THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XLV MARCH 1966 NUMBER 3

Digital Computer Simulation of Sampled-Data Communication Systems Using the Block Diagram Compiler: BLODIB

By ROGER M. GOLDEN

(Manuscript received November 23, 1965)

Digital computer simulation of communication systems is accomplished readily by means of the system-oriented programming language called BLODIB (for BLOck DIagram Compiler, B). The language is designed for programming sampled-data systems which may be represented either in block diagram form or in the mathematical representation of the z-transform calculus. Contained within the language are 40 basic "building" blocks from which an entire communication system can be built. In addition, new blocks may be defined as consisting of complex configurations of basic blocks. The structure of BLODIB allows convenient specification of system parameters as well as permitting these parameters to be varied in order to study changes in system performance. The use of BLODIB is demonstrated by its application in the simulation of a voice-coding (vocoder) system.

I. INTRODUCTION

The study of complex communications systems by digital computer simulation is facilitated by use of the Block Diagram Compiler, BLODIB.[1] The source language accepted by the compiler is oriented towards communication system engineers and analysts who are already adept at handling block diagrams (or transfer functions) of complete systems. BLODIB is easily learned even by persons who have had little or no previous programming experience. The language of the compiler was designed specifically for simulating a wide variety of sampled-data

system problems as well as simulating sampled-data approximations to continuous (analog) systems. The capabilities of BLODIB make possible the simulation of systems which were impossible for the old BLODI[2] compiler to handle. In particular, simulation of voice-communication and speech bandwidth compression systems such as vocoders, are accomplished more easily. The digital computer simulation of these systems was pioneered at Bell Telephone Laboratories using the original BLODI language.[3,4] While such systems were programmed in a few weeks time with a minimum of difficulty, it is now possible to program the same systems in about one day.

II. NEED FOR SIMULATION

The fantastic growth of the demand for information handling and processing makes it absolutely essential that our communication systems be as efficient as technically possible. This means that all forms of information exchange including speech communications should be encoded in such a way as to assure the most efficient use of a communication or data channel. The complexity of terminal equipment associated with such coding systems is such that a great deal of design, testing, and redesign must take place before these systems can be used on a regular basis. The testing and debugging (including redesign where necessary) of such equipment requires many costly man-hours. Much of this cost may be reduced or even avoided if the entire system is first evaluated by simulation on the digital computer. This technique provides the flexibility for systematically optimizing system design. The advantage of using a digital computer as a system simulator lies principally in the ease with which system parameters can be changed. Hence, many designs may be tested before choosing the one that would be built for the actual system. It is just this technique that is now being applied to the design and evaluation of voice-communication systems.

III. PREREQUISITES FOR COMPUTER SIMULATION

Three prerequisites must be completed in order to accomplish a digital computer simulation of a voice-communication system. The first is the determination of an appropriate sampled-data (or digital) representation for the system to be studied; the second is the preparation and compilation of a BLODIB computer program which causes the computer to perform the same operations as would be performed by the actual system; and the third prerequisite is the digitalization of real speech for processing by the computer.

These prerequisites will now be treated with respect to a specific voice-communication system for bandwidth compression. This system is more commonly known as a vocoder (an acronym for voice-coder). The vocoder, invented over 35 years ago by Homer Dudley[5] at Bell Telephone Laboratories, codes voices or speech signals for transmission over a communication channel. Parameters describing the speech are transmitted to a receiver where they are used to synthesize a replica of the original speech sounds. Bandwidth compression is achieved because the bandwidth required to transmit these parameters is considerably less than that required to transmit the original speech sounds. It is helpful at this point to give a brief review of vocoder operation before describing how the above prerequisites for simulation are accomplished.

Fig. 1 shows a functional block diagram of a spectrum channel vocoder and in itself can be used as an aid in the writing of the simulation program. The vocoder consists of two main parts: an analyzer for extracting that information which must be transmitted, and a synthesizer for reconstituting the original speech signal.

