 1500 mcps and plotted as a function of distance as shown in Figure 10. The data for tropospheric scatter assumes smooth earth and no obstacles on line-of-sight. The median basic transmission loss for chaff transmission was based on a median value of $\bar{\sigma}_{1\text{-}2}$ of 3500 square meters obtained for 3.75 (10^6) dipoles and 90 degree polarization (See Figure 6). Chaff drop point is assumed to be directly over either terminal at an altitude of $D^2/2 + 6000$ feet where D is the distance between sites. This altitude is 6000 feet above the line-of-sight to the far terminal and provides a transmission time of 30 minutes for a chaff fall rate of 200 feet per minute.

At 50 miles the transmission loss for both tropospheric and chaff scatter are approximately equal. However, at 80 miles chaff communication has a 10 db advantage which increases to 20 db at 200 miles. This decrease in transmission loss would produce significant savings in transmitted power or antenna gain.

An important feature of this comparison is the modest altitude at which chaff is dispensed. For a distance of 200 miles, an altitude of only 26,000 feet is required to provide 30 minutes transmission time. A small inexpensive rocket employing a simple guidance technique can dispense chaff at this altitude.

Conclusion

The experimental effort reported in this paper has demonstrated the feasibility of communications using a cloud of chaff dipoles as the scattering media. For distances of 80 to 200 miles and 3.75×10^6 dipoles chaff scatter can have a 10 to 20 db advantage over tropospheric scatter. Chaff signals exhibit fading characteristics of other scattering propagation modes; i.e., exponential distribution in power and fast fading frequencies in the range of 0 to 15 cps. It was determined that the bistatic cross section is proportional to the number of dipoles. Polarization effects were noted on one observation by a variation of 2.5 db in bistatic cross section when the receiving antenna polarization was varied 90 degrees (transmitting antenna polarization fixed). Chaff fall rates of 100 to 200 feet per minute were measured as were cloud dimensions of 2000 to 3000 yards.

Additional tests are scheduled for the near future for a site separation of approximately 54 miles. Experiments are planned where both transmitting and receiving antennas will be synchronously rotated in polarization. The important parameter of cloud information bandwidth will be measured and compared to cloud size.

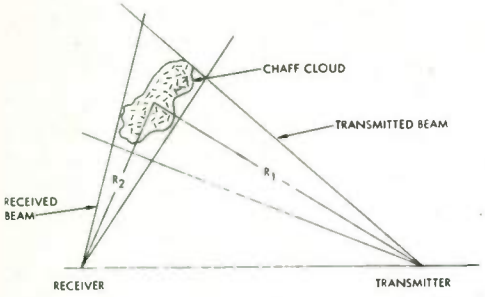
References

1. Hessemer, R. A., Jr., Scatter Communication with Radar Chaff, Vol. AP-9

No. 2, March 1961, p. 211-217.

2. Blom, B. V., Communications by Re-Radiation From Chaff, 1960 IRE Conference Proceeding on Military Electronics, p. 542-546.

3. Kerr, Propagation of Short Radio Waves, McGraw-Hill, p. 553-587.



$$\sigma_{1-2} = \frac{64\,\pi^3\,P_R\,(R_1 R_2)^2}{P_T\,G_T\,G_R\,\lambda^2\,K} \qquad (1)$$

WHERE

σ_{1-2}	=	BISTATIC CROSS SECTION
P_T	=	TRANSMITTER POWER
P_R	=	RECEIVED POWER
G_T	=	TRANSMITTER ANTENNA GAIN
G_R	=	RECEIVER ANTENNA GAIN
λ	=	WAVELENGTH
K	=	SYSTEM LOSSES
R_1	=	TRANSMITTER TO CLOUD RANGE
R_2	=	CLOUD TO RECEIVER RANGE

Fig. 1. Chaff scatter geometry.

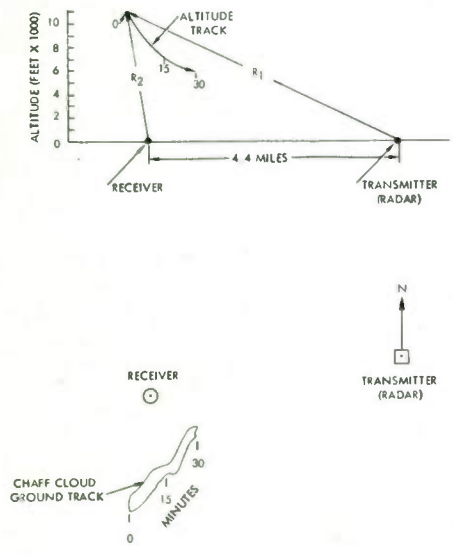


Fig. 2. Path profile and path geometry.

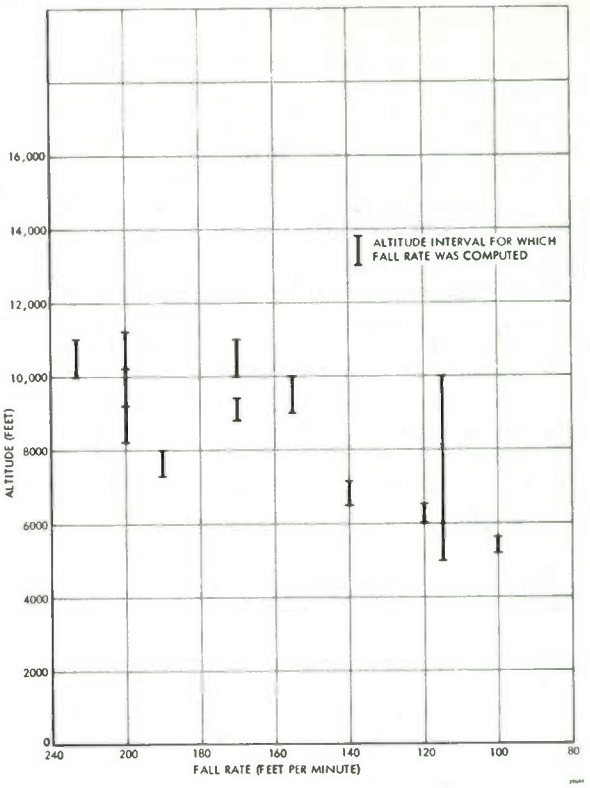


Fig. 3. Chaff fall rate as a function of altitude.

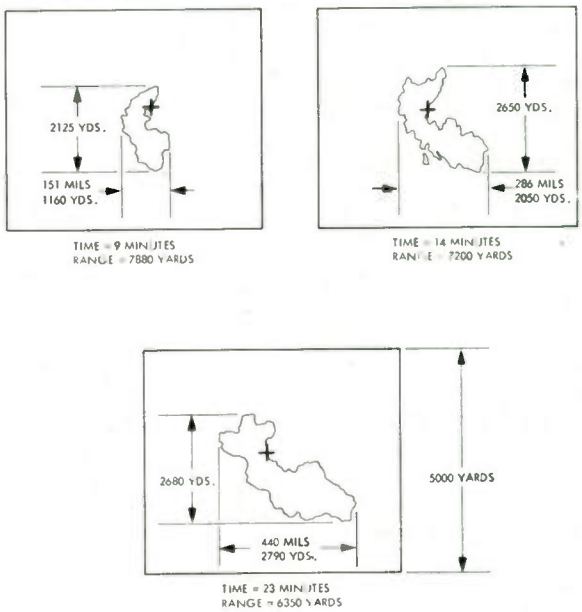


Fig. 4. Cloud size and shape in range and azimuth.

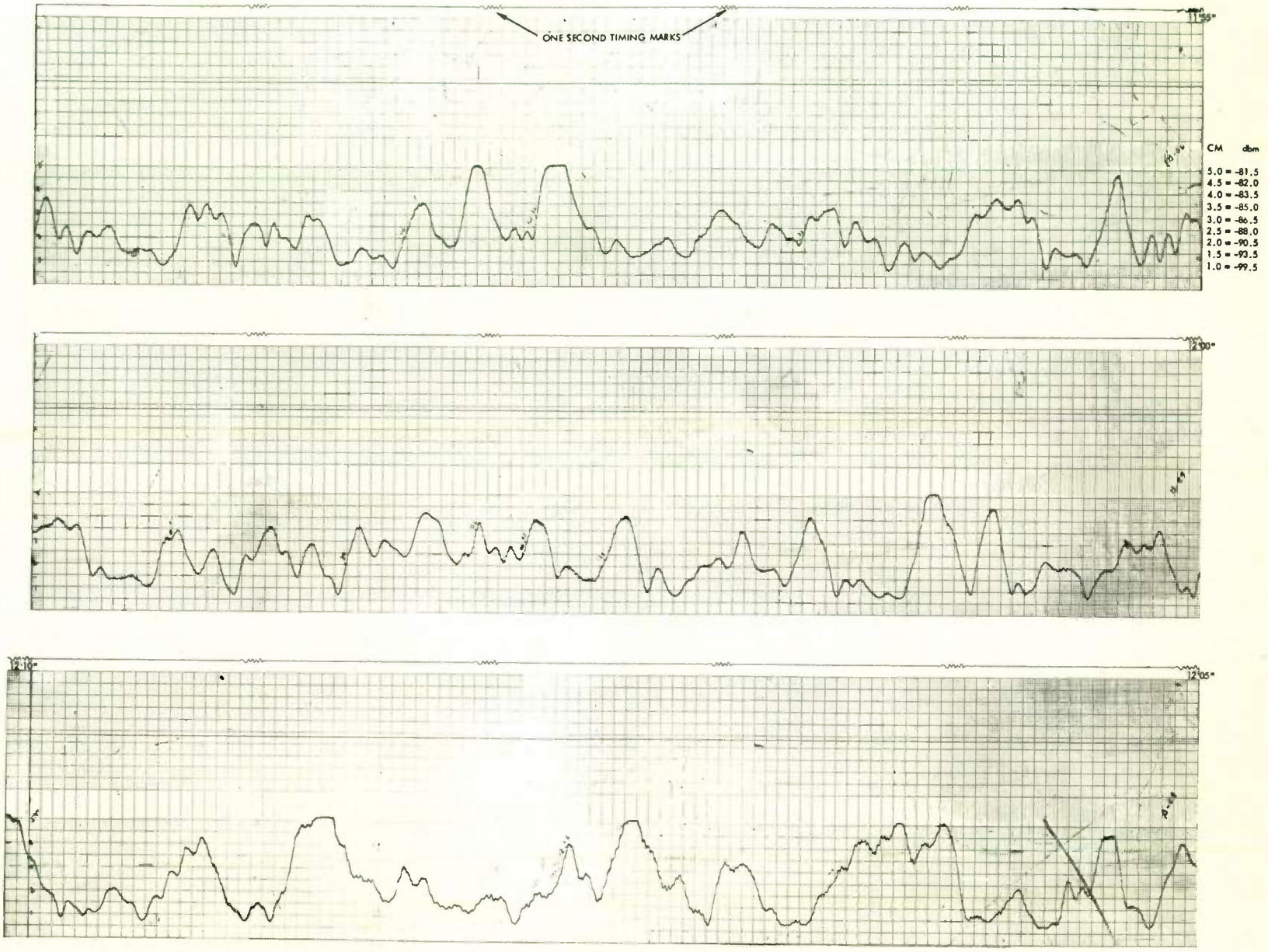


Fig. 5. Sample recording of AN/GRC-50 receiver output.

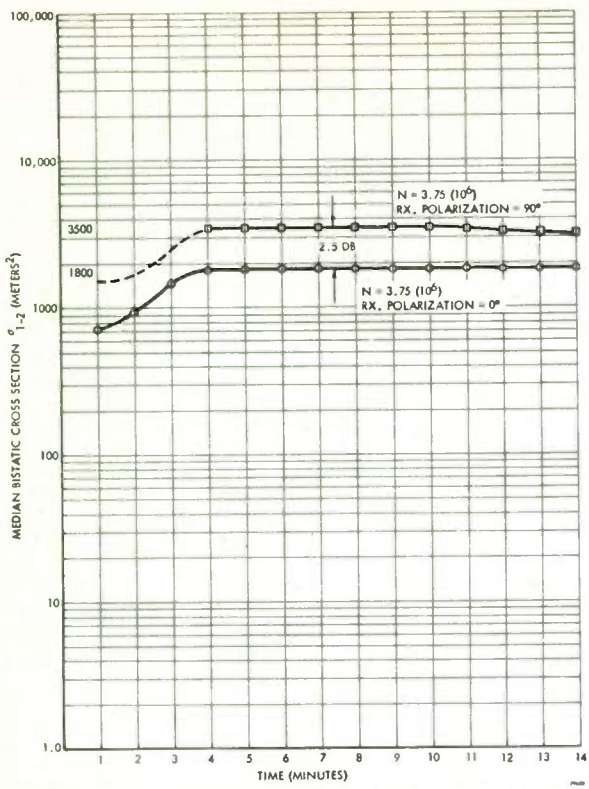


Fig. 6. Bistatic cross section for two polarization settings of the receiving antenna.

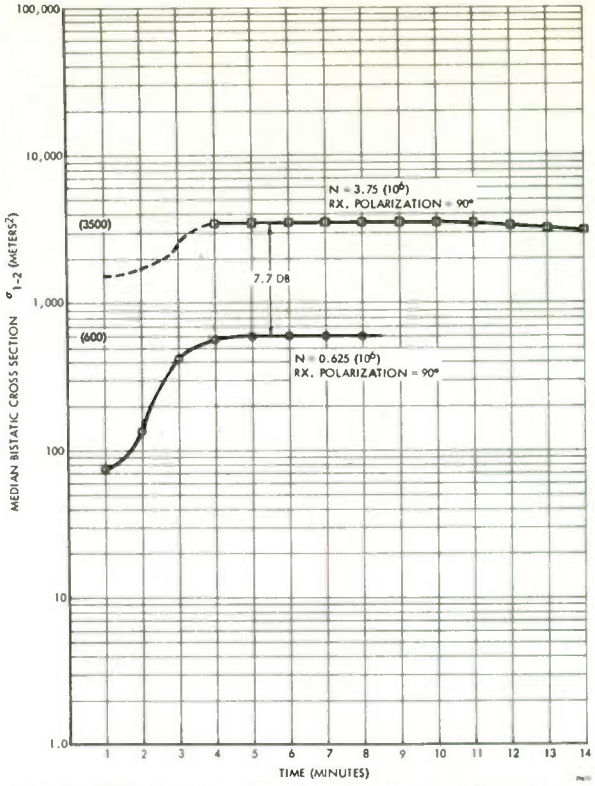


Fig. 7. Bistatic cross section for 3.75 (10^6) and 0.625 (10^6) dipoles.

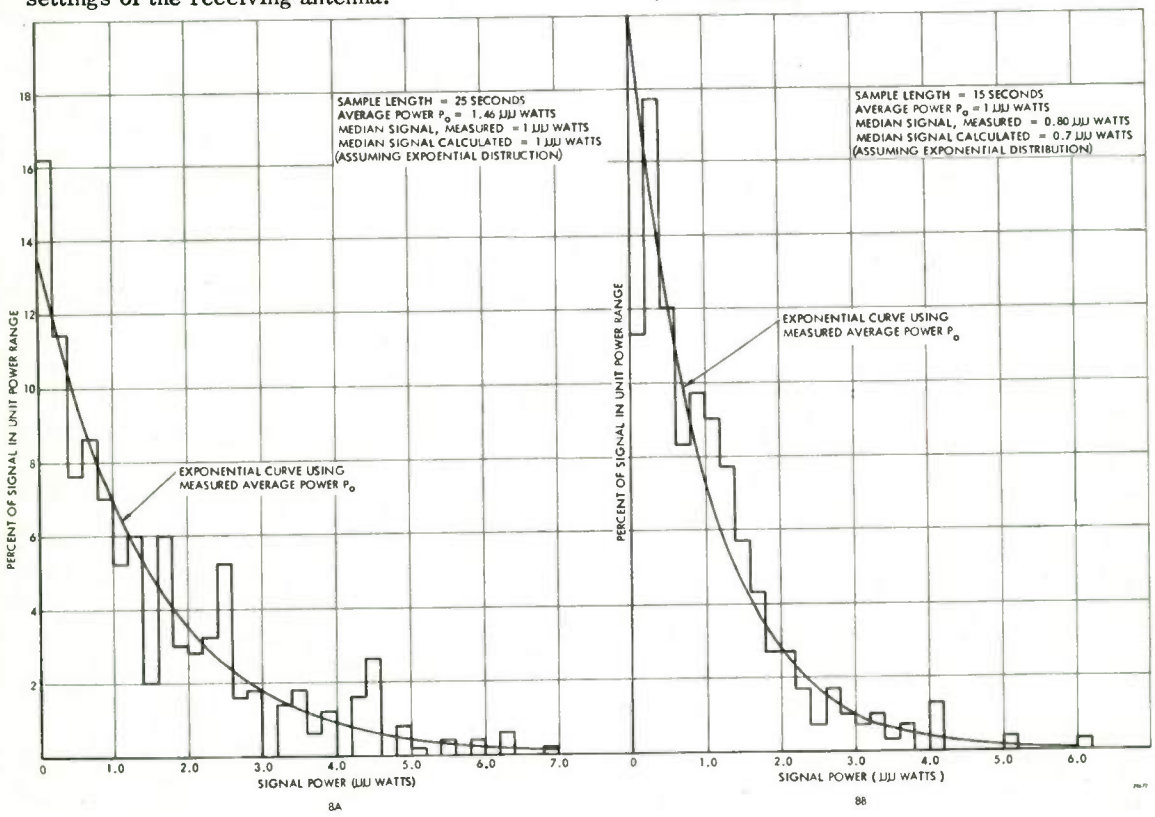


Fig. 8. Typical probability distributions in power for chaff scattered signals.