The analyzer portion of the vocoder consists of two basic parts: a short-time frequency spectrum analyzer and an excitation detector. The spectrum analyzer consists of a number of bandpass filters for separating contiguous bands of frequencies. The output of each analyzing filter is rectified and then smoothed by low-pass filters. The outputs of the smoothing filters are low-frequency signals which represent the short-

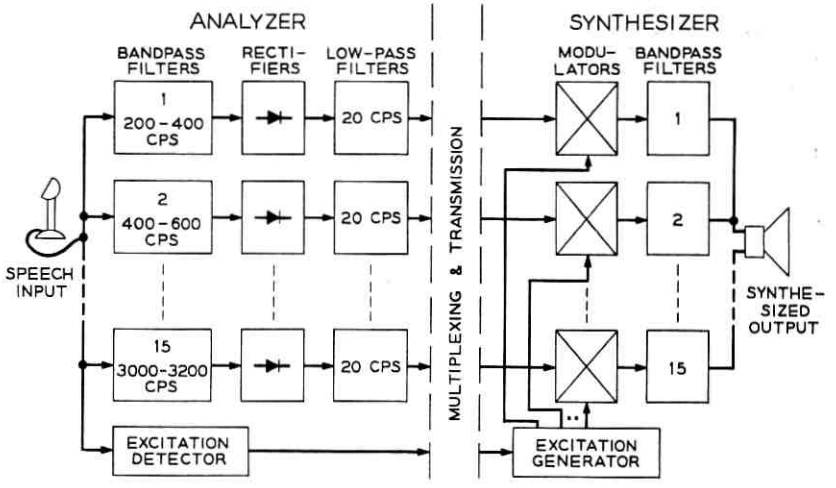


Fig. 1 — Functional block diagram of vocoder.

time spectral energy in each of the analyzed bands. These signals represent the slowly varying characteristics of a speech signal. Physically, they are just the average energy of the signals from the bandpass filters.

The output of the excitation detector (be it conventional pitch detector with voiced-unvoiced decision or a narrow band of the input speech) is transmitted by a suitable frequency or time multiplexing method along with the slowly varying spectrum signals. This information is used by the synthesizer portion of the vocoder to regenerate the input speech.

Synthesis is accomplished by amplitude modulating the frequency harmonics of a locally generated excitation signally the corresponding low-frequency signals. The harmonics of the excitation function are obtained from a set of filters (not shown in Fig. 1) similar to those used for analyzing the input speech signal. The modulated signals are then band limited in order to remove spurious sidebands introduced by the modulation process. The combined output of these filters is a replica of the input speech. Hence, wideband speech is synthesized from a set of low-frequency signals having a relatively narrow over-all bandwidth. This, then, is the complex communication system for which a suitable sampled-data representation is sought.

IV. SAMPLED-DATA REPRESENTATION OF A CONTINUOUS SYSTEM

In order that a suitable linear time-invariant sampled-data representation be found for the vocoder shown in Fig. 1, it is necessary that the transfer functions be known for the various blocks in the system. The representation of these transfer functions may be in either the time domain (a relation between the input time signal and the output time signal) or the frequency domain (a relation between the frequency spectrums of the input and output signals). For example, the various bandpass and low-pass filters are first represented by their frequency domain transfer functions in the form

$$H(s) = \frac{C_0 \prod_{k=1}^{K} (s - \alpha_k)}{\prod_{l=1}^{L} (s - \beta_l)} \tag{1}$$

where

C_0 = constant,
α_k = complex zeros of $H(s)$,
β_l = complex poles of $H(s)$,

K = degree of numerator,
L = degree of denominator, and
s = complex Laplace transform variable.

Conversion of this type of function to an appropriate sampled-data form is accomplished by applying either the z-transform or bilinear z-transformation[6,7] to the above transfer function.

The resulting z-transfer function is then written as a partial fraction expansion in the form,

$$\mathcal{H}(z) = \sum_{m=1}^{M} \frac{A_{1_m} z^{-1} + A_{0_m}}{B_{2_m} z^{-2} + B_{1_m} z^{-1} + 1} \tag{2}$$

where

A_{1_m}, A_{0_m}, B_{2_m}, B_{1_m} are constants,
z^{-1} = exp $(-sT)$ = the unit delay operator, and
T = unit sampling interval.