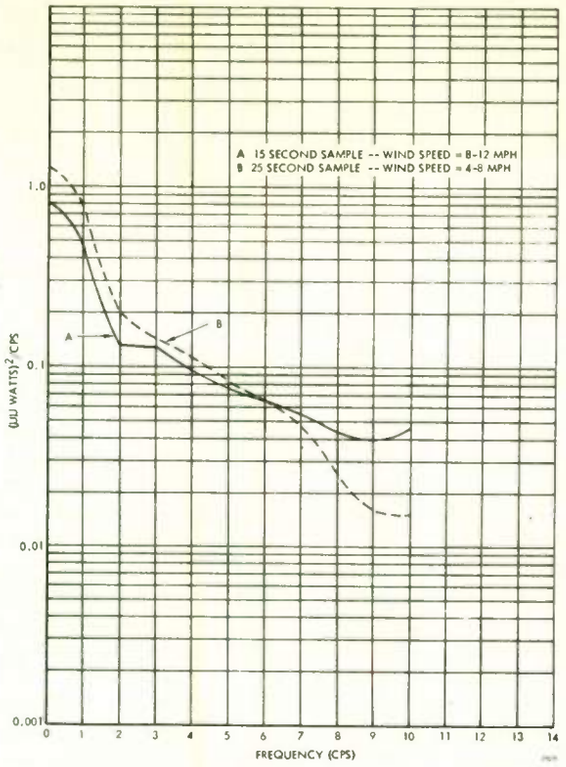


Fig. 9. Typical power frequency spectra for fluctuations of chaff scattered signals.

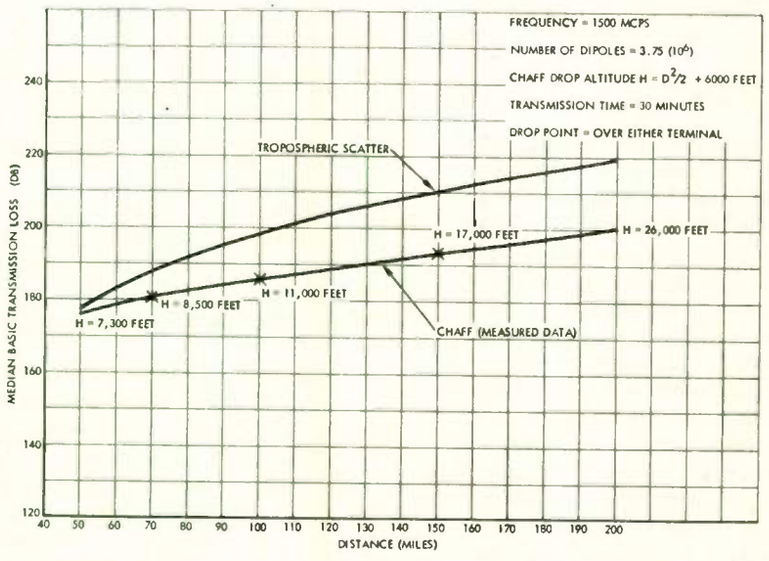


Fig. 10. Median basic transmission loss for tropospheric and chaff scatter systems.

PREDETECTION DIVERSITY COMBINING WITH SELECTIVELY FADING CHANNELS

Phillip Bello and Bert D. Nelin
ADCOM, Inc.
Cambridge, Mass.

Summary. The effective use of predetection (or i.f.) linear combining of diversity channels requires that all frequency components be added in-phase. In view of the time-variant phase characteristics of fading channels, some form of instantaneous phase correction must be provided prior to the diversity combining so that proper in-phase addition will result. If the instantaneous fading is not uniform over the width of the channel, the resulting channel incoherence will reduce the efficacy of the phase correction. One way to avoid this difficulty is to divide the selectively fading channel into "flatly" fading subchannels and perform the necessary phasing and combining operations per subchannel. The question arises as to how much frequency selective fading can be tolerated within a channel, or a subchannel, before serious deterioration in performance results. This paper considers this question with reference to FDM-SSB binary data transmission and matched filter reception, wherein a pilot tone is used both for deriving proper phase correction and for providing the weighting necessary for maximal ratio combining. Analytical expressions are derived for error probabilities which show the deleterious effects of frequency selectivity and pilot tone additive noise.

Introduction

The method of coherent combining to be studied is indicated in the simplified receiver block diagram shown in Fig. 1, wherein the predetection portion of only one receiver is shown. As mentioned above, a pilot tone is transmitted with the frequency division multiplexed channel signals to provide a phase reference for coherent combining. The optimum place to locate this pilot tone is as near as possible to the center of the set of multiplexed signals. In this way the worst phase incoherence relative to the pilot tone is that for either outside channel, both of which are separated from the pilot tone by half the total multiplex signal bandwidth. Clearly, any other pilot tone location would produce an outside channel with a greater frequency separation and thus (generally) a greater phase incoherence relative to the pilot tone. For purposes of analysis we will confine our attention to the determination of the error probabilities associated with a multiplex channel which is spaced F cycles/sec. from the pilot tone. Consequently the results of the analysis can be applied to any data channel and any pilot tone location.

Although we allow selective fading between frequency components spaced F cycles/sec. apart we will assume that the individual data channels occupy a small enough bandwidth so that flat fading may be assumed for the frequency components within a single data channel. In such a case the frequency selectivity may be characterized in terms of an appropriate correlation coefficient between the received pilot tone and a hypothetical received tone which would be transmitted at the center of the data channel, F cycles/ sec. away.

In the receiver of Fig. 1 the pilot tone is extracted from the multiplexed process by a (pilot tone) filter after appropriate RF and IF filtering. It is assumed that signal detection will take place with the multiplexed band centered at some convenient offset frequency, say f_0 cycles/sec. To accomplish this centering and the necessary phase correction, the filtered pilot tone, located at f_1 cycles/sec. (the i.f. frequency), is mixed with a local oscillator at f_0 cycles/sec. and the difference frequency component is mixed with the multiplexed process. Extraction of the resulting difference frequency component then yields the phase corrected multiplex process centered at f_0 cycles/sec. The diversity combining method illustrated in Fig. 1 will provide exact maximal ratio combining when the frequency selective fading and the pilot tone noise-to-signal ratio are negligible, and when the additive noise levels are the same for all diversity channels.

In our analysis we will consider three forms of matched filter detection; two forms of coherent detection and non-coherent detection. The two forms of coherent detection are distinguished by the fact that one method (which provides only approximate coherent detection) uses a fixed phase reference adjusted to coherently detect the non fading component in the diversity combined signal while the other (which provides exact coherent detection) uses a time variable phase reference adjusted to compensate for the instantaneous phase fluctuations in the diversity

combined signal.

Formulation of Predetection Operations

In the discussion to follow we will use complex notation for signals, so as to present our derivations in the most efficient manner. As discussed in the introduction we desire to determine the error probabilities associated with a data channel spaced F cycles/sec from the pilot tone. The complex transmitted signal corresponding to this data channel is represented in Fig. 2 as

$$S(t)e^{j2\pi(f_c + F)t} \tag{1}$$

where S(t), the "complex envelope" of the data signal is given by

$$S(t) = \sum_{x} x_k(t - kT) \tag{2}$$

in which

$$x_k(t) = \begin{cases} S_1(t)\ ;\ \text{"1" transmitted in interval} \\ kT < t < (k+1)\,T \\ \\ S_0(t)\ ;\ \text{"0" transmitted in interval} \\ kT < t < (k+1)\,T \end{cases} \tag{3}$$

and T is the reciprocal of the data rate.

The transmitted pilot tone, centered at f_c cycles/sec. is represented in Fig. 2 as $\exp\left[j2\pi f_c t\right]$. We have chosen to normalize the pilot tone power at the outset so as not to carry unnecessary variables through the derivations. The pilot tone and data signals are shown transmitted through a selectively fading medium. Due to the fading a transmitted sinusoid is received as a narrow band process whose amplitude and phase are slowly varying. It will be assumed that no specular component exists and that $g_1(t)$ the complex envelope of the resulting narrow band process is a complex valued normally distributed process. Thus the received pilot tone may be represented as

$$g_1(t)e^{j2\pi f_c t} \tag{4}$$

where $g_1(t)$ is a complex valued normally distributed process, whose magnitude is the conventional envelope of the received pilot tone and whose angle is the phase of the pilot tone relative to $2\pi f_c t$. If the medium were flat-fading, a change in f_c, the frequency of the pilot tone, would leave the received complex envelope unchanged except for a possible time invariant amplitude change and a phase shift dependent upon f_c only. Such changes in two complex processes do not affect the magnitude of their correlation coefficient and as a result a flat fading medium produces unity magnitude for the correlation coefficient between the received complex envelope of sinusoids transmitted at different frequencies.

For a selective fading medium, the magnitude of the correlation coefficient between the received complex envelopes of pilot tones located at different frequencies is no longer unity and in fact drops to zero as the frequency separation between the pilot tones increases. Thus, if a pilot tone were transmitted at the frequency $f_c + F$ cycles/sec. one expects a received process with complex representation

$$h_1(t)e^{j\left[2\pi(f_c + F)t + \theta\right]} \tag{5}$$

where θ is a constant phase shift and $h_1(t)$ is a complex valued normally distributed process which is somewhat decorrelated with respect to $g_1(t)$.

As already mentioned it is assumed that S(t) occupies a narrow enough band of frequencies about the frequency $f_c + F$ cycles/ sec so that across this small band at least the fading may be assumed flat. In this case the received signal may be represented as†

$$h_1(t)S(t)e^{j\left[2\pi(f_c + F)t + \theta\right]} \tag{6}$$

An additive noise disturbance $\eta_1(t)e^{j2\pi f_c t}$ is shown combined with the output of the selectively fading medium, yielding as input to the receiver the process z(t) given by

† Strictly speaking S(t) in (6) should be replaced by $S(t-\tau_o)$ where τ_o is some mean path delay. However, as may readily be verified, the results of our analysis will be unchanged if we make the convenient assumption $\tau_o = 0$.

$$z(t) = g_1(t)e^{j2\pi f_c t} + h_1(t)S(t)e^{j\left[2\pi(f_c+F)t+\theta\right]} + \eta_1(t)e^{j2\pi f_c t} \tag{7}$$

The r.f. and i.f. filtering of z(t) is not explicitly indicated in Fig. 2, because this filtering is assumed to pass the signal components of z(t) without distortion. However, this filtering does limit the bandwidth of the input white noise and to account for this, $\eta_1(t)$ is assumed to be band limited white noise. Only a heterodyning operation is shown in Fig. 2 in place of the r.f. and i.f. stages, so as to show a translation of the input process from a center frequency of f_c cycles/sec to one of f_1 cycles/sec.

The pilot tone filter output is given by

$$\left[g_1(t) + \xi_1(t)\right]e^{j2\pi f_1 t} \tag{8}$$

where $\xi_1(t)$ is the complex envelope of the output noise accompanying the extracted pilot tone. Since the bandwidth of the pilot tone filter must be wide enough to pass $g_1(t)e^{j2\pi f_1 t}$ it must be of the order of the fading bandwidth. It follows that $\xi_1(t)$ will have a bandwidth of the same order as that of $g_1(t)$.

Figure 2 shows that the output of the pilot tone filter is multiplied by $e^{-j2\pi f_0 t}$. This has the desired effect of translating the spectrum from a center frequency of f_1 to $(f_1 - f_0)$ cycles/sec. Finally the resulting translated complex process is conjugated (the asterisk below the multiplier symbol indicates that the multiplier input at that point is transformed into its complex conjugate) and multiplied by the output of the IF amplifier. The operation just described is the equivalent in complex notation of the real operation of mixing two narrow band processes and extracting their difference frequency component.

Following the above operations through we find that the complex data signal becomes $w_1(t)\exp\left[j2\pi(f_0 + F)t\right]$ where the complex envelope $w_1(t)$ is given by

$$w_1(t) = \left[g_1^*(t) + \xi_1^*(t)\right]\left[h_1(t)S(t)e^{j\theta} + \eta_1(t)e^{-j2\pi Ft}\right] \tag{9}$$

After combination with the corresponding predetection outputs of the other diversity receivers, the data signal becomes $w(t)\exp\left[j2\pi(f_o + F)t\right]$ where the complex envelope w(t) is given by

$$w(t) = \sum_{k=1}^{L}\left[g_k^*(t) + \xi_k^*(t)\right]\left[h_k(t)S(t)e^{j\theta} + \eta_k(t)e^{-j2\pi Ft}\right] \tag{10}$$

wherein the subscript k has been used to denote functions associated with diversity channel k, and L is the order of the diversity.

It will be assumed that the various channels fade independently but with identical fading statistics. Thus

$$\frac{\overline{h_k g_k^*}}{\sqrt{\overline{|h_k|^2}\ \overline{|g_k|^2}}} = r < 1$$

$$\overline{|h_k|^2} = 2\sigma^2 = \overline{|g_k|^2} \tag{11}$$

Since it is assumed that h_k and g_k are complex-valued normally distributed processes, the moments in (11) are sufficient to specify the individual and joint statistical properties of (h_k, g_k) at the same time instant. As a general rule the correlation coefficient in (11) would be complex. However, we may assume that r in (11) is real and positive, since any angle in the correlation coefficient may be assumed automatically included in the angle θ.

Examination of Eq. (10) shows that the complex envelope of the diversity combined signal may be expressed in the form

$$w(t) = Z(t)S(t) + N(t) \tag{12}$$

where

$$Z(t) = e^{j\theta} \sum_{k=1}^{L} h_k(t) \left[g_k^*(t) + \xi_k^*(t) \right]$$

$$N(t) = e^{-j2\pi Ft} \sum_{k=1}^{L} \eta_k(t) \left[g_k^*(t) + \xi_k^*(t) \right] \qquad (13)$$

Any phase fluctuations in the diversity combined signal caused by the medium appear in (12) is a time varying angle of Z(t). However, because of the diversity action the combined signal has a non-fading component. The complex envelope of this non-fading signal component is just equal to the ensemble average value of Z(t) S(t) over the fading (i.e., considering S(t) deterministic). Thus the non-fading component is given by

$$\overline{Z(t)S(t)} = e^{j\theta} \sum_{k=1}^{L} \overline{h_k g_k^*} \quad S(t)$$

$$= 2rL\sigma^2 e^{j\theta} \quad S(t) \qquad (14)$$

where

$$\overline{h_k \xi_k^*} = 0 \qquad (15)$$

has been used because of the assumed independence and zero mean values of h_k and ξ_k.

In the fixed phase-reference coherent detection method, the detection is adjusted to coherently detect the non-fading component of the combined signal, while in the variable phase-reference coherent detection method coherent detection of the complete combined signal is assumed.

Error Probabilities for Non-Coherent Detection

The error probabilities in the non-coherent detection case may be evaluated in two steps. One may first calculate the error probabilities assuming g_k, h_k, and ξ_k known and then average this result over g_k, h_k, ξ_k (k = 1, ... L). When g_k, h_k, ξ_k are known we can use the well known result that P, the error probability for non-fading incoherent (zero threshold) matched filter detection of orthogonal equiprobable, equal energy, binary signals in white noise is given by

$$P = \frac{1}{2} \exp\left[\frac{E}{4N_o} \right] \qquad (16)$$

where E is the energy (integral of the square) of the typical (real) received binary waveform and N_o is the (two sided) spectral intensity of the (real) white noise.

It may be readily demonstrated that we may use the error probability expression (16) provided the bandwidths of g_k, h_k and ξ_k are very small compared to the bandwidths of the binary signals (i.e. slow fading) and provided the power spectral density of the additive noise term $n_k(t)e^{-j2\pi Ft}$ is flat over the frequency region that S(t) has significant spectral amplitudes. By evaluating the power spectral density of N(t) (with g_k, ξ_k given) one finds that

$$N_o = n_o \sum_{k=1}^{L} |g_k + \xi_k|^2 \qquad (17)$$

where n_o is the (two sided) spectral intensity for the white noise whose complex envelope is $\eta_k(t)$.

The energy of a typical received binary waveform may be expressed as

$$E = E_t |Z|^2 = E_t \left| \sum_{k=1}^{L} h_k^* (g_k + \xi_k) \right|^2 \qquad (18)$$

where E_t is the energy in the corresponding transmitted binary waveform. Thus

$$\frac{E}{4N_o} = \frac{E_t}{4N_o} \frac{\left| \sum_{k=1}^{L} h_k^* (g_k + \xi_k) \right|^2}{\sum_{k=1}^{L} |g_k + \xi_k|^2} \qquad (19)$$

and our error probability is found by averaging P (Eq. 16) with respect to h_k, g_k, ξ_k, k=1, ..., L. To perform this average we first determine the density function of the variable

$$x = \frac{\left| \sum_{k=1}^{L} h_k^* (g_k + \xi_k) \right|^2}{\sum_{k=1}^{L} |g_k + \xi_k|^2} \qquad (20)$$

and then our desired error probability p_L is given by

$$p_L = \frac{1}{2}\int_o^\infty W(x)e^{-sx}\,dx \tag{21}$$

where

$$s = \frac{E_t}{4N_o} \tag{22}$$

and W(x) is the density function of x. It is interesting to note that $2p_L$ is just the Laplace transform of the density function W(x).