Each term of the above expansion is then represented in block diagram form as shown in Fig. 2(a). Fig. 2(b) shows the realization of the complete transfer function. This technique is used to realize all of the linear transfer functions in the system. (Incidentally, the design and

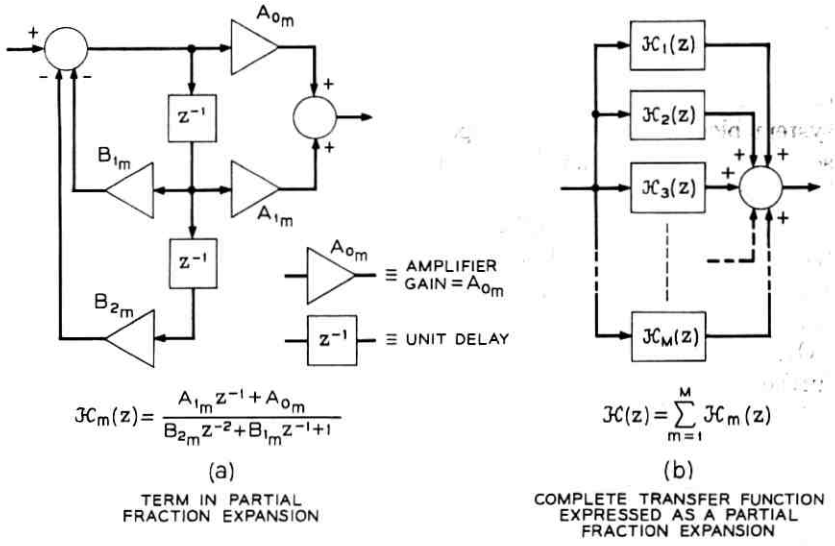


Fig. 2 — Block diagram representation for a partial fraction expansion of a sampled-data transfer function.

BLODIB realization of such functions is readily carried out by digital computer. The output from such a design program is illustrated in the following section.

The function performed by the rectifier is

$$r_{\text{out}}(t) = | r_{\text{in}}(t) | . \tag{3}$$

The sampled-data representation of this function in the time domain is simply

$$r_{\text{out}}(nT) = | r_{\text{in}}(nT) | . \tag{4}$$

The "chopper-modulators" perform the function

$$m_{\text{out}}(t) = m_{\text{in}_1}(t) \cdot \text{sgn } (m_{\text{in}_2}(t)) \tag{5}$$

which is realized by

$$m_{\text{out}}(nT) = m_{\text{in}_1}(nT) \cdot \text{sgn } (m_{\text{in}_2}(nT)). \tag{6}$$

Similarly, appropriate sampled-data representations were made for the excitation detector and generator. This, then, completes the first prerequisite for computer simulation. Next follows the presentation of a BLODIB computer program using the above information.

V. THE BLODIB PROGRAMMING LANGUAGE

As stated earlier, the BLODIB[1] programming language is system oriented in that it allows a direct representation of a communication system block diagram. More specifically, the language is a verbal description of the various blocks in the diagram and information concerning how the blocks are interconnected. In addition, BLODIB makes possible the "construction" or definition of new types of blocks (using facilities called MACROS or SUPERS) from the basic types available. (Special boxes not in the basic set of 40 may also be introduced by supplying an external algorithm coded as a FAP subroutine or as a FORTRAN function. However, this technique will not be discussed further here.)

The use of the MACRO facility as well as the simplicity of BLODIB coding can be demonstrated easily with respect to the realization of the filter transfer functions. Fig. 3 shows the actual BLODIB configuration for the basic terms required in the partial fraction expansion given by (2). Since this "term" or function is required many times, it will be defined as a new type — ADB — and thereafter used as a basic building block. Coding in BLODIB requires giving each block a name (chosen by

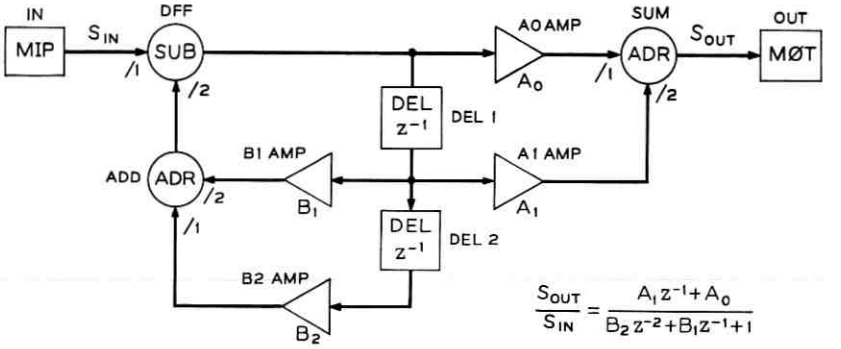


Fig. 3 — BLODIB representation for the transfer function
$\mathfrak{C}(z) = (A_1 z^{-1} + A_0)/(B_2 z^{-2} + B_1 z^{-1} + 1)$