In Appendix 1 it is shown that the density function W(x) is given by

$$W(x) = \sum_{j=0}^{L-1} \nu^j(1-\nu)^{L-1-j}C_j^{L-1}\frac{1}{j!}\left(\frac{x}{2\sigma^2}\right)^j\frac{e^{-x/2\sigma^2}}{2\sigma^2} \tag{23}$$

and that F(s), the Laplace Transform of W(x) is given by

$$F(s) = \frac{(s2\sigma^2(1-\nu)+1)^{L-1}}{(s2\sigma^2+1)^L} = \sum_{j=0}^{L-1}\frac{\nu^j(1-\nu)^{L-1-j}C_j^{L-1}}{(s2\sigma^2+1)^{j+1}} \tag{24}$$

where

$$\nu = \frac{r^2}{1+\sigma_p^2/\sigma^2} \tag{25}$$

$$2\sigma_p^2 = \overline{|\xi_k|^2}$$

and C_m^n is the binomial coefficient $\frac{n!}{m/(n-m)!}$

It follows that p, the error probability for non-coherent detection may be expressed as

$$p' = \frac{\left[\frac{(1+1/\rho_p - r^2)}{(1+1/\rho_p)}\cdot\frac{\rho}{2}+1\right]^{L-1}}{2\left(\frac{\rho}{2}+1\right)^L} \tag{26}$$

where we have defined two signal/noise ratios,

$$\rho = \frac{\sigma^2 E}{n_o} = \frac{2\sigma^2 E/T}{2n_o/T}$$

$$= \frac{\text{Signal Power in Data Channel}}{\text{Noise Power in Data Channel}} \tag{27}$$

$$\rho_p = \frac{\sigma^2}{\sigma_p^2} = \frac{\text{Signal Power in Pilot Tone Channel}}{\text{Noise Power in Pilot Tone Channel}}$$

In computing the noise power in the data channel we calculate the noise power in a rectangular bandpass filter of unity gain and bandwidth 1/T (where T is the duration of the binary waveform).

If the signal/noise ratios ρ, ρ_p are high and r is near one, i. e.,

$$\frac{1}{d} = \frac{1}{1-r} >> 1,\ \rho >> 1,\ \rho_p >> 1 \tag{28}$$

where d is defined as the "decorrelation", then the coefficient of ρ in the numerator in (26) is approximately

$$\frac{1+1/\rho_p - r^2}{2(1+1/\rho_p)} \approx d + \frac{1}{2\rho_p} \tag{29}$$

Since the pilot tone signal/noise ratio ρ_p may usually be made large with small cost in transmitter power, one may neglect the $\rho_p/2$ term in (29) and use

$$p_L = \frac{1}{2}\frac{(\rho d+1)^{L-1}}{\left(\frac{\rho}{2}+1\right)^L} \quad ;\ \rho_p >> \frac{1}{d} >> 1 \tag{30}$$

as the expression for error probability for incoherent detection when the decorrelation is small and the pilot tone noise/signal ratio is negligible in comparison to the decorrelation.

Examination of (30) shows that for $\rho d >> L-1$ the error probability as a function of signal/noise ratio takes the asymptotic form

$$p_L \approx \frac{(2d)^{L-1}}{\rho} \quad ;\ \rho d >> L-1 \tag{31}$$

This expression should be compared with the no-diversity expression (L=1) for large signal/noise ratio.

$$p_1 \approx \frac{1}{\rho} \tag{32}$$

and the large signal/noise ratio error probability expression for L^{th} order diversity with zero decorrelation

$$p_L \approx \frac{2^{L-1}}{\rho^L} \tag{33}$$

Examination of Eq's (21)-(33) show that for large signal/noise ratios the presence of a non zero decorrelation causes the asymptotic expression for L^{th} order diversity to have the same behavior as no diversity. This effect may be observed in the graph of Fig. 3 where a family of error probabilitie curves are plotted as a function of signal/noise ratio for no diversity, 2^{nd}, and 4^{th} order diversity with d = 0, .1, .01. Note however, that the signal/noise degradation caused by d = .01 and 2^{nd} order diversity is higher than the degradation caused by d = .01 and 4th order diversity. It appears as a general result that high order diversity provides a beneficial side-effect in counteracting selective fading. This effect is seen most clearly in Fig. 4 where, for an error probability of 10^{-3}, a family of signal/noise degradation curves is plotted as a function of decorrelation d for different orders of diversity (N = 2, 4, 6). Similar curves are shown in Fig's 5, 6 and 7 for error probabilities of 10^{-4}, 10^{-5} and 10^{-6}, respectively.

Error Probabilities For Exact Coherent Detection

We may evaluate error probabilities in the exact coherent detection case in a manner entirely analogous to that used for the non-coherent detection case. Thus, since the error probability for non-fading, coherent, matched filter detection (zero threshold) of orthogonal equiprobable, equal energy, binary signals in white noise is given by

$$p_L = \frac{1}{2} \operatorname{Erfc} \sqrt{\frac{E}{4N_o}} \tag{34}$$

our desired error probability p_L is given by

$$p_L = \frac{1}{2} \int_o^\infty W(x) \operatorname{Erfc}(\sqrt{sx}\,)\, dx \tag{35}$$

where Erfc ( ) is the complementary error function, s is given by Eq. (22) and W(x) by Eq. (23). Using (23) in (35) and making the change of variable $y = x/2\sigma^2$ one finds

$$p_L = \frac{1}{2} \sum_{j=0}^{L-1} \nu^j (1-\nu)^{L-1-j} \frac{C_j^{L-1}}{j!} \int_o^\infty y^j e^{-y} \operatorname{Erfc}\left[\sqrt{\frac{\rho y}{2}}\right] dy \tag{36}$$

The integral in (36) is most conveniently evaluated by equating the expression p_L in (36) to the error probability expression in the case of fixed reference coherent detection (Eq. (59) for $d = 1/\rho_p = 0$ (or $\nu = 1$)) since in this case both expressions must degenerate to the same error probability, i.e., that for maximal ratio i.f. combining with exact coherent detection. Thus

$$\frac{1}{2} \frac{1}{(L-1)!} \int_o^\infty y^{L-1} e^{-y} \operatorname{Erfc}\left[\sqrt{\frac{\rho y}{2}}\right] dy = \left[\frac{\sqrt{1+2/\rho}-1}{2\sqrt{1+2/\rho}}\right]^L \sum_{m=o}^{L-1} \left[\frac{\sqrt{1+2/\rho}+1}{2\sqrt{1+2/\rho}}\right]^m C_m^{L-1+m} \tag{37}$$

Since (37) is true for any $L \geq 1$, it is true in particular for L-1=j, which yields the desired integral in (36). Thus one determines that for exact coherent detection

$$p_L = \sum_{\zeta=0}^{L-1} \nu^j (1-\nu)^{L-1-j} C_j^{L-1} \left[\frac{\sqrt{1+2/\rho}-1}{2\sqrt{1+2/\rho}}\right]^{j+1} \sum_{m=0}^{j} \left[\frac{\sqrt{1+2/\rho}+1}{2\sqrt{1+2/\rho}}\right]^m C_m^{L-1+m} \tag{38}$$

Examination of (38) shows that decorrelation produces the same type of asymptotic effect for large signal/noise ratios in the coherent detection case as was observed for the non-coherent detection case. Thus setting

$$1-\nu \approx 2d \quad ; \quad \rho_p \gg \frac{1}{d} \gg 1$$
$$\frac{\sqrt{1+2/\rho}-1}{2\sqrt{1+2/\rho}} \approx \frac{1}{2\rho} \quad ; \quad \rho \gg 2 \tag{39}$$

one finds that

$$p_L \approx \frac{(2d)^{j-1}}{2\rho} \qquad (40)$$

Note that the asymptotic behavior is the same as for non-coherent detection except that the coherent case has a 3db signal/noise advantage.

It may be shown, as for the non-coherent detection case, that the higher the order of diversity the less the signal/noise ratio degradation caused by a given decorrelation.

Error Probabilities for Fixed Reference Coherent Detection

The fixed reference coherent matched filter detector is shown in Fig. 8. Two matched filters are indicated, each matched to one of the two possible transmitted binary pulses. The matched filter outputs are subtracted, mixed to d.c. by a reference carrier whose phase is properly adjusted, and then low pass filtered. After low pass filtering the waveform is sampled and then compared with a threshold, usually set at zero, such that if the sample exceeds threshold a "1" is printed, and if it is less than threshold a "0" is printed. The sampling times occur periodically and must be "phased" so that they occur at the trailing edges of successive input pulses (with due compensation for a possible delay in the low pass filter.)

The detector in Fig. 8 becomes an exact coherent detector when the phase of the reference carrier (with respect to $2\pi f_o t$) is chosen equal to the angle of $Z(t)$ in Eq's (12) and (13). It becomes a fixed reference coherent detector when the phase of the reference carrier is set equal to the angle of $\overline{Z(t)}$, i.e. equal to Θ. In the absence of selective fading and pilot tone noise the angle of $Z(t)$ becomes constant and equal to Θ. Thus for $d = 1/\rho_p = 0$ the fixed reference coherent detector becomes an exact coherent detector.

Since we do not have exact coherent detection one can not average over the non-fading coherent detection error probability, Eq. (34), to obtain the error probability for fixed reference coherent detection. Although the same type of two-step procedure could be used to evaluate the error probabilities for fixed reference coherent detection as was used for incoherent and exact coherent detection, we prefer to use a one-step procedure in which we average over the variables g_k, ξ_k, h_k and η_k simultaneously. An interesting feature of this one step procedure is that it brings to light a previously unreported correspondence between the expression for error probability in the cases of maximal ratio coherent diversity combining with coherent detection and square-law diversity combining.

Since we are assuming that the fading is a stationary (and ergodic) process, it will be sufficient to consider error probabilities on an ensemble basis. Thus we will consider error probabilities associatd with a pulse whose trailing edge occurs at t = T. Fig. 9 shows a functional block diagram of the matched filter detector in which only complex envelopes of waveforms are involved. As far as complex envelopes are concerned, the detection system is linear and time invariant up to the thresholding action. Thus, to simplify the discussion, the filtering and sampling operations can be combined as shown in Fig. 9 where these combined operations are indicated by integrals over the range $0 < t < T$. In complex envelope notation, multiplication by the reference carrier in the actual system becomes multiplication by $e^{-j\Theta}$, and selection of the low frequency components becomes a real part extraction operation. (The factor of 2 shown in Fig. 4 is merely a convenient normalization). Since the fading is slow compared to a bit duration, one may regard $g_k(t)$, $\xi_k(t)$, and $h_k(t)$ as constant in the integrations indicated in Fig. 9.

Using the expression for w(t) (Eq. (10)) in the various operations indicated in Fig. 9, plus the slow fading assumption, it is readily found that q, the input to the threshold device, may be represented as

$$q = \sum_{k=1}^{L} \left[U_k V_k^* + U_k^* V_k \right] \qquad (41)$$

where

$$U_k = h_k \int_0^T S(t) \left[S_1^*(t) - S_0^*(t) \right] dt + \int_0^T \eta_k(t) e^{-j2\pi Ft} \left[S_1^*(t) - S_0^*(t) \right] dt \qquad (42)$$

$$V_k = g_k + \xi_k$$

In (42) we have omitted the arguments of the functions h_k, g_k and ξ_k, since, because these functions are assumed stationary processes, their

arguments do not play a part in the subsequent derivations.

There are two error probabilities of concern here,

$$\Pr\left[q > 0/S(t) = S_0(t);\ 0 < t < T\right]$$

$$\Pr\left[q < 0/S(t) = S_1(t);\ 0 < t < T\right] \tag{43}$$

i.e. the probability that "1" is printed given that "0" is sent and vice-versa. We shall determine these error probabilities by integrating over the probability density function of q. Our quadratic form (41) is a special case of the quadratic form

$$q = \sum_{k=1}^{L}\left[a\left|X_k\right|^2 + b\left|Y_k\right|^2 + cX_k^* Y_k + c^* X_k Y_k^*\right] \tag{44}$$

where a, b are real and the pairs of identically distributed zero mean complex-valued gaussian variables (X_k, Y_k) are independent of one another. In Appendix 2 it is shown that W(q) the density function of q in (44), is given by

$$W(q) = \begin{cases} e^{-\alpha q}\sum_{m=0}^{L-1} \dfrac{(\alpha\beta)^L}{(\alpha+\beta)^{L+m}} C_m^{L-1+m} \dfrac{q^{L-1-m}}{(L-1-m)!} & ;\ q>0 \\ e^{\beta q}\sum_{m=0}^{L-1} \dfrac{(\alpha\beta)^L}{(\alpha+\beta)^{L+m}} C_m^{L-1+m} \dfrac{(-q)^{L-1-m}}{(L-1-m)!} & ;\ q<0 \end{cases} \tag{45}$$

where

$$\alpha = \sqrt{\left[\frac{p_{xx}+p_{yy}}{2\det P}\right]^2 - \frac{1}{\det P}} + \frac{p_{xx}+p_{yy}}{2\det P}$$

$$\beta = \sqrt{\left[\frac{p_{xx}+p_{yy}}{2\det P}\right]^2 - \frac{1}{\det P}} - \frac{p_{xx}+p_{yy}}{2\det P} \tag{46}$$

in which p_{xx}, p_{yy} are the diagonal elements, and detP is the determinant, of a matrix P. This matrix is equal to the product of the matrix of the elementary quadratic form in (44) with the typical moment matrix for the pair (X_k, Y_k), i.e.,

$$P \equiv \begin{bmatrix} p_{xx} & p_{xy} \\ p_{xy}^* & p_{yy} \end{bmatrix} = \begin{bmatrix} m_{xx} & m_{xy} \\ m_{xy}^* & m_{yy} \end{bmatrix}\begin{bmatrix} a & c \\ c^* & b \end{bmatrix} \tag{47}$$

where

$$m_{xx} = \overline{\left|X_k\right|^2}\ ; \quad m_{yy} = \overline{\left|Y_k\right|^2}$$

$$m_{xy} = \overline{X_k^* Y_k} \tag{48}$$

The probability that the variable q is positive, $\Pr(q > 0)$, and the probability that it is negative $\Pr(q < 0)$, are given by integrating over W(q) as indicated below,

$$\Pr(q>0) = \int_0^\infty W(q)\,dq$$

$$= \left[\frac{1+\gamma}{2+\gamma}\right]^L \sum_{m=0}^{L-1} \frac{1}{(2+\gamma)^m} C_m^{L-1+m}$$

$$\Pr(q<0) = \int_{-\infty}^0 W(q)\,dq \tag{49}$$

$$= \frac{1}{(2+\gamma)^L} \sum_{m=0}^{L-1} \left[\frac{1+\gamma}{2+\gamma}\right]^m C_m^{L-1+m}$$

where the parameter γ is given by

$$\gamma = \frac{\beta}{\alpha} - 1$$

$$= \frac{(p_{xx}+p_{yy})/\det P}{\sqrt{\left[\frac{p_{xx}+p_{yy}}{2\det P}\right]^2 - \frac{1}{\det P}} - \frac{p_{xx}+p_{yy}}{2\det P}} \tag{50}$$

The quadratic form (44) reduces to the form (41) by setting a = b = 0, c = 1, and identifying $U_k = X_k$, $V_k = Y_k$. In this case one readily finds that

$$p_{xx} + p_{yy} = \overline{U_k V_k^*} + \overline{U_k^* V_k} \tag{51}$$

$$- \det P = \overline{|U_k|^2}\,\overline{|V_k|^2} - \overline{|U_k V_k^*|^2}$$

If we substitute (51) in (50) and define the correlation coefficient

$$\mu = \frac{\overline{U_k^* V_k}}{\sqrt{\overline{|U_k|^2} \cdot \overline{|V_k|^2}}} \tag{52}$$

we find that in our case γ may be expressed as

$$\gamma = \frac{2\mathrm{Re}(\mu)}{\sqrt{1 - \mu^2 + \mathrm{Re}^2(\mu)} - \mathrm{Re}(\mu)} \tag{53}$$

To evaluate μ we need the moments

$$\overline{|U_k|^2} = \overline{|h_k|^2}\left|\int_0^T S(t)\left[S_1^*(t) - S_0^*(t)\right]dt\right|^2 + \overline{\left|\int_0^T \eta_k(t) e^{-j2\pi Ft}\left[S_1^*(t) - S_0^*(t)\right]dt\right|^2}$$

$$\overline{|V_k|^2} = \overline{|g_k|^2} + \overline{|\varepsilon_k|^2} \tag{54}$$

$$\overline{V_k^* U_k} = \overline{g_k^* h_k}\int_0^T S(t)\left[S_1^*(t) - S_0^*(t)\right]dt$$