the programmer); designating the type of block from the list of basic types, specifying any parameters associated with the block; and finally, listing the names of other blocks to which the output should be connected. Thus, the actual coding for Fig. 3 is given in Table I. The first line of coding defines ADB as a SUPER which may have up to 4 inputs (I1, I2 etc.) and has 4 parameters A1, A0, B2, B1. The boxes MIP and MOT are required input/output boxes for SUPERS. The term END designates the end to the coding for ADB. The order of appearance of the boxes within the SUPER is immaterial since the compiler builds internally a connection matrix from which the blocks are ordered to produce an efficient program.

The partial fraction expansion terms in the various filter transfer functions are then realized using this definition for ADB. Since all of the filters, bandpass and low-pass, are used in several places, these

TABLE I — BLODIB CODING FOR THE TRANSFER FUNCTION
$\mathfrak{C}(z) = (A_1 z^{-1} + A_0)/(B_2 z^{-2} + B_1 z^{-1} + 1)$

ADB	MACRO	I1, I2, I3, I4, A1, A0, B2, B1
IN	MIP	1, DFF
DFF	SUB	A0AMP, DEL1
DEL1	DEL	1, B1AMP, A1AMP, DEL2
DEL2	DEL	1, B2AMP
B1AMP	AMP	B1,, ADD
B2AMP	AMP	B2,, ADD/2
ADD	ADR	DFF/2
A0AMP	AMP	A0,, SUM
A1AMP	AMP	A1,, SUM/2
SUM	ADR	OUT
OUT	MOT	
	END	

too are coded as SUPERS. Hence, wherever a filter is required, it can be referenced as a single block. In this way, a complete set of filter blocks is built up quickly from the basic BLODIB language.

Figs. 4 and 5 show, respectively, the frequency and time response of one of the bandpass filters required in the simulation. (These graphs along with the BLODIB programming required for simulation were produced by a special computer program for determining sampled-data filter transfer functions. The block diagram for this filter is shown in Fig. 6. Table II presents the actual BLODIB coding.

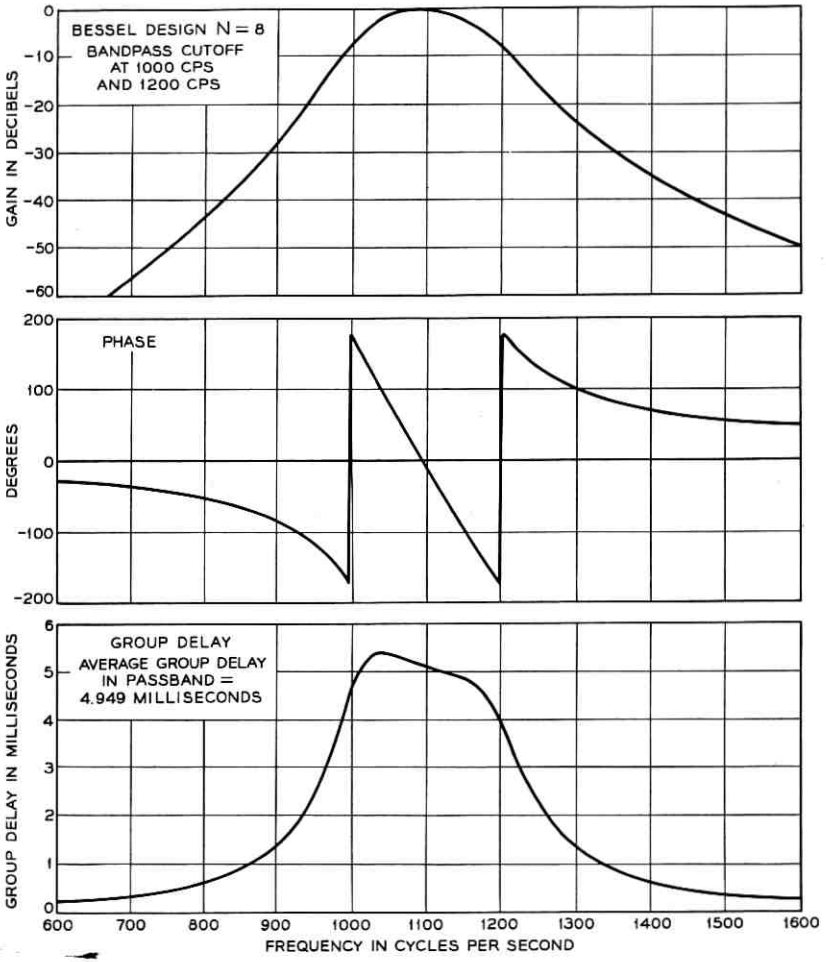


Fig. 4 — Frequency response for a vocoder-channel bandpass filter.

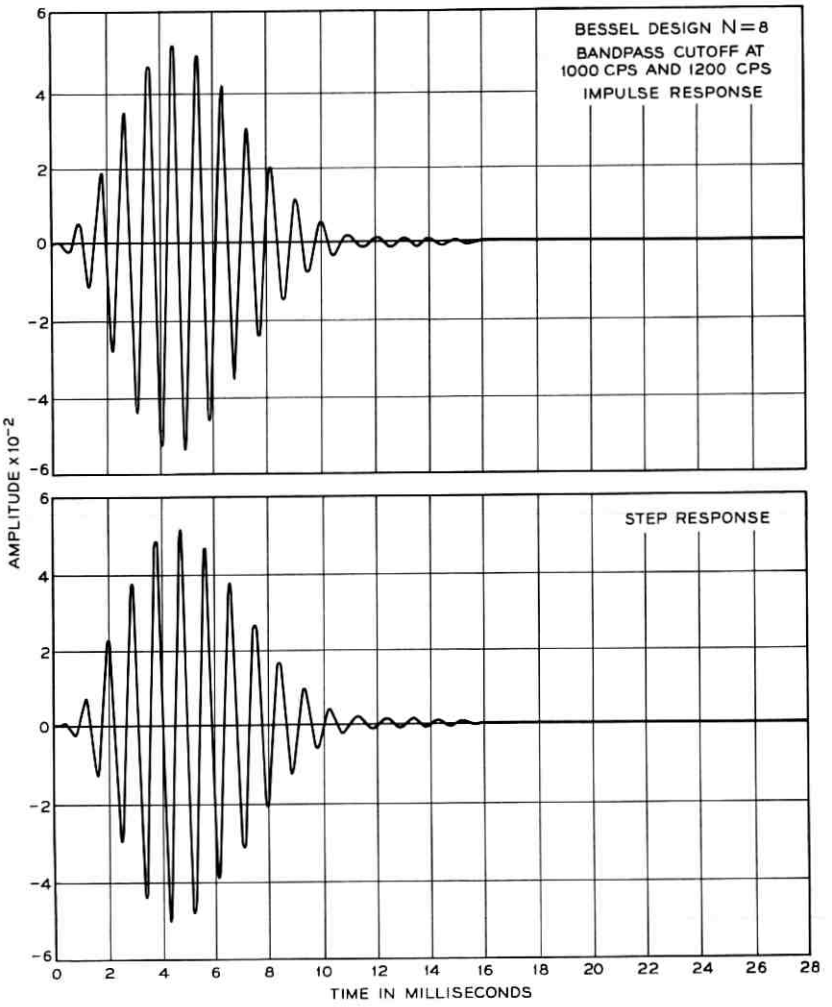


Fig. 5 — Time response for a vocoder-channel bandpass filter.

Having determined the representation for the various filters, the next step required is that of coding each of the 15 channels in the vocoder. From Fig. 1, it is seen that each channel (omitting multiplexing and transmission) consists of a bandpass filter, a full-wave rectifier, a low-pass filter, a "chopper" modulator, and another bandpass filter. In addition, a third bandpass filter is used between the output of the excitation generator and the modulator. A block diagram for one of these

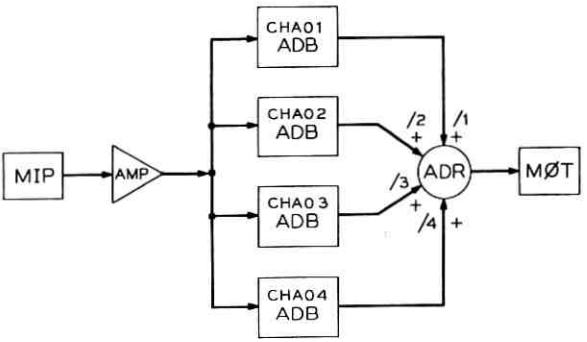


Fig. 6 — BLODIB representation for a vocoder-channel bandpass filter.

channels is shown in Fig. 7 along with the BLODIB coding necessary for simulation. The power of BLODIB with its "SUPER" capability is now readily apparent. Not only is the actual coding for each channel simplified, but any or all of the filters may be changed (by substitution of different SUPERS) without the necessity for recoding the entire system.