Assuming orthogonal, equal energy signals, S_1 and S_0, the error probabilities p_1 and p_0 become equal. Then, confining our attention to the case in which $S(t) = S_1(t)$ for $0 < t < T$,

$$\int_0^T S(t)\left[S_1^*(t) - S_0^*(t)\right]dt = \int_0^T |S_1(t)|^2 dt = 2E_t \tag{55}$$

The noise $\eta_k(t)$ is flat with power spectral density $4n_o$ (the factor of 4 makes the real noise have two sided spectral density n_o). Then assuming the bandwidth of $\eta_k(t)$ much bigger than that of $S_1(t)$, it is quickly shown that

$$\overline{\left|\int_0^T \eta_k(t) e^{-j2\pi Ft}\left[S_1^*(t) - S_0^*(t)\right]dt\right|^2} = 16\, n_o E_t \tag{56}$$

Using (11), (55) and (56), and the definition (25) in (54) and then evaluating μ, we find that

$$\mu = \frac{4\sigma^2 r E_t}{\sqrt{(8\sigma^2 E_t^2 + 16 n_o E_t)(2\sigma^2 + 2\sigma_p^2)}} = \frac{r}{\sqrt{(1 + \frac{2}{\rho})(1 + \frac{1}{\rho_p})}} = \sqrt{\frac{\nu}{1 + \frac{2}{\rho}}} \tag{57}$$

Thus from (53)

$$\gamma = \frac{2\sqrt{\frac{\nu}{1 + 2/\rho}}}{1 - \sqrt{\frac{\nu}{(1 + 2/\rho)}}} \tag{58}$$

The error probability for the system is obtained by substituting γ of (58) into the lower equation of (49). Thus

$$p_L = \left[\frac{\sqrt{\frac{1}{\nu}(1 + 2/\rho)} - 1}{2\sqrt{\frac{1}{\nu}(1 + 2/\rho)}}\right]^L \sum_{m=0}^{L-1}\left[\frac{\sqrt{\frac{1}{\nu}(1 + 2/\rho)} + 1}{2\sqrt{\frac{1}{\nu}(1 + 2/\rho)}}\right]^m C_m^{L-1+m} \tag{59}$$

It is interesting to compare our results with those of a system employing non-coherent combining and detection. Pierce[1] has evaluated error probabilities for FSK transmission, matched filtering, square law envelope detection, and linear post-detection combining. His results are actually valid not only for FSK but for any pair of equal energy orthogonal waveforms. An examination of his expression for error probability (his Eq. (9)) shows it to be identical with ours if γ is set equal to the signal-to-noise ratio ρ. (Our ρ is equal to his $S_0 T/n_o$). This identity of form is a result of the fact that the receiver operation for square law envelope detection and linear combining involves the comparison of a quadratic form with a zero threshold wherein the quadratic form is the special case of (44) that results when $a = 1$, $b = -1$, $c = 0$.

The large signal/noise behavior of p_L in (59) is different from that of the exact coherent and the non-coherent detection cases, in that the error probability becomes asymptotically independent of ρ. Specifically, from (59) it may be determined that

$$p_L \approx \frac{C_L^{2L-1}}{2^L} \left(d + \frac{1}{\rho}\right)^L \; ; \; \rho_p >> \frac{1}{d} >> 1 \qquad (60)$$

However, one may show, as for the previous two detection methods, that the higher the order of diversity the less the degradation caused by a given decorrelation.

It will be recalled that $d = 1 - r$ where r is the magnitude of the correlation coefficient that would exist between the complex envelope of the received pilot tone and the complex envelope of a tone transmitted at the center of the data channel. While the correlation coefficient r may be measured directly, it is common practice to measure r_e, the correlation coefficient between the envelopes of the two tones. In Appendix 3 it is shown that for r and thus r_e close to unity

$$r_e \approx \frac{1 + r - \frac{\pi}{2}}{2 - \frac{\pi}{2}} \qquad (61)$$

Appendix 1

For simplicity of notation we define

$$z_k = g_k + \xi_k \qquad (62)$$

so that

$$x = \frac{\left| \sum_{k=1}^{L} z_k h_k^* \right|^2}{\sum_{k=1}^{L} |z_k|^2} \qquad (63)$$

If we define the quadratic forms

$$q_1 = \mathrm{Re}\left[\sum_{k=1}^{L} z_k h_k^*\right] = \frac{1}{2}\sum_{k=1}^{L}\left[z_k h_k^* + z_k^* h_k\right]$$

$$q_2 = \mathrm{Im}\left[\sum_{k=1}^{L} z_k h_k^*\right] = \frac{-j}{2}\sum_{k=1}^{L}\left[z_k h_k^* + z_k^* h_k\right]$$

$$q = \sum_{k=1}^{L} |z_k|^2 \qquad (64)$$

The random variable x may be expressed as

$$x = \frac{q_1^2 + q_2^2}{q} \qquad (65)$$

We shall evaluate the density function of x by first expressing the joint density function of q_1, q_2, q as the inverse transform of its characteristic function, making the change of variables

$$q_1 = R \cos \phi$$

$$q_2 = R \sin \phi \qquad (66)$$

and evaluating the joint density function of R, ϕ, q. Following this the variable ϕ will be integrated out yielding the joint density function of R and q. The density function of x may be represented as a single integration over this joint density function.

The joint characteristic function of q_1, q_2, and q is given by

$$F(t_1, t_2, t_3) = \overline{\exp\left[j(t_1 q_1 + t_2 q_2 + t_3 q)\right]} = \prod_{k=1}^{L} \overline{\exp(jQ_k)} \qquad (67)$$

where the Hermitian quadratic form Q_k is given by

$$Q_k = \frac{1}{2}(t_1 - jt_2) z_k h_k^* + \frac{1}{2}(t_1 + jt_2) z_k^* h_k + t_3 |z_k|^2 \qquad (68)$$

The average $\overline{\exp\left[jQ_k\right]}$ is just the characteristic function of the quadratic form Q_k evaluated at unity. From Eq. (94) in Appendix 2 we determine that

$$\overline{\exp\left[jQ_k\right]} = 1/\det\left[I - jM_kQ\right] \tag{69}$$

where the moment matrix M_k is given by

$$M_k = \begin{bmatrix} \overline{|z_k|^2} & \overline{z_k^* h_k} \\ \overline{z_k h_k^*} & \overline{|h_k|^2} \end{bmatrix} \tag{70}$$

and the quadratic form matrix Q is given by

$$Q = \begin{bmatrix} t_3 & \frac{1}{2}(t_1 - jt_2) \\ \frac{1}{2}(t_1 + jt_2) & 0 \end{bmatrix} \tag{71}$$

Then defining

$$u = \mathrm{Re}\left[\overline{z_k^* h_k}\right], \quad \delta = \mathrm{Im}\left[\overline{z_k h_k^*}\right], \quad m = \overline{|z_k|^2}$$

$$\Delta = \overline{|z_k|^2}\ \overline{|h_k|^2} - \overline{|z_k^* h_k|^2} \tag{72}$$

one readily determines that

$$F(t_1, t_2, t_3) = \frac{1}{\left[1 - jt_1u - jt_2\delta + \frac{1}{4}(t_1^2 + t_2^2)\Delta - jt_3m\right]^L} \tag{73}$$

The joint density function of q_1, q_2 and q is then

$$W(q_1, q_2, q) =$$

$$\frac{1}{(2\pi)^3} \int\int\int_{-\infty}^{\infty} \frac{e^{-j(t_1q_1 + t_2q_2 + t_3q)} dt_1 dt_2 dt_3}{\left[1 - jt_1u - jt_2\delta + \frac{1}{4}(t_1^2 + t_2^2)\Delta - jt_3m\right]^L} \tag{74}$$

Noting that $W_1(R, \phi, q)$ the joint density function of R, ϕ, and q is related to $W(q_1, q_2, q)$ by

$$W_1(R, \phi, q) = RW(R\cos\phi,\ R\sin\phi, q) \tag{75}$$

and changing the integration variables as follows

$$\begin{aligned} t_1 &= r\cos\theta \\ t_2 &= r\sin\theta \\ -jt_3 &= w \end{aligned} \tag{76}$$

one readily finds that

$$W_1(R, \phi, q) =$$

$$\frac{1}{(2\pi)^2} \int_0^{\infty}\int_0^{2\pi} \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} dw\, \frac{e^{-jRr\cos(\theta - \phi)}\, e^{wq}\, dr\, d\theta}{\left[1 - jrP\cos(\theta - \theta') + \frac{1}{4}r^2\Delta + wm\right]^L} \tag{77}$$

where

$$\overline{z_k h_k^*} \equiv Pe^{j\theta'} = u + j\delta \tag{78}$$

The integral over w is of the form

$$\frac{1}{2\pi j} \int_{-j\infty}^{j\infty} \frac{e^{wq}}{(w + z)^L}\, dw = \frac{1}{(L-1)!}\, q^{L-1} e^{-zq}; \quad q > o \tag{79}$$

and the subsequent integration over θ requires the integral

$$\frac{1}{2\pi} \int_0^{2\pi} e^{-jA\cos(\theta - \phi) - jB\cos(\theta - \theta')} dt = J_o\left[\sqrt{A^2 + B^2 + 2AB\cos(\phi - \theta')}\right] \tag{80}$$

where J_o is the zeroth order Bessel function of the first kind. The last integration requires the integral

$$\int_0^{\infty} e^{-sY} J_o(\alpha\sqrt{Y})\, dy = \frac{1}{s} e^{\frac{-\alpha^2}{4s}} \tag{81}$$

Using (79), (80) and (81) to evaluate (77) one finds

$$W_1(R, \phi, q) =$$

$$\left(\frac{q}{m}\right)^{L-2} \frac{e^{-\frac{q}{m}(1 + \frac{P^2}{\Delta})}}{m\Delta} \frac{2Re^{-\frac{R^2m}{q\Delta} - \frac{2RP}{\Delta}\cos(\phi - \theta')}}{(L-1)!} \tag{82}$$

The joint density function of R and q, $W_2(R, q)$

is given by

$$W_2(R, q) = \int_0^{2\pi} W(R, \phi, q)\, d\phi =$$

$$\frac{q}{m}^{L-2} \quad \frac{e}{m\Delta}^{-\frac{q}{m}(1+\frac{p^2}{\Delta})} \quad \frac{2Re}{(L-1)!}^{-\frac{R^2 m}{q\Delta}} \quad I_o(2R\frac{P}{\Delta}) \tag{83}$$

where I_o is the zeroth order modified Bessel function of the first kind. In (83) use has been made of the integral

$$\frac{1}{2\pi}\int_0^{2\pi} e^{-\beta\cos(\phi-\theta')}d\phi = I_o(\beta) \tag{84}$$

which may be obtained from (80) by noting that $J_o(j\beta) = I_o(\beta)$

It is readily seen that the density function for x may be expressed as

$$W(x) = \frac{d}{dx}\left[\Pr(R^2 < qx,\ o < q < \infty)\right] \tag{85}$$

By expressing the probability in (85) as an integration over the joint density function $W_2(R, q)$ and then differentiating with respect to x there results

$$W(x) = \frac{1}{2}\int_0^{\infty}\sqrt{\frac{q}{x}}W_2(\sqrt{qx}, q)dq \tag{86}$$

Using (83) in (86) one finds

$$W(x) = \frac{e}{\Delta}^{-\frac{xm}{\Delta}} \int_0^{\infty}\left(\frac{q}{m}\right)^{L-1} \frac{e}{(L-1)!}^{-\frac{q}{m}(1+\frac{p^2}{\Delta})} I_o(\frac{2p}{\Delta}\sqrt{qx})dq \tag{87}$$

Instead of carrying this integration through and directly evaluating W(x) it is easier to calculate the Laplace transform F(s) first and then inverse transform to obtain W(x). Thus

$$F(s) = \int_0^{\infty}\left(\frac{q}{m}\right)^{L-1} \frac{e}{(L-1)!\Delta}^{-\frac{q}{m}(1+\frac{p^2}{\Delta})} \int_0^{\infty} e^{-x(s+\frac{m}{\Delta})} I_o(\frac{2p}{\Delta}\sqrt{qx})dx\, dq \tag{88}$$

The integral over x may be obtained from (81) by noting that $J_o(j\alpha) = I_o(\alpha)$ and the integral over q is just the transform relation inverse to that in Eq. (79). Performing these integrations and noting that

$$\frac{p^2}{\Delta} = \frac{\nu}{1-\nu}\ ;\ \frac{m}{\Delta} = \frac{1}{2\sigma^2}\cdot\frac{1}{1-\nu} \tag{89}$$

where ν is given by Eq. (25), we find

$$F(s) = \frac{(s2\sigma^2(1-\nu)+1)^{L-1}}{(s2\sigma^2+1)^L} = \sum_{k=o}^{L-1} \frac{\nu^k(1-\nu)^{L-1-k}\,C_k^{L-1}}{(s2\sigma^2+1)^{k+1}} \tag{90}$$

Then inverse transforming the sum in (90) term by term (Eq. (79)) one obtains

$$W(x) = \sum_{k=o}^{L-1}\left(\frac{x}{2\sigma^2}\right)^k \frac{e}{k!2\sigma^2}^{-\frac{x}{2\Delta^2}} \nu^k(1-\nu)^{L-1-k}C_k^{L-1} \tag{91}$$

Appendix 2

The probability density function for q, Eq. (44), will be evaluated by first determining its characteristic function and then Fourier transforming. Note that q may be expressed as

$$q = \sum_{k=1}^{L} q_k \tag{92}$$

where

$$q_k = a|X_k|^2 + b|Y_k|^2 + cX_k^*Y_k + c^*X_kY_k^* \tag{93}$$

Because of the assumed independence and identical statistics of (X_k, Y_k) and (X_ℓ, Y_ℓ), $k \neq \ell$, the elementary Hermitian quadratic forms q_k and q_ℓ are statistically independent and identically distributed.

Turin[2] has shown that the characteristic function $\phi_D(jt)$ of a Hermitian quadratic form D in zero-mean complex-valued normal random variables is given by

$$\phi_D(jt) = e^{\overline{jtD}} = \frac{1}{\det\left[I - jt\,MQ\right]} \tag{94}$$

where M is the moment matrix of the complex normal variables, Q is the matrix of the quadratic form, and I is the identity matrix. Let the matrix P be defined as

$$P = MQ \tag{95}$$

where the elements of P are defined In Eq. (47). Evaluating the determinant in (94) for P as given in Eq. (47) shows that the characteristic function of q_k is given by

$$f(t) = \frac{\alpha\beta}{(it + \alpha)(it + \beta)} \tag{96}$$

where α and β are given by Eq. (46). Since the characteristic function of a sum of independent random variables is the product of the individual characteristic functions, the characteristic function for q, F(t), is given by

$$F(t) = \left[f(t)\right]^L = \left(\frac{\alpha}{it+\alpha}\right)^L \left(\frac{\beta}{it-\beta}\right)^L \tag{97}$$

If we make use of the fact that multiplication in the transform domain becomes convolution in the inverse domain, we may evaluate the density function W(q) by convolving the two functions

$$W_1(q) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\alpha^L}{(it+\alpha)^L}\, e^{-itq}dq$$

$$= \begin{cases} \dfrac{q^{L-1}}{(L-1)!}\, e^{-\alpha q} & ;\ q \geq 0 \\ 0 & ;\ q < 0 \end{cases} \tag{98}$$

$$W_2(q) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\beta^L}{(it-\beta)^L}\, e^{-itq}dq$$

$$= \begin{cases} 0 & ;\ q \geq 0 \\ \dfrac{(-q)^{L-1}}{(L-1)!}\, e^{\beta q} & ;\ q < 0 \end{cases}$$

Thus

$$W(q) = \int_0^{\infty} W_1(y)\, W_2(q-y)dy \tag{99}$$

Eq. (45) is obtained upon carrying through the convolution in (99).

Appendix 3

The envelope correlation coefficient r_e may be expressed as

$$r_e = \frac{\overline{|g_k|\,|h_k|} - \overline{|g_k|}\;\overline{|h_k|}}{\sqrt{\left[\overline{|g_k|^2} - \overline{|g_k|}^2\right]\left[\overline{|h_k|^2} - \overline{|h_k|}^2\right]}} \tag{100}$$

We may use a result of Lawson and Uhlenbeck (pg. 62, Eq. 74 of Ref. 3) to determine that

$$\overline{|g_k|\,|h_k|} = \sigma^2\left[2E(r) - (1 - r^2)K(r)\right]$$

$$= \frac{\pi}{2}\sigma^2\left[1 + \frac{r^2}{4} + \frac{r^2}{64} + \ldots\right] \tag{101}$$

where K and E are the complete elliptic integrals of the first and second kind. From

$$\left[\overline{|g_k|\,|h_k|}\right]_{r=0} = \overline{|g_k|}\;\overline{|h_k|}$$

$$\overline{|g_k|^2} = \overline{|h_k|^2} = 2\sigma^2 \tag{102}$$

we obtain

$$r_e = \frac{2E(r) - (1 - r^2)\,K(r) - \frac{\pi}{2}}{2 - \frac{\pi}{2}} \qquad (103)$$

Jahnke and Emde[4] give an expansion of K and E in ascending powers of $(1 - r^2)$.† If these expansions are used in (103) and only first order terms in d = 1 - r retained, Eq. (61) results.

Acknowledgement

The work reported in this paper was performed at ADCOM, Inc. under subcontract with ITT Communication Systems, Inc., Paramus, New Jersey. The authors are greatful to Messrs. William S. Litchman and Joseph W. Halina of ITT Communication Systems for their permission to publish this paper.

† On the bottom of page 73 in Ref. 4 identify $k' = \sqrt{1 - r^2}$.

References

1. Pierce, J. N., "Theoretical diversity improvement in frequency-shift keying," Proc. I.R.E., Vol. 46, pp. 903-910, May, 1958.

2. Turin, G. L. "Some computations of error rates for selectively fading multipath channels," N.E.C., Vol. 15, pp. 431-440.

3. Lawson, J. L., and Uhlenbeck, G. E. Threshold Signals, McGraw-Hill, 1950.

4. Jahnke, E., and Emde, F., Tables of Functions, Dover, 1945.

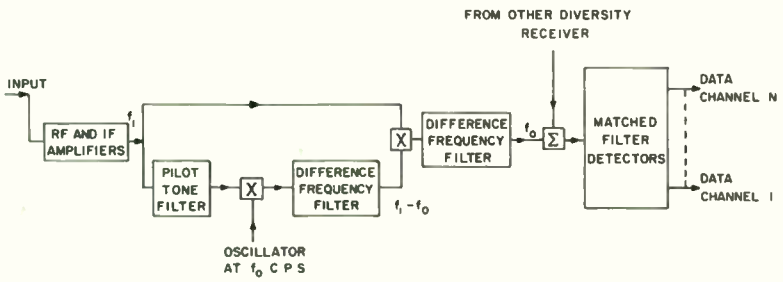


Fig. 1. Block diagram of coherent combining method.

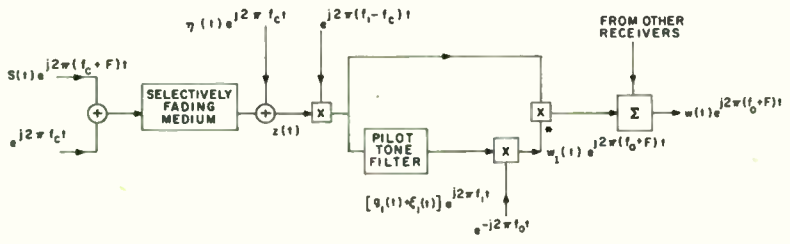


Fig. 2. Waveforms in predetection portion of system.

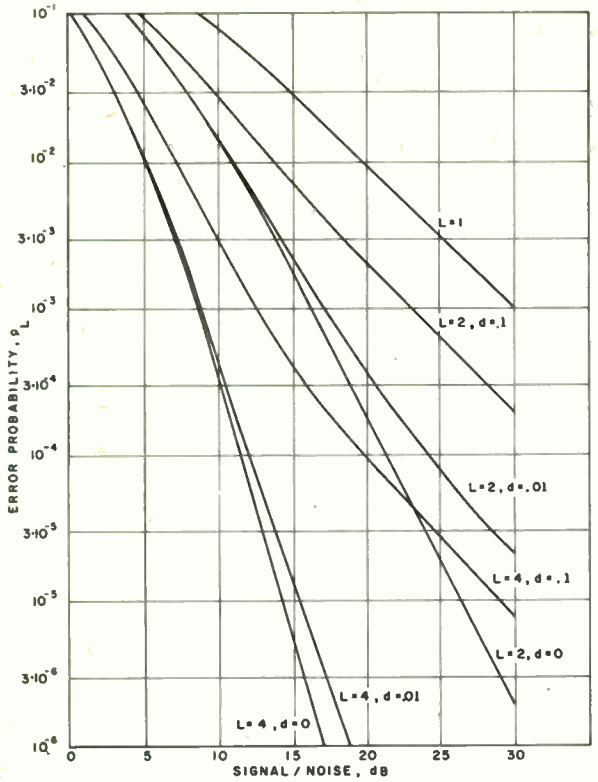


Fig. 3. Error probability as a function of signal/noise ratio.

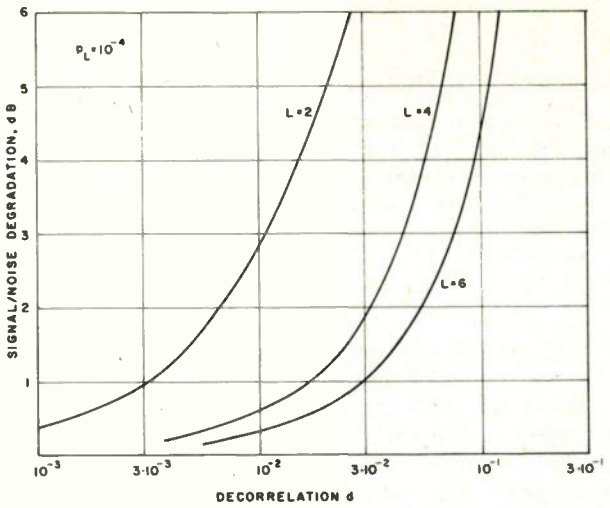


Fig. 5. Signal/noise degradation as a function of decorrelation for different orders of diversity with $p_L = 10^{-4}$.

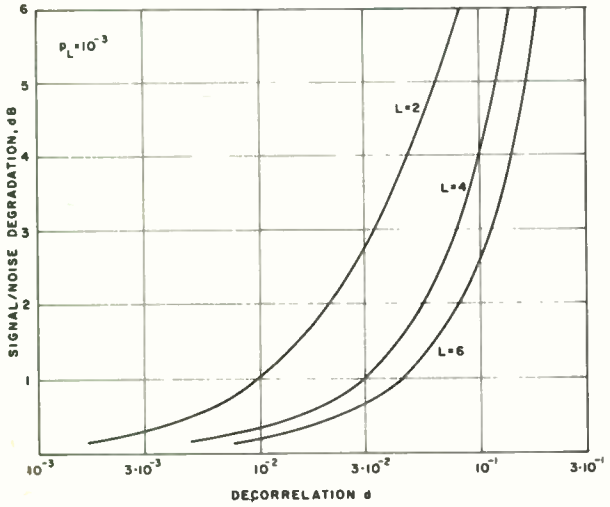


Fig. 4. Signal/noise degradation as a function of decorrelation for different orders of diversity with $p_L = 10^{-3}$.

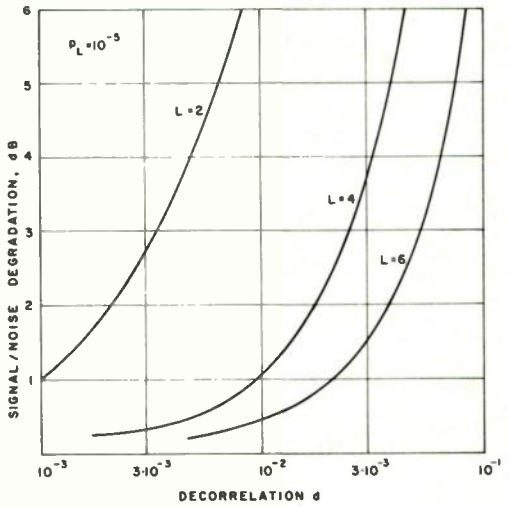


Fig. 6. Signal/noise degradation as a function of decorrelation for different orders of diversity with $p_L = 10^{-5}$.

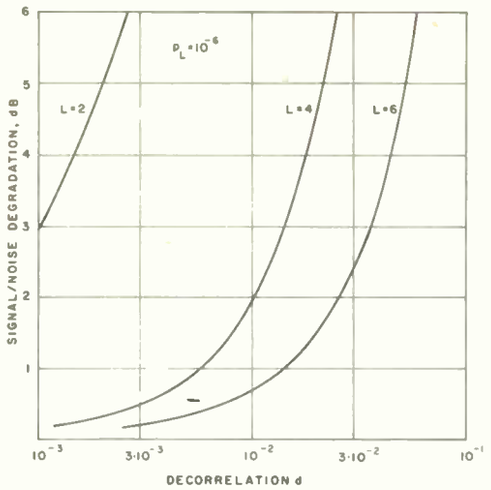


Fig. 7. Signal/noise degradation as a function of decorrelation for different orders of diversity with $p_L = 10^{-6}$.

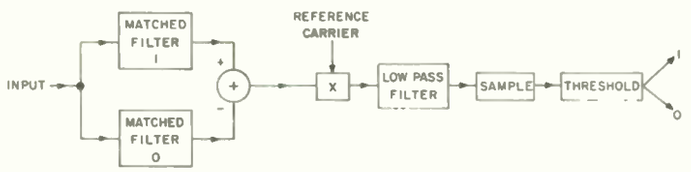


Fig. 8. Fixed reference coherent detector.

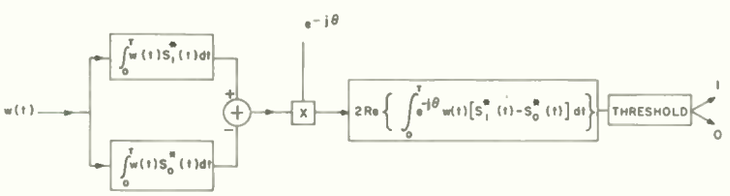


Fig. 9. Mathematical operations in fixed-reference coherent detector.

PIECE-WISE DIVERSITY COMBINING

Robert T. Adams
Sichak Associates
Nutley, N. J.

Barry M. Mindes
ITT Federal Labs.
Nutley, N. J.

Abstract

By subdividing a broad channel in narrower bands, and subsequently re-assembling the division, predetection diversity combining can be extended to wide-band signals in a multipath environment, where a single phase adjustment is not adequate for proper combining.

A piecewise dual diversity combiner has been tested in the laboratory, using one sideband of signals received from BBC broadcasts. The results and the tests are described.

While initially investigated for a specific HF radio application, the piecewise combining technique has direct application to extend the useful bandwidth of systems such as troposcatter, which are presently limited by multipath considerations.

A TIME DIVERSITY TECHNIQUE FOR SPEECH TRANSMISSION OVER RAPIDLY FADING CHANNELS (ECHOPLEX)

Leonard R. Kahn
Kahn Research Laboratories, Inc.
Freeport, L.I., New York

Introduction

Diversity systems may be used to overcome the deleterious effects of fading. In order to accomplish this objective, it is necessary that the component diversity signals fade independently. There are three basic methods which may be used to obtain the required independent fading characteristic.

The first and most widely used is space[1] or antenna diversity. In this system, two or more antennas are spaced a considerable distance apart so that the wave received by each antenna fades independently. This independent fading characteristic may also be achieved by a related technique, polarization diversity.

A second basic diversity technique is frequency diversity. In this system, the signal is transmitted on two or more frequencies. The use of this method has gradually declined because of the excessive required bandwidth.

The third basic technique for obtaining independent fading signals is time diversity. Time diversity systems are not in general use because the methods proposed to date require additional spectrum and much additional equipment. The reason for this is that these time diversity systems transmit the intelligence two or more times and thus require additional circuits.

Echoplex is a new time diversity technique that does not require increased spectrum space nor does it require additional antennas, receivers or transmitters. However, in its present form, its use is restricted to the transmission of signals having a high degree of frequency redundancy, such as voice or music.

Block Diagram Description

Figure 1 is a simplified block diagram of the Echoplex system. The voice spectrum is split into small segments by audio bandpass filters. The outputs of filters "F_1 to F_2" and "F_4 to F_5" are fed directly to the radio transmission system. The outputs of filters "F_2 to F_3" and "F_5 to F_6" are fed to a time delay network having a delay of, for example, 1/2 second. Filters "F_3 to F_4" and "F_6 to F_7" are fed to a time delay device having a delay of, for example, one second. The outputs of the two time delay networks are also fed to the associated radio transmission equipment. Thus, the modulating wave fed to the transmitter is a complex time function of the audio input signal. A simple description equation may be written:

$$e = f(t) \Big|_{w_1}^{w_2} + f(T-t) \Big|_{w_2}^{w_3}$$

$$+ f(T-2t) \Big|_{w_3}^{w_4} + f(T) \Big|_{w_4}^{w_5} + \ldots.$$

This wave has two very long echo components and the wave cannot be easily understood without the use of special decoding equipment.

The decoding equipment which is installed in the receiver, is similar to the equipment used at the transmitter. In this case, however, the segments of the voice wave that were subject to delays of less than a second at the transmit end, are delayed sufficiently to produce an overall delay of all components of one second. Thus, components

which were delayed 1/2 second are delayed an additional 1/2 second and those components that were not delayed at the transmitter are delayed one second at the receiver. Those segments that were delayed a second in the transmitter are not further delayed at the receiver. In this manner all of the components of the speech wave arrive at the output of the equipment at the same instant and the resulting output is undistorted and of good quality.

The individual segments may be combined by use of various diversity combiner techniques[2,3] or they may be simply added together in a linear network. It should be noted, however, that the diversity combining problem is somewhat unique in that the signals to be combined carry different information. Therefore, the decision as to whether a component is to be favored or not is largely dependent upon the frequency flatness of the fades experienced. If the fades are quite broad in terms of frequency, then the combining can be done on the basis of the amplitude of the carrier component or other constant amplitude transmitted waves. Simple linear addition of the signal components should only be used if the automatic volume control circuitry is made slower than the fading rate, otherwise the noisier signal will be favored in the combination process. A simple procedure would be to use fixed gain or very slow AGC and linear addition of the components.

Advantages of the Echoplex System

There are three distinct advantages to be gained by use of this system. The first results from the fact that the probability of a sharp fade lasting for more than one second is small. Since redundant segments of the voice are transmitted at different time intervals, the system has effective anti-fade characteristics.

Squelch techniques are commonly used in systems where sharp fades are experienced or systems which use modulation techniques having sharp thresholds, such as frequency modulation. The justification for using squelch techniques is that it is better to temporarily interrupt the signal than to suffer sudden bursts of noise. The interruption in speech may cause the loss of an occasional word but the likelihood is that the noise burst would have masked the signal anyway and therefore it appears best to provide the squelch action.

Echoplex may be used to improve squelch system performance. During the periods when the squelch operates the receiver to the Off position, the chances are that some of the segments will be received if Echoplex is used. The only situation which would cause the loss of all of the information would be if three noise bursts happen to occur at exactly the proper time to interfere with the three transmissions of segments of speech. Thus, the probability is high that the listener will be provided with at least the major portion of the speech at all times and therefore will not suffer the loss of intelligence that he would when using a conventional squelch system.

The second advantage of the system is that the peak-to-average ratio of the Echoplex encoded wave is considerably smaller than a normal speech wave. It is well known that voice signals have a very large peak-to-average ratio. Also, there are syllabic pauses in the signal during which the transmission equipment is not being effectively utilized. The Echoplex system, by spreading out the voice wave in time, reduces the crest factor of the wave and fills in those idle periods.

It should be noted that the encoding process used in Echoplex makes the speech wave become more random in character and more closely approximates white noise. Information theorists have pointed out that the ideal coding system would make the signal approximate random noise in character so that the encoded signal more fully utilizes the transmission facility.

Because of the Echoplex encoding technique, the crest factor of speech is reduced by approximately 6 db. Since FM, AM and SSB modulation methods, as well as other modulation systems are limited by the peak level of the signal, reduction in the crest factor provides a gain in signal-to-noise ratio of approximately 6 db.

The third advantage is that the system provides a degree of privacy in that the encoded wave is extremely difficult to understand without special equipment. Thus, it would provide protection against the casual listener eavesdropping. It should be emphasized that the Echoplex system as shown is not a secrecy system but merely a privacy system.

Amount of Time Delay Used

The longer the time delay used, the greater the fading gain. However, when the delay is over a second or two there is very little improvement to be realized by increasing the time delay. The time delay required, in order to achieve the gain due to the decrease in crest factor, is in the order of .1 or .2 seconds; increasing the time delay to a few seconds does not provide appreciable additional gain in crest factor reduction.

For applications which cannot accept a time delay of one second, lesser time delays may be used, although such operation offers less improvement under fading conditions. However, even when time delays of as little as one-tenth of a second are used, considerable fading gain is achieved and almost the full crest factor reduction gain is achieved. If the time delay is increased over 1/2 second the interference between echoes is reduced and the intelligibility of encoded speech is increased reducing the privacy effect.

Summary

In comparison with the conventional diversity systems, Echoplex with a time delay of one second should equal or better the fading gain of space or frequency diversity. In addition, it provides a gain of 6 db which is over and above the gain of conventional diversity techniques.

It does, however, introduce a time delay and it is also necessary that the modulating signal be redundant in terms of frequency. The main application of Echoplex should be in voice communications where push-to-talk operation is used, broadcast relays, mobile applications and in voice transmission to space vehicles. In these and many other applications where the time delay is not a problem, this new technique appears to have appreciable advantages in size, cost and signal-to-noise gain over conventional diversity methods.

Acknowledgments

The author wishes to acknowledge the assistance provided by Heino L. Pull formerly of Kahn Research Laboratories. He also wishes to thank Messrs. Lloyd A. Ottenberg, Stanley F. Turner and James T. Krowl of Kahn Research Laboratories for their assistance in the preparation of this paper, as well as Prof. Walter Lyons of RCA Communications.

References

1. H. O. Peterson, H. H. Beverage, and J. B. Moore, "Diversity Telephone Receiving System of RCA Communications, Inc.", Proc. I.R.E., vol. 19, p.562, April 1931.