Completion of the above second prerequisite is usually all that is required in the programming of an entire system. However, two other features of BLODIB which also are used advantageously will be discussed briefly. These are the modular programming feature called SSUBR and the subroutine feature called SUBR. The SSUBR feature allows a block of BLODIB coding (similar to that contained within a SUPER) to be used as an external "module" to a main BLODIB program. However, unlike SUPERs which must be specified and compiled within the main program, SSUBRs are coded and compiled separately. These programs are then loaded as subroutines for use by the main program.

For example, the MOD (chopper) box shown in Fig. 7 was coded as an SSUBR and used in conjunction with the main vocoder program. This allowed different modulator configurations to be tested by simply "plugging" them into the program deck at run time. Multiplexing systems also were tested by using different external SSUBRs.

The SUBR feature permits coding an entire simulation so that it can be controlled by a main or "executive" program. (Such a program may be coded in FORTRAN or FAP.) This permits changing system parameters or using intermediate analysis programs as an integral part of the simulation.

At present, the simulated vocoder consists of a main FORTRAN program which supplies external parameters and provides the calling

TABLE 2 — BLODIB CODING FOR A VOCODER-CHANNEL BANDPASS FILTER

```
CHA       MACRO    I1, I2, I3, I4
IN        MIP      1, CHAI
CHAI      AMP      1., −4, CHA01, CHA02, CHA03, CHA04
CHA01     ADB      4
                   −448.005981
                     74.456865
                    114.566368
                   −189.968182
                    CHA05/2
CHA02     ADB      4
                    276.439266
                    193.008114
                    113.922181
                   −184.224327
                    CHA05/3
CHA03     ADB      4
                    187.295624
                   −112.815660
                    118.592467
                   −198.106075
                    CHA05/4
CHA04     ADB      4
                    −11.099112
                   −154.649326
                    117.119220
                   −180.637222
                    CHA06/2
CHA05     ADR      CHA06
CHA06     ADR      MOT
MOT       MOT
          END
```

sequences for the following BLODIB subroutines (SUBRs): an input routine which allows either the conventional mode or the voice-excited vocoder mode to be used; the main vocoder program; and, the output routine. The main vocoder program uses BLODIB SSUBRs for the excitation detector and generator and for the multiplexing and transmission networks. Fig. 8 shows how the various programs are interconnected. Variation of parameter or network configuration is accomplished by changing either parameters within a given block or changing the block completely. In this manner, optimization of system performance was achieved by optimizing the appropriate "modules" in the simulation. Hence, the programming of a complete vocoder system is achieved in an extraordinarily short amount of time using BLODIB.

Having completed the first two important prerequisites for simulation — namely sampled-data representation of the vocoder and its translation into the BLODIB programming language — there remains only the final step of preparing input speech data to be processed by the com-

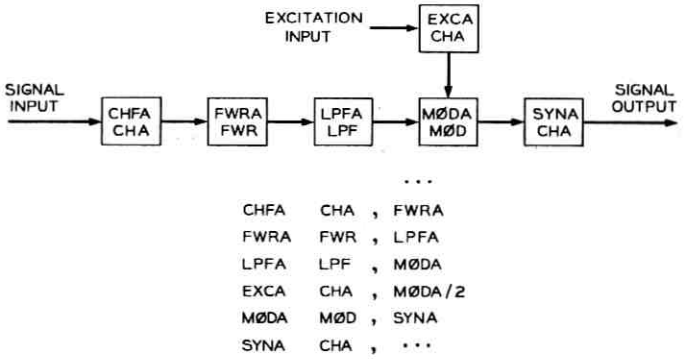


Fig. 7 — BLODIB representation and coding for a complete vocoder channel.

puter. This is accomplished by an analog-digital translator as indicated in Fig. 9. The translator, which can sample prerecorded input speech at rates up to 20 kcps, quantizes the input speech signal to one of 4096 levels (including sign). The sampled-and-quantized data is recorded at 800 bits/inch on standard digital magnetic tape in a format used by the BLODIB input/output routines. Running time on the IBM 7094$^{\text{Mod II}}$ for typical vocoder simulations is about 100 times real time (i.e., 1 second of speech requires 100 seconds for processing). The results from the simulation programs are similarly recorded on digital tape and then

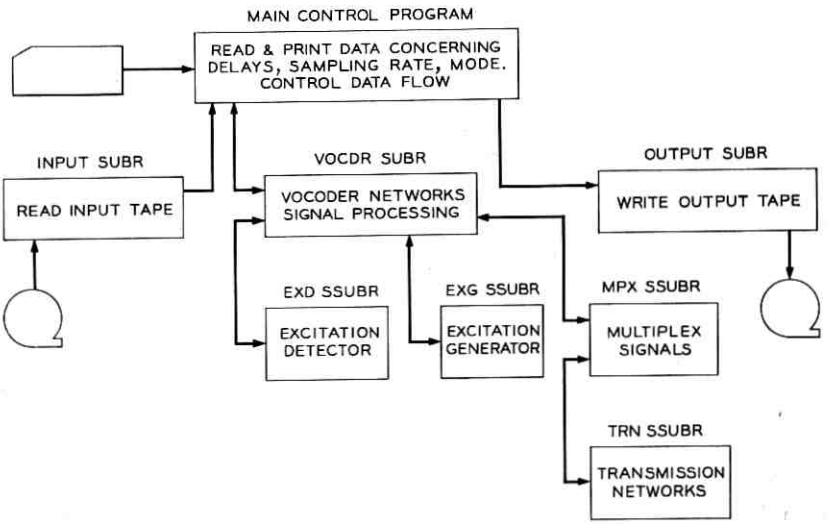


Fig. 8 — Flow diagram for vocoder simulation program.

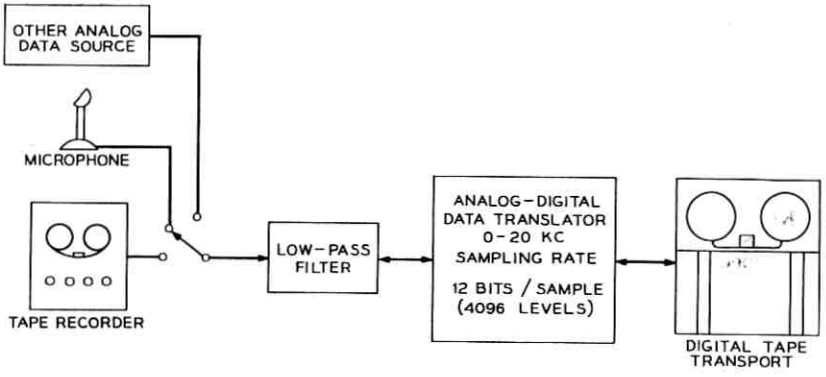


Fig. 9 — Preparation of digital input data for use with computer simulation of speech communication systems.

played back through the data translator. Of course, digital input tapes can be used over and over again, thereby removing this third prerequisite from succeeding simulations.

Thus far, BLODIB computer simulations have indicated: what filter designs yield good synthesized speech, what type of multiplexing and transmission systems might be satisfactorily used on the spectrum channel signals, how many spectrum channels are required, etc. As a result of these findings, improvement of system performance has been accomplished in a relatively short time and without the inherent delay and cost of building complex equipment.

VI. SUMMARY

Digital computer simulation of communication systems has been simplified and made more flexible by use of the BLODIB programming language.

Three prerequisites to actual simulation have been presented. They are:

(*i*) the finding of an appropriate sampled-data representation for the particular communication system,

(*ii*) the preparation of a BLODIB computer program for the above sampled-data representation, and

(*iii*) the preparation of input signals for processing by the program.

Details with respect to the BLODIB programming of the second step above were illustrated by their applications in the simulation of a vocoder system. Hence, new and complex communication systems can

be evaluated, optimized, and redesigned quickly and efficiently before final construction in hardware.