2. L. R. Kahn, "Ratio Squarer", Proc. I.R.E., p.1704, November 1954, vol. 42.

3. D. G. Brennan, "Linear Diversity Combining Techniques", Proc. I.R.E., pp.1075-1102; June 1959.

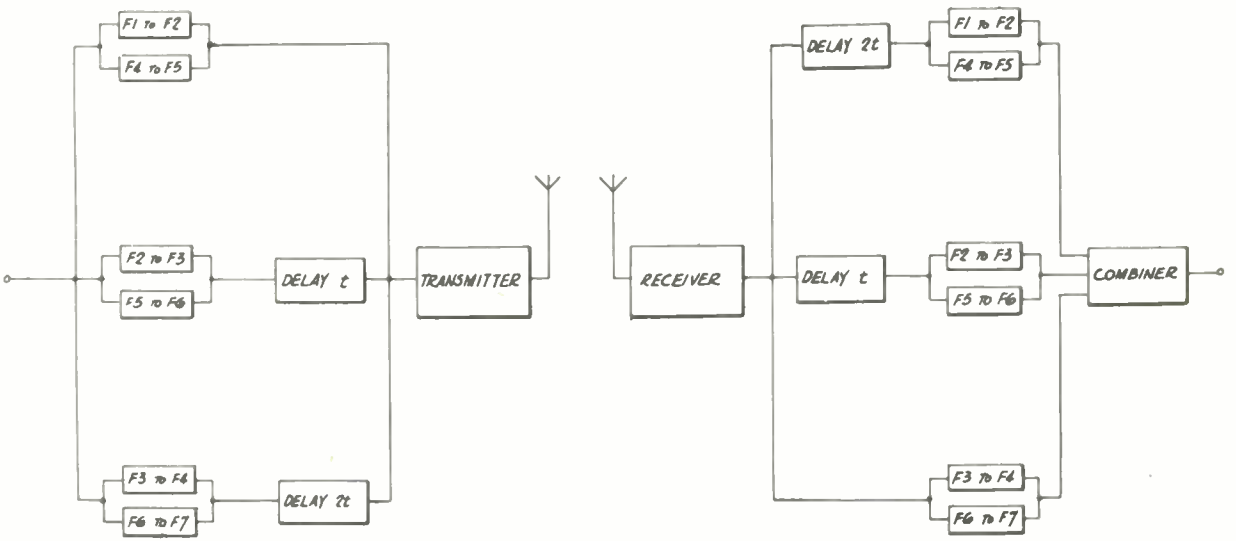


Fig. 1. Simplified block diagram of Echoplex.

MULTIPATH AND SCATTER COMMUNICATION TECHNIQUES

Donald P. Harris
Lockheed Missiles & Space Company
Palo Alto, California

Summary

This paper presents some results of an investigation of random-multipath- and scatter-communication channels and of communication techniques applicable to such channels. The limiting bounds imposed on the performance of adaptive matched-filter receiver techniques by channel-sounding errors are evaluated for a wide range of behavior in practical multipath-channel models. The performance characteristics of a class of incoherent communication techniques that require no channel-sounding provisions are considered. Power requirements and bandwidth requirements per unit information transmission rate are evaluated and compared for these techniques and for differentially coherent narrow-band techniques under similar conditions.

Introduction

Random multipath channels have presented interesting and challenging problems to statistical communication researchers and to practical systems designers as well. Although statistical theorists have been steadily developing better ways of undertaking multipath system synthesis from the very foundations, most of the recent advances in practical systems have been based on refinement and automation of previously established techniques – such items as better detectors, better diversity-combining techniques, automated propagation sounders and channel selectors, error-correcting equipment, and information-feedback techniques. With increasing interest in feedback communication, it is likely that this problem will soon be approached from a fundamental point of view with the methods of statistical analysis.

There are still, however, many gaps between theory and practice with regard to the forward multipath circuit alone. In this paper, attention is directed toward the following important limitations in the present understanding of the forward communication circuit involving random multipath propagation:

- Most theoretical studies to date have not treated the problems of performance analysis and evaluation in sufficient detail to justify and guide subsequent development of practical systems.

- In multipath-system synthesis, it has generally been assumed that certain a priori correct knowledge of short-term channel behavior is available at the receiver. In practice, this knowledge is never precisely correct; the effects of channel-measurement errors need to be considered as well as the influence of channel-sounding requirements on the overall system-optimization problem. The unsolved problems involving channel measurements were recognized clearly in the work of Price[1] and Turin[2]. The same sort of problem arises implicitly in more recent work of Kailath,[3] in connection with an assumption that a time-varying correlation matrix pertaining to the channel transfer function is known a priori at the receiver.

- The problems described above need to be treated in connection with channel models that realistically represent existing troublesome multipath-propagation phenomena. In addition, the performance results achieved with various techniques need to be determined and illustrated in such a manner that meaningful comparisons may be made among them.

The work reported in this paper has been undertaken in an attempt to partially resolve some of the above limitations. It is hoped that interest will be stimulated in practical application of theory to multipath problems, and that the immediate results will give some guidance to practical systems development. The first part of the paper deals with the overall system optimization problem and the performance potentials offered by an adaptive-correlation receiving technique when detailed consideration is given to practical channel-sounding problems. In the second section, some attention is given to a more conventional narrow-band differentially coherent modulation technique and its future possibilities. The case where the received signals are regarded as incoherent noise, with no attention paid to the details of the distortion wrought by the channel, is treated in the third section.

Definition of Terms

The following is a partial list of definitions applying to symbols used in this paper: Additional definitions appear in the text and illustrations.

B_s .... Doppler-spread bandwidth of received signal when a single-frequency sine wave is transmitted over a random multipath-propagation circuit

Γ .... Gamma function

$\mathcal{E}$ .... Probability of receiver decision error per symbol, or error rate

H(y) .. Entropy per second or per degree of freedom of a stochastic function y, as appropriate

m .... Total number of available possibilities from which a transmitted signal waveshape is selected

n .... Number of taps on a delay-line channel model; additive noise signal; number of degrees of freedom of a Chi Squared probability distribution

n_o .... Power spectral density of additive receiver noise, in watts per cps.

S .... Average received signal power during communication period

$\hat{S}$.... Average received signal power during a pulse-observation period

P(x) . Probability distribution of a stochastic function x

R.... Rate of communication of message information, in bits per second

$\hat{R}$.... Rate at which bits may be encoded for transmission over channel

T.... Duration of a transmitted symbol or pulse, in seconds

T_m.. Time dispersion of channel, i.e., maximum differential propagation-path delay

T_s .. Average repetition time of symbol or pulse

W ... Total bandwidth occupied by a system (or subsystem), in cps

Performance of Multipath-Communication Techniques

The rate of communication of information in a digital system may be conveniently used to characterize its performance. To determine R , one must first specify the characteristics of the transmitted signal, of the channel, and of the receiver that is to be used. A suitable analytical approach is to specify only the general form of the transmitted signal and the receiver, and to study the effects on R as variations are made in such parameters as SNR , signal bandwidth and alphabet size; in decision error rates and other internal system-design parameters; and in the statistical properties of the random multipath channel.

The Adaptive Matched-Filter Receiver

The philosophy of the adaptive matched-filter receiver for randomly varying multipath channels with additive random noise has been presented by Price and Green[4], and a rather special experimental binary version has been built and operated. While it has not been established that such a technique is generally optimum, except possibly in the limit as channel-transfer-function measurement errors vanish, it is a potentially valuable technique, and there has been a distinct need for better understanding of its performance characteristics as a function of measurable propagation conditions.

A recent investigation[5] of a very general version of the adaptive matched-filter receiver has led to a better understanding of some of the mathematical steps that are necessary to determine realistically its performance in practical multipath applications. While no attempt is made to present the detailed derivations and proofs in this paper, a description of the methods and of the conditions under which the analysis was made is presented, and a study of some of the direct results will be undertaken.

Description of System

Because the special nature of the experimental Rake[4] multipath system is not well suited to signaling alphabets much larger than binary, consideration has been given to a more general model in which a set of independent, discrete frequency-sounding signals are uniformly spaced throughout the occupied band, as originally suggested by Price[1]. The message-bearing signal components have been restricted to intermediate intervals of frequency in such a manner that no interaction takes place between the channel-sounding function and the direct message-communication function of the system. An enumeration of other applicable conditions and assumptions is made below.

1) A fraction k of total transmitter power is devoted to the message-bearing components of the signal.

2) All short-term information pertaining to the time-varying, channel-transfer function is obtained by cross-correlating the received signal with a stored replica of the sounding signal at appropriate increments of delay. Hence, the available channel information is independent of which message (or symbol) is transmitted.

3) The transmitted message-signal component has maximum randomness (maximum entropy) within the constraints imposed on power and bandwidth occupancy. While the investigation of R for such a signal, has implications similar to those of Shannon's[6] classic formula for communication capacity of an additive noise channel, it appears that the practical limitations imposed by finite alphabet size and duration can be handled in the same way as for a fixed unipath channel with purely additive noise – for example, see Rice[7].

4) Additive noise is gaussian, white within W , and of known power-spectral density.

5) With regard to the channel model representation shown in Fig. 1, no single tap multiplier dominates in its contribution to received signal power.

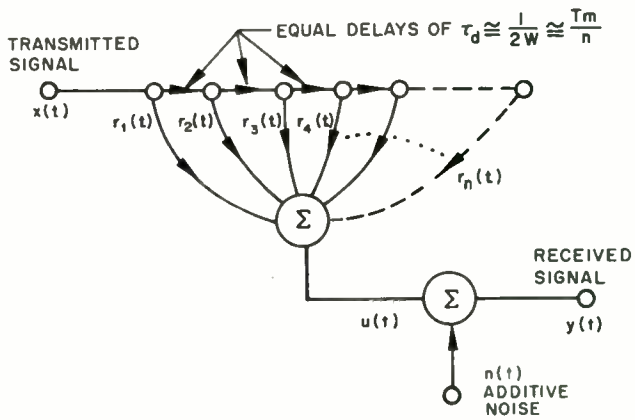


Fig. 1—Uniformly tapped delay-line model of a random multipath channel.

Kailath[8] has shown that an arbitrary linear filtering operation may be represented as closely as desired by a discrete filter model such as is shown in Fig. 1. The randomly time-varying nature of the multipath channel is accommodated by allowing the tap multipliers $r_i(t)$ to be random functions of time. If a practical multipath circuit is to be analyzed correctly in terms of such a model, the $r_i(t)$ must not be restricted as to polarity. The rate of change of the $r_i(t)$ may be characterized by a low-pass frequency spectrum limited to $B_s/2$ cps, which in turn establishes the required bandwidth of filters used for measuring the $r_i(t)$. The approximate length of the tapped delay line, and consequently the number of taps, is directly related to the differential propagation delay of the physical channel it represents.

Analysis of Performance. In the detection procedure used to determine which one of a set $x^{(j)}(t)$, $j = 1, 2, \ldots$, of possible message waveforms has been transmitted, the adaptive matched-filter receiver constructs an approximate replica of the channel delay-line model in accordance with measured values for the functions $r_i(t)$. It is necessary to modify the measured values of tap-multiplier functions in accordance with certain auxiliary weighting functions, to be subsequently discussed, to obtain the best results. The adaptive matched-filter receiver processes each of the $x^{(j)}(t)$ by passing it through a channel-filter replica constructed in the manner described. A decision as to which of the $x^{(j)}(t)$ was transmitted is then made on the basis of which modified symbol is most similar (according to a least-mean-square error criterion) to the received signal $y(t)$.

The need for applying the auxiliary weighting functions is not a consequence of Brennan's[9] theorem relating to primary weighting functions for the optimum combining of noisy signals, as might first be suspected, but goes one step beyond this previous result. The process of matched-filter detection itself inherently satisfied Brennan's requirement that various coherent contributions to a signal having equal noise levels be weighted in amplitude by their respective mean signal-to-noise amplitude ratios. The need for the additional weighting functions to be applied here arises because of practical limitations in the ability to implement Brennan's derived result – the correct values for the primary weighting functions are, in a practical random-channel application, unknown at the receiver. They can be obtained only approximately from uncertain receiver estimates of the SNR ratios of the various signal components to be combined, and thus, an additional corrective factor must be applied to get the best possible results.

Optimum selection of auxiliary weighting functions, which are in general functions of $r_i(t)$ and thus functions of time themselves, is influenced by the particular joint probability distribution applicable to the magnitudes of the $r_i(t)$. It is quite possible, however, that few applications will justify the use of auxiliary weighting functions based on anything more than a characteristic first-order, long-term probability distribution of tap-sample values for the general classification of multipath channel involved.

For the following results, it is assumed that a restricted optimization of the auxiliary weighting functions has been permitted. Specifically, it is assumed that the $r_i(t)$ are of equal variance, are gaussian and independent, and that the auxiliary weighting functions are constrained to be independent of time. Aside from analytical simplifications occasioned by this particular set of assumptions, the following practical implementation advantages arise for a system optimized under the assumed conditions:

1) The assumed $r_i(t)$ statistics appear to be reasonably valid for many types of scatter multipath channels.

2) The system implementation is greatly simplified.

3) The resulting performance calculations for such a system are found to be valid when it is used, without any adjustment or modification, on a wide range of practical ionospheric- and scatter-multipath-propagation circuits.

In connection with the above advantages, it is noted that the Rake system described by Price and Green[4] had little conceptual provision for auxiliary weighting functions. It was possible to apply only a rudimentary on-off control by manually disconnecting delay-line taps when an operator noticed that they were not contributing much to the received signal.

Detailed consideration has been given to the inherent nature and limitations of adaptive, matched-filter techniques in deriving the following bounds on R. For the specified conditions, it has been found that[5]

$$\left(1 - T_m B_s\right) W \log \left[\frac{0.316\, kS + n_o W \left(1 - T_m B_s\right)}{\frac{kS}{1+\gamma} + n_o W \left(1 - T_m B_s\right)}\right] \tag{1}$$

$$\leqq R \leqq \left(1 - T_m B_s\right) W \log \left[\frac{kS + n_o W \left(1 - T_m B_s\right)}{\frac{kS}{1+\gamma} + n_o W \left(1 - T_m B_s\right)}\right]$$

where γ is the mean signal-to-noise ratio obtained in measuring the $r_i(t)$, given by

$$\gamma = \frac{(1-k)\, S}{2 T_m B_s \left[(1-k)\, S + n_o W\right]} . \tag{2}$$

Note that the lower limit of Eq. (1) can go negative, in which case a bound of $R \geqq 0$, arising from the fundamental relation $H(y) \geqq H_x(y)$, is more appropriately employed.

To determine the effects of varying certain system design parameters, it is convenient to express the results in terms of a power requirement factor F reflecting the received signal energy required per bit of received message information, normalized in units of n_o.

$$\text{i.e.,} \quad F = \frac{S}{n_o R} = \frac{W}{R}\left(\frac{S}{n_o W}\right) = \frac{W\, SNR}{R} \tag{3}$$

By combining Eqs. (1) and (2) and expressing the results in the indicated manner, one obtains

$$\frac{\text{SNR}}{(1 - T_m B_s)\log_2 \left\{ \dfrac{k(\text{SNR}) + (1 - T_m B_s)}{\dfrac{k\,(\text{SNR})}{1 + \dfrac{(1-k)\,\text{SNR}}{T_m B_s\left[(1-k)\,\text{SNR} + 1\right]}} + (1 - T_m B_s)} \right\}} \leq F$$

and

$$F \leq \frac{\text{SNR}}{(1 - T_m B_s)\log_2 \left\{ \dfrac{0.316\,k\,(\text{SNR}) + (1 - T_m B_s)}{\dfrac{k(\text{SNR})}{1 + \dfrac{(1-k)\,\text{SNR}}{T_m B_s\left[(1-k)\,\text{SNR} + 1\right]}} + (1 - T_m B_s)} \right\} \qquad (4)$$

The bandwidth allocation required to perform a given communication task in a given time is often of considerable importance. In a later section, some consideration will therefore be given to the following normalized bandwidth-requirement factor, which is readily obtained from Eq. (3).

$$\frac{W}{R} = \frac{F}{\text{SNR}} \qquad (5)$$

Results. An example of the influence that design parameter k has on required signal power is illustrated in Fig. 2. In this figure, and in those to be presented subsequently, a three db degradation factor has been applied to γ to reflect practical limitations in the synthesis of ideal tap-measurement filters.

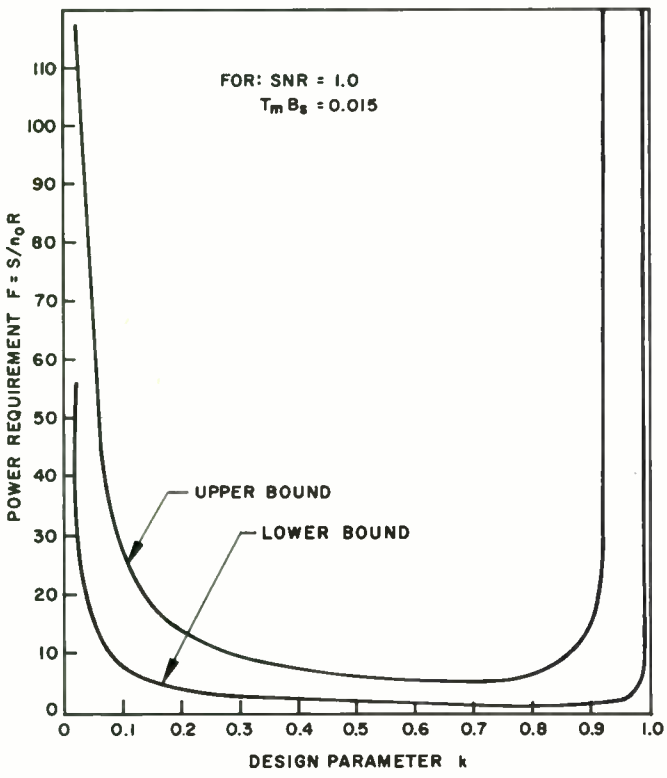


Fig. 2—Power requirement vs. k .

In investigating the nature of system performance bounds as the channel characteristics and the SNR are varied, a large number of curves similar to those of Fig. 2 have been calculated and plotted. The following interesting observations have been made in this process:

- In most cases, the performance bounds are flat over surprisingly wide ranges of k adjacent to the minima. The decrease in message power resulting from greater concentration on the channel-sounding function appears to be almost exactly compensated by the improved channel knowledge in this region. In a few cases, bounds on F change almost imperceptably for k values ranging all the way from 0.1 to 0.9.

- As the product $T_m B_s$ increases (between 0 and 1), the optimum points occur for smaller values of k , indicating that more effort should be devoted to the sounding function if an adaptive matched-filter technique is to be used. Similarly, the minima occur at larger values of F , indicating the increased degradation caused by the channel.

- There are cases where the bounds exhibit very slight but distinct double minima as k is varied.

- The minima for the upper and lower bounds in any given case may occur for values of k that are quite widely separated, but this result is of little significance in view of the flatness of the curves.

In Figs. 3 through 6 the minimum points from Fig. 2 and numerous other similar sets of curves are plotted as a function of SNR for various values of the product $T_m B_s$.

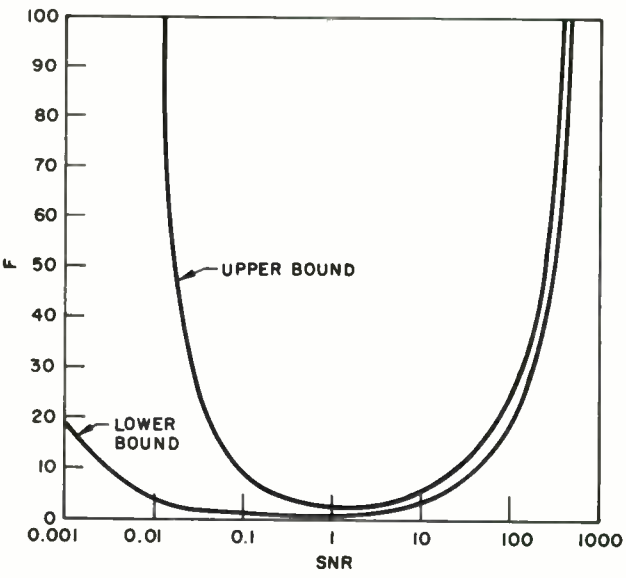


Fig. 3—Bounds on F for $T_m B_s$ = 0.003.

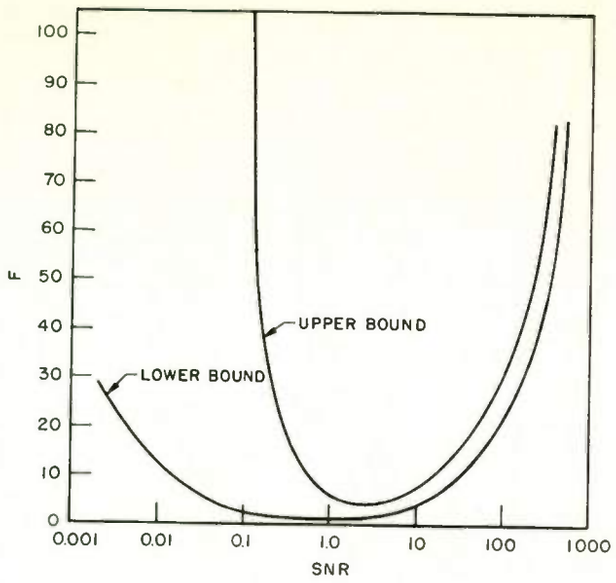


Fig. 4—Bounds on F for $T_m B_s = 0.015$.

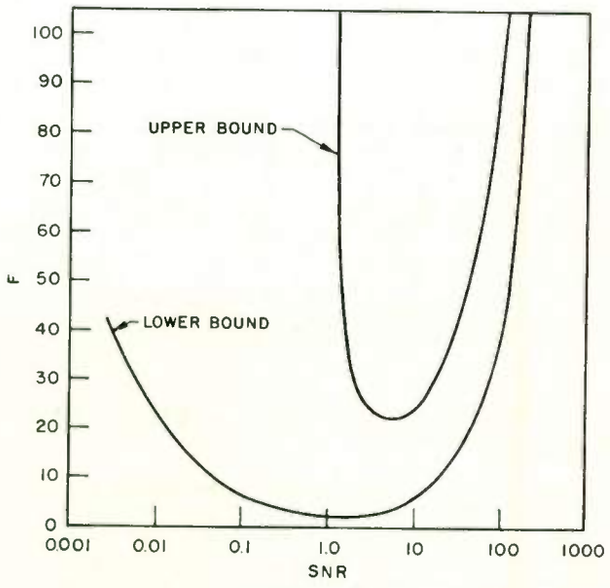


Fig. 5—Bounds on F for $T_m B_s = 0.075$.

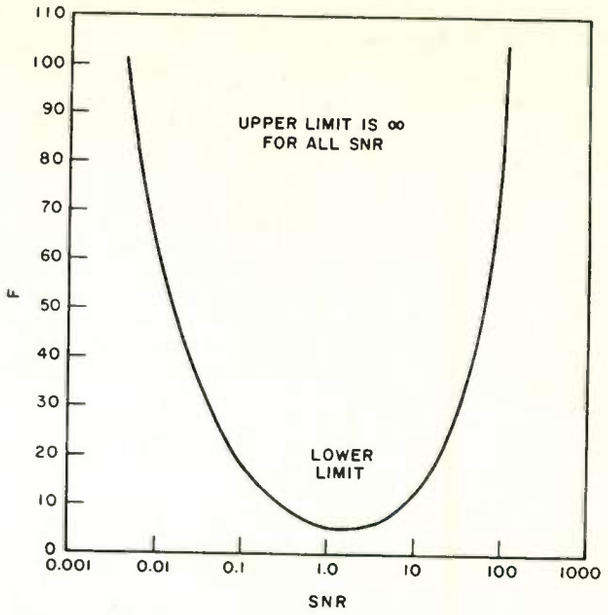


Fig. 6—Bounds on F for $T_m B_s = 0.25$.

A degree of conservatism is recommended in interpreting the foregoing sets of curves. Though considerable confidence can be placed in the upper bound on F, it should be remembered that the effects of alphabet limitations[7] have not yet been introduced. Further, in order to achieve a desired rate of communication with a limited alphabet, symbol durations short relative to T_m may be indicated. Under such circumstances, special techniques[4,5] may be required to avoid coherent intersymbol influences. In spite of these adverse factors, and in spite of the optimization restrictions imposed on the system that has been analyzed, the foregoing results reflect very favorably on the use of adaptive matched-filter receiving techniques for random-multipath and scatter-multipath channels having moderate $T_m B_s$ products.

It appears that the analytical techniques described may be readily extendable to situations involving greater freedom to optimize the system for the particular characteristics of various types of channels. It is quite likely that such a step will give better insight into the potential communication value of unusual types of propagation modes, and that it will suggest practical ways of exploiting detailed statistical knowledge that can be made available.

Narrow-Band Techniques

In this section, the performance potentials of some relatively simple but appropriate, narrow-band, multipath techniques will be determined and expressed in a manner that can be compared with other types of systems. Provided the bandwidth of signals is small relative to $1/T_m$, intersymbol influence problems due to time dispersion are minor and may be avoided by using special filtering techniques[10] and sacrificing a small proportion of received signal power. For a differential phase-modulation system, channel-amplitude variations affect the SNR and, hence, cause error-rate fluctuations,

but do not otherwise influence the receiver-decision process. The accompanying narrow-band, channel-phase fluctuations contribute direct errors to the differential phase measurement between successive symbols, but this effect may be kept very small provided $T_m B_s$ is much less than unity.[11]

The effects of random multipath fading on these systems may be determined approximately by averaging the error rate over the expected distribution of received-signal amplitudes. This may be done conveniently by defining a rate $\hat{R}$ at which message bits are transmitted, which differs from the maximum rate of communication of information over the system by the equivocation due to receiver decision errors. It is useful to employ a corresponding power-requirement factor

$$\hat{F} \triangleq \frac{S}{n_o \hat{R}} = \frac{\text{Received energy per transmitted bit}}{n_o} \qquad (6)$$

For binary, polarity-modulated, sinusoidal signals with additive gaussian noise, the dependence of error rate on $\hat{F}$ is expressed by the simple relation[12]

$$\mathcal{E} = 0.5\, e^{-C\hat{F}} \qquad (7)$$

with C equal to unity. A similar exponential expression with $C = 0.551$ is found to give the corresponding relation for a $\pi/2$ differential-phase-shift system operating as two co-channel, binary circuits[10] — a system which makes more effective use of available bandwidth, but which is accordingly more susceptible to the influence of channel phase fluctuations. An ideal binary system (assuming availability of a perfect phase reference, which cannot be obtained for the random-channel model) has a theoretical, error-rate curve that can be approximated quite accurately by a similar expression with $C = 1.242$.

The Rayleigh fading channel yields a distribution for $\hat{F}$ given by

$$p(\hat{F}) = \frac{1}{\hat{F}_o}\, e^{-(\hat{F}/\hat{F}_o)} \qquad (8)$$

where $\hat{F}_o$ corresponds to mean strength of the fading signal. The average error rate, given by

$$\bar{\mathcal{E}}_{\text{Rayleigh fading}} \cong \int_0^\infty \mathcal{E}(\hat{F})\, p(\hat{F})\, d\hat{F} \qquad (9)$$

may be readily evaluated for functions of the form of Eqs. (7) and (8), yielding,

$$\bar{\mathcal{E}}_{\substack{\text{Rayleigh} \\ \text{fading}}} \cong \frac{0.5}{C(\hat{F}_o + 1)} \qquad (10)$$

Several error rate curves obtained in the manner described and from typical experimental measurements[10] are plotted in Fig. 7.

To obtain the desired relations between energy and information transferred through the channel, one must take into account error equivocation. For a symmetrical

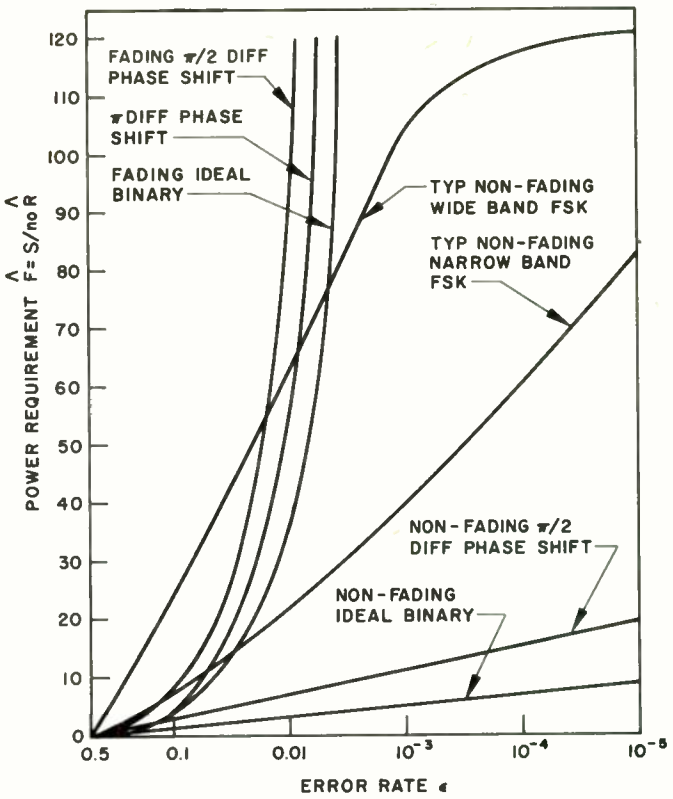


Fig. 7 — Relations between symbol energy and error rate.

digital system with alphabet size m, error probability $\mathcal{E}$, and independent occurrence of errors one has

$$R = \frac{1}{T}\left[H(x) - H_y(x)\right]$$

$$= \hat{R}\left\{1 - \left[\frac{(1-\mathcal{E})\log\left(\frac{1}{1-\mathcal{E}}\right) + \mathcal{E}\log\left(\frac{m-1}{\mathcal{E}}\right)}{\log(m)}\right]\right\} \qquad (11)$$

In Fig. 8, the curves of Fig. 7 have been modified to reflect the relation given in Eq. (11).

The assumption of independent, random-error occurrence in obtaining performance curves yields results for the fading cases that are pessimistic in some respects. The effect of errors, given by the second term of Eq. (11), may be considerably less than that calculated, for two reasons. In the first place, a degree of error dependence is introduced by the common level of fading experienced by adjacent symbols, lessening the conditional entropy of the channel. Considerable progress has been made, for example, in devising systematic error-correcting codes to take advantage of error clustering.[13, 14] In the second place, the receiver may derive a symbol-by-symbol estimate of the time-varying error probability function from observations of the amplitude of received signals and noise.

It is clear from Fig. 8 that an individual narrow-band fading channel must be operated at a relatively high error rate in order to optimize power requirements. Presently available error-coding techniques for reducing such error rates to the much lower values usually desired in practice are not very efficient, but greater application of advances in digital data-handling and storage capabilities can be expected to improve the situation, especially where error dependence and error knowledge of the kind described are more fully exploited.

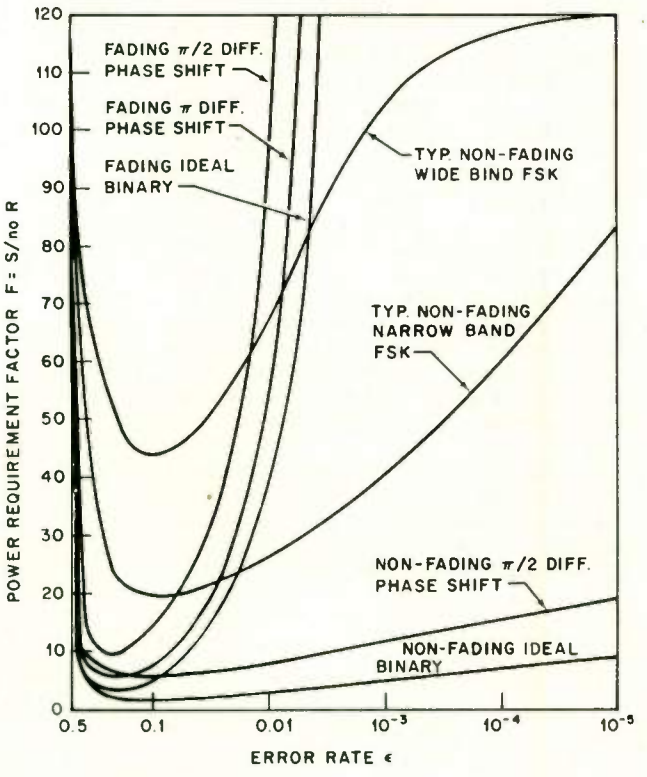


Fig. 8—Power requirement factor vs. error rate.

It should be noted that the common forms of post-detection diversity reception used for digital communication over fading channels are elementary, highly redundant forms of parallel error coding. For a given effective antenna aperture, they offer little special advantage over purely serial coding except in the elimination of information storage requirements in coding and detecting. Other than for some special power considerations involving scattering volumes and antenna gains on some circuits[15], themathematical treatment of parallel diversity schemes is the same as that of purely serial (pure time diversity) error-correction coding on a single, narrow-band channel, and thus needs no special consideration.

Incoherent Detection Techniques

As the product of differential propagation time delay and doppler spreading bandwidth of a channel increases toward and beyond unity, most communication techniques tend to become less satisfactory and, eventually, unworkable. It becomes impossible at the receiver to learn anything very specific about the transmitted signal from observations made of the received signal. A cause-and-effect relation still exists, however, between the energy of a transmitted symbol and that observed at the receiver in a corresponding time interval and doppler- spread frequency band. A promising technique[15] under such circumstances involves sinusoidal signals with pulse-time and/or frequency-shift modulation, with the shift size selected so that the energies of different possible received pulse locations may be independently measured.

If mean received power is $\hat{S} + N$ during a pulse observation interval T that is long compared to channel fading rate B_s, the observed energy may be represented by a sum x of $n = 2TB_s$ equally spaced instantaneous power observations. For Rayleigh fading sinusoidal signals in gaussian additive noise, this sum is described by the Chi Squared distribution,

$$P_{\hat{S}+N}(x) \cong \frac{x^{\frac{n}{2}-1} \exp\left[-\frac{x}{2(\hat{S}+N)}\right]}{\left[2(\hat{S}+N)\right]^{n/2}\Gamma(n/2)} \quad . \qquad (12)$$

A similar expression with variance N describes the corresponding sum of power observations when no signal pulse is present.

$$P_N(x) \cong \frac{x^{\frac{n}{2}-1} \exp\left[-\frac{x}{2N}\right]}{(2N)^{n/2}\Gamma(n/2)} \qquad (13)$$

At the receiver, there will be m energy measuring circuits corresponding to the product of the number of possible arrival times and the number of possible frequencies available to each pulse. The largest energy occurring among the $m - 1$ measurements of noise alone has a probability distribution corresponding to

$$P_{N_{max}}(x) = (m-1)\,P_N(x)\left[\int_0^x P_N(\gamma)\,d\gamma\right]^{m-2} \quad . \qquad (14)$$

A receiver decision error occurs whenever the largest noise observation exceeds an observation of signal plus noise. Thus

$$\mathcal{E} = \int_0^\infty \left[P_{\hat{S}+N}(x)\int_x^\infty P_{N_{max}}(\beta)\,d\beta\right] dx \quad . \qquad (15)$$

For a symmetrical modulation system, information may be encoded for transmission at a rate

$$\hat{R} = \frac{1}{T_S}\log_2(m) \quad . \qquad (16)$$

The bandwidth of a transmitted pulse is approximately $1/T$, and the corresponding noise power accompanying a doppler-spread received pulse is

$$N \cong n_o\left(\frac{1}{T} + B_s\right) = \frac{n_o}{T}\left(1 + \frac{n}{2}\right) . \qquad (17)$$

Average received signal power S is given by

$$S = \hat{S}\,\frac{T}{T_s} \quad . \tag{18}$$

By defining pulse $\overline{SNR} = \hat{S}/N$, one obtains

$$\hat{F} = \frac{S}{n_o R} = \frac{\hat{S}T}{n_o \log_2(m)} = \frac{\overline{SNR}\left(1+\frac{n}{2}\right)}{\log_2(m)} \quad . \tag{19}$$

Given a value of $\overline{SNR}$, the number n of independent power observations per pulse, and an alphabet size m, one can determine $\mathcal{E}$ and $\hat{F}$ from the relations presented. Some computed results of $\hat{F}$ vs. SNR for various combinations of m and $\mathcal{E}$ are shown in Fig. 9.

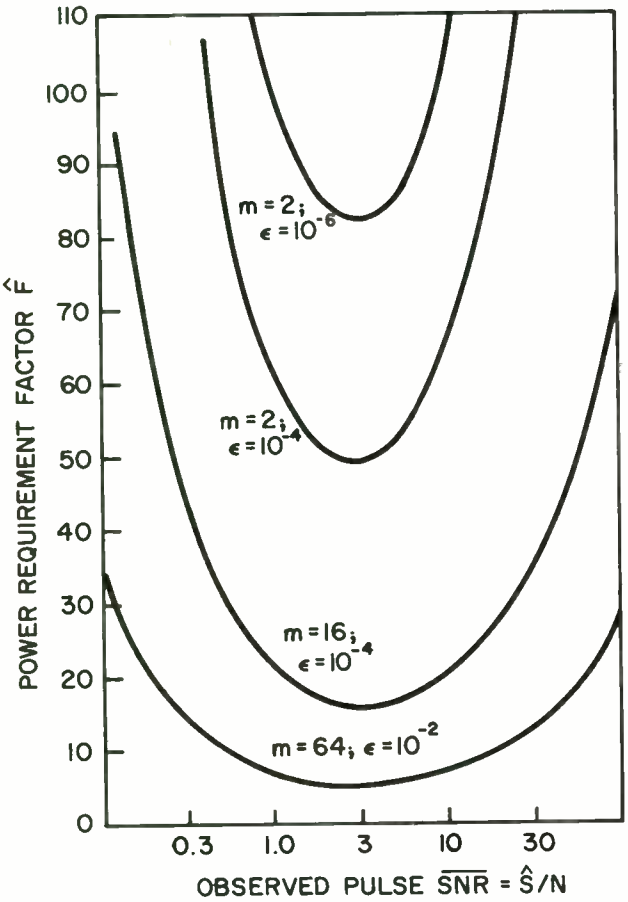


Fig. 9—Power requirement factor $\hat{F}$ vs. $\overline{SNR}$.

The curves of Fig. 9 could have been plotted as a function of the number of independent observations n per pulse (corresponding roughly to the order of time diversity used in reception) instead of as a function of $\overline{SNR}$, since the two parameters are directly related for given values of $\hat{F}$ and m through Eq. (19).

The curves of Fig. 9 illustrate a basic characteristic(that may not have been recognized heretofore) of most post-detection diversity techniques. These particular results were obtained, in effect, for a general m^{th} order orthogonal alphabet system with n^{th} order time diversity (it would be $n/2^{th}$ order if envelope observations were made). Except for increasing pulse time dispersion difficulties, these results apply equally well to an m^{th}-order alphabet system with q^{th}-order frequency diversity and n/q^{th}-order time diversity. The basic observation to be made about such systems is that no matter what error rate is involved, nor what alphabet size is used, and no matter what combination of time-and-frequency diversity (and even space diversity for a fixed total-antenna aperture) is to be used, the minimum received energy requirement per transmitted bit occurs when the SNR of received signals in each diversity channel is very nearly three. This same result arises implicitly, for example, in an investigation of the optimum order of diversity for a binary system by Pierce,[16] where division of average "total SNR" by the "optimum number of diversity branches" yields a result very close to the magic number three in every case, even though the analytical approach to the problem is quite different from that employed here. By way of qualification, it should be mentioned that the results of this investigation give minimum F for $\overline{SNR}$ values somewhat less than three in a few very special cases involving systems with very small alphabet size, when operating at very high error rates.

If the minimum values of $\hat{F}$ from a large number of performance curves, such as those of Fig. 9, are modified to reflect error equivocation in accordance with Eq. (11) and plotted as a function of $\mathcal{E}$, the results appear as shown in Fig. 10.

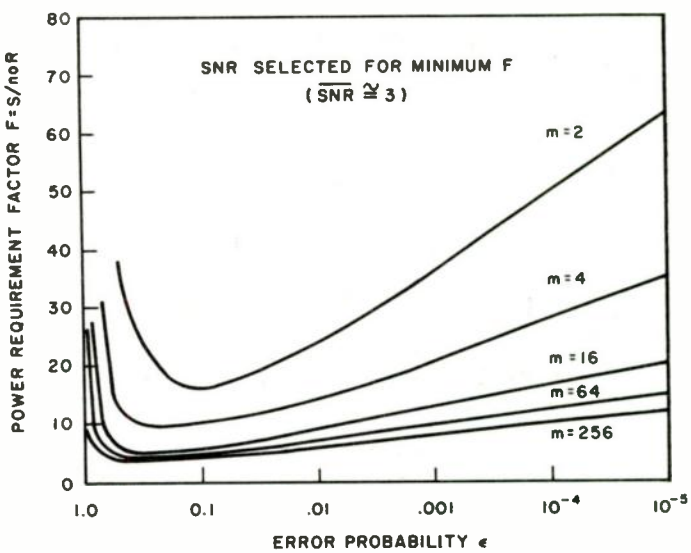


Fig. 10—Minimum power requirement factor F vs. error rate.

The degradation associated with operation at SNR values other than three, for given m and $\mathcal{E}$, may be determined to a good degree of approximation from the empirical curve shown in Fig. 11.

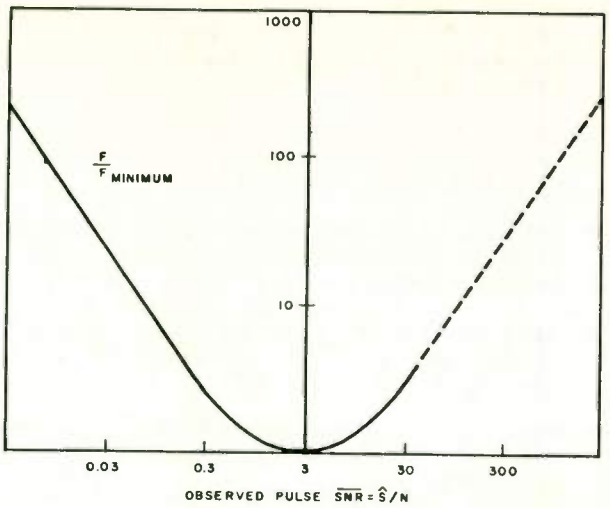


Fig. 11—Normalized power requirement factor vs. $\overline{SNR}$.

Under most desirable operating conditions, the number n of signal observations may be made quite large without sacrificing power. From the equation,

$$T = \frac{n}{2B_s} = \left(\frac{n/2}{T_m B_s}\right) T_m \quad , \qquad (20)$$

one observes that T may be made much greater than T_m for $T_m B_s > 1$, thus effectively reducing degradation caused by multipath time dispersion of transmitted pulses.

System Bandwidth Requirements

Emphasis so far has been placed on minimizing power requirements for transmission of information. In many practical applications, however, a need for efficient use of available frequency allocations may force one to compromise power efficiency. To assist in making such a compromise, it is helpful to determine the power requirement for various techniques as a function of the total bandwidth occupied per unit rate of information transfer. Curves illustrating this functional dependence are shown in Fig. 12 for some of the techniques and conditions that have been considered.

The curves given for adaptive matched-filter techniques correspond to the lower bound on information rate given in Eq. (1). In the case of incoherent detection techniques, it is assumed that the modulation system is designed to take full advantage of all available space in a frequency-time plane. Any failure to do so causes a corresponding increase in W/R. Under the assumed conditions, total incoherent system bandwidth,

$$W = m\left(B_s + \frac{1}{T}\right)\frac{T}{T_s} \quad ,$$

and

$$\frac{W}{R} = \left[\frac{m\left(B_s + \frac{1}{T}\right)T}{T_s}\right]\left(\frac{n_o F}{S}\right) = \frac{mF}{\overline{SNR}} \quad . \qquad (21)$$

In the case of narrow-band differentially coherent modulation schemes without diversity, varying SNR affects the bit-error rate, and hence R, even though the symbol transmission rate and bandwidth remain essentially constant for the curves given.

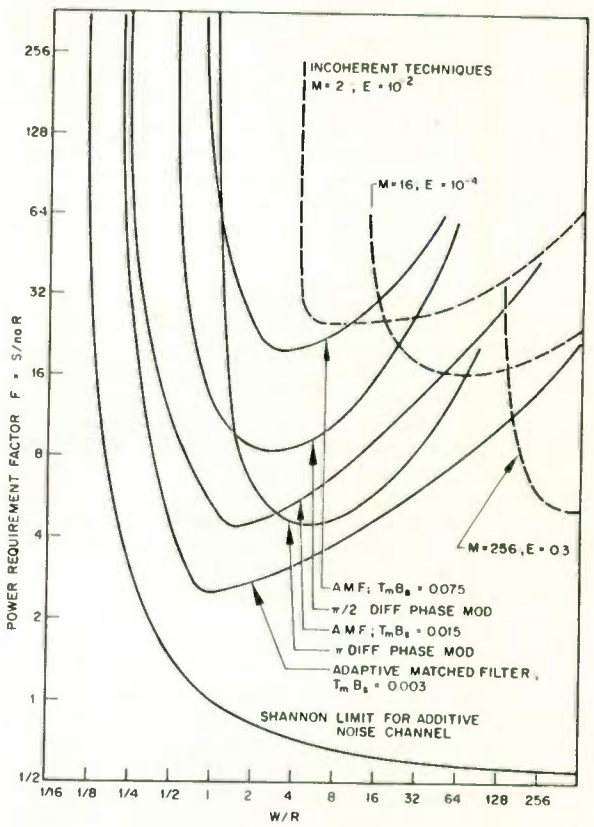


Fig. 12—Power requirement vs. bandwidth requirement.

The Shannon capacity limit shown in Fig. 12 is obtained for a nonvarying channel with additive white gaussian noise, in which case

$$F = \frac{S}{n_o R} \geqq \frac{SNR}{\log_2 (1 + SNR)} \qquad (22)$$

and

$$\frac{W}{R} = \frac{F}{SNR} \geqq \frac{1}{\log_2 (1 + SNR)} \quad . \qquad (23)$$

It is interesting to try to locate an operating point for the experimental Rake[4] multipath communication technique in the chart of Fig. 12. The abscissa under

normal operating conditions, would be 440, which falls nearly off the chart. Sufficient information is not available at this writing to determine an ordinate accurately, but a rough estimate, based on limited experimental comparison data, places a Rake operating point well above the upper bound given for the adaptive matched-filter system. A value of $T_m B_s$ of 0.003 is approximately correct for the conditions under which the experimental tests were run.

Conclusions

When interpreting the results that have been presented, one should not lose track of the limitations and restrictions under which various types of systems have been analyzed. In particular, note that specific alphabet and/or error-rate constraints have been imposed in obtaining results for some of the techniques, while the curves for the adaptive matched-filter receiver imply a much greater design latitude. The methods and approach used to compare techniques, and the general trends indicated are perhaps more significant in many cases than the exact numerical values obtained. A few generalizations that appear to be valid are as follows:

- Conservative appraisal of the performance potentials of adaptive matched-filter techniques indicates that attractive results may be achieved when the product $T_m B_s$ for a communication circuit is reasonably small. For somewhat larger values of $T_m B_s$, it is possible that appropriately designed narrow-band techniques may be reasonably competitive.

- To do an adequate channel-sounding job for an adaptive matched-filter receiver, it appears that one should avoid spreading the signal power over an excessively wide band of frequencies.

- Incoherent detection techniques, requiring no measurement of channel transfer function, may provide efficient use of available power even for channels with $T_m B_s$ considerably greater than unity, but bandwidth allocations much greater than those normally needed for other techniques will be required. From a standpoint of mere power conservation, it appears that the incoherent techniques can be designed to be reasonably competitive with adaptive matched-filter techniques even where $T_m B_s$ is very small.

It is hoped that more and better performance data will become available for various theoretical and practical solutions to the random multipath-communication problem as time goes by. Only in this way can the merits of such techniques be appraised and the full benefits of extensive research and development efforts be received.

Acknowledgment

The author gratefully acknowledges the assistance of Mr. Daniel G. Drath in programming and running the computations necessary for parts of this report. In addition, appreciation is expressed to Professors A. M. Peterson and Von R. Eshleman of Stanford University and to Dr. Forrest R. Fulton, Jr. for helpful discussions bearing on some of the problems.

References

1. Price, R., Statistical Theory Applied to Communication Through Multipath Disturbances, M.I.T. R.L.E. T.R. No. 266; September 3, 1953.

2. Turin, G. L., Communication Through Noisy, Random-Multipath Channels, M.I.T. Lincoln Laboratory, T.R. No. 116; May 14, 1956.

3. Kailath, T., "Correlation Detection of Signals Perturbed by a Random Channel," I.R.E. P.G.I.T. IT-6 (June, 1960), pp. 361–366.

4. Price, R., and Green, P. E., "A Communication Technique for Multipath Channels," Proc. I.R.E. 46 (March, 1958), pp. 555–570.

5. Harris, D. P., Communication Through Random Multipath Media, Stanford University T.R. No. SEL-62-031; April, 1962.

6. Shannon, C. E., "Communication in the Presence of Noise," Proc. I.R.E. 37 (January, 1949), pp. 10–21.

7. Rice, S. O., "Communication in the Presence of Noise," B.S.T.J. 29 (January, 1950), pp. 60–93.

8. Kailath, T., Sampling Models for Linear Time-Variant Filters, M.I.T. R.L.E. T.R. No. 352; May, 1959.

9. Brennan, D. G., "On the Maximum Signal-to-Noise Ratio Realizable from Several Noise Signals," Proc. I.R.E., 43 (October, 1955), p. 1530.

10. Doelz, M. L., Heald, E. T., and Martin, D. L., "Binary Data Transmission Techniques for Linear Systems," Proc. I.R.E. 45 (May, 1957), pp. 656–661.

11. Juda, H. J., and Skalbania, F. R., "Frequency Error Effects on a Binary Communication System Utilizing Coherent Integrators," Record of the Seventh National Communication Symposium, Utica, N. Y., October 2–4, 1961.

12. Lawton, J. G., "Comparison of Binary Data Transmission Systems," I.R.E.-P.G.M.E. Proceedings of the Second National Convention on Military Electronics, (June, 1958), pp. 54–61.

13. Abramson, N. M., "A Class of Systematic Codes for Non-Independent Errors," I.R.E. P.G.I.T. IT-5 No. 4 (December, 1959), pp. 150–156.

14. Reiger, S. H., "Codes for the Correction of Clustered Errors," I.R.E. P.G.I.T. IT-6 No. 1 (March, 1960), pp. 16–20.

15. Harris, D. P., "Techniques for Incoherent Scatter Communication," Paper presented at I.R.E. WESCON Convention, San Francisco, California, August 22–25, 1961.

16. Pierce, J. N., "Theoretical Limitations on Frequency and Time Diversity for Fading Binary Transmissions," I.R.E. P.G.C.S. CS-9 No. 2 (June, 1961), p. 186.