REFERENCES

1. Karafin, B. J., The New Block Diagram Compiler for Simulation of Sampled-Data Systems, AFIPS. Conference Proceedings, *27*, pt. 1, 1965, Fall Joint Computer Conference, Spartan Books, Washington, D.C., pp. 55–61.
2. Kelly, J. L., Jr., Lochbaum, C., Vyssotsky, V. A., A Block Diagram Compiler, B.S.T.J., *40*, May, 1961, pp. 669–676.
3. Schroeder, M. R., Logan, B. F., Prestigiacomo, A. J., New Methods for Speech Analysis-Synthesis and Bandwidth Compression, Fourth International Congress on Accoustics, Copenhagen, August 21–28, 1962.
4. Golden, R. M., Digital Computer Simulation of a Sampled-Data Voice-Excited Vocoder, J. Acoust. Soc. Amer., *35*, Sept., 1963, pp. 1358–1366.
5. Dudley, H., Automatic Synthesis of Speech, Proc. Nat. Acad. Sci., *25*, July, 1939, pp. 377–383.
6. Kaiser, J. F., Design Methods for Sampled-Data Filters, Proc. First Allerten Conference on Circuit and System Theory, Monticello, Illinois, Nov., 1963.
7. Golden, R. M. and Kaiser, J. F., Design of Wideband Sampled-Data Filters, B.S.T.J., *43*, July, 1964, Part 2, pp. 1533–1546.

The Capacity of the Band-Limited Gaussian Channel

By A. D. WYNER

(Manuscript received December 27, 1965)

Shannon's celebrated formula $W \ln(1 + P_o/N_oW)$ for the capacity of a time-continuous communication channel with bandwidth W cps, average signal power P_o, and additive Gaussian noise with flat spectral density N_o has never been justified by a coding theorem (and "converse"). Such a theorem is necessary to establish $W \ln(1 + P_o/N_oW)$ as the supremum of those transmission rates at which one may communicate over this channel with arbitrarily high reliability as the coding and decoding delay becomes large.

In this paper, a number of physically consistent models for this time-continuous channel are proposed. For each model the capacity is established as $W \ln(1 + P_o/N_oW)$ by means of a coding theorem and converse.

I. INTRODUCTION

As an idealized model for the time-continuous Gaussian channel (with bandwidth W cycles per second, two-sided noise spectral density $N_o/2$, and average power P_o), Shannon[1,2] employed the mathematical time-discrete channel which passes $2W$ real numbers x per second, with the average of x^2 restricted to be P_o. Each input x is perturbed by an independent "noise" random variable which is Gaussian with mean zero and variance N_oW. If by "channel capacity" we mean the maximum rate at which a channel is capable of transmitting information with arbitrarily small error probability as the coding and decoding delay becomes large, then the capacity of this time-discrete channel is given by the celebrated formula $W \log_2 (1 + P_o/N_oW)$ bits per second (or $W \ln (1 + P_o/N_oW)$ nats per second).

In order to show that the capacity is given by this formula, it is necessary to prove a coding theorem (showing the possibility of achieving "error-free" communication at any rate less than $W \log_2 (1 + P_o/N_oW)$), and a "converse" (showing the impossibility of achieving "error-free" coding at a rate exceeding this quantity). For this — purely mathematical — channel these theorems have been proved, and there is no question as to the meaning and validity of the capacity formula.

The way in which Shannon arrived at this time-discrete model for a "physical" time-continuous channel is described in detail in Section II. It will suffice to remark here that there remain questions as to the relation of this time-discrete model (and the resulting capacity formula) to a physically meaningful time-continuous channel. These difficulties center on the fact that the inputs and outputs of the time-continuous channel are band-limited signals which are not physically realizable. As we shall see in Section II, such assumptions lead to a number of anomalies and absurdities.

Our purpose in this paper is to find physically consistant mathematical models for the time-continuous band-limited Gaussian channel, and to establish their capacity by means of a coding theorem and converse. Schematically our results are of the following form:

Let $a(T,W,P_o)$ be a class of functions which are "approximately band-limited to W cycles per second and approximately time-limited to T seconds", and which have "average power" P_o. The channel inputs must be members of a. The noise is additive, stationary, and Gaussian with flat two-sided spectral density $N_o/2$ in the band $0 - W$ cycles per second (or "approximately" given as above). Then the channel capacity, defined as the maximum rate for which arbitrarily high reliability is possible (using signals from a) as T becomes large, is given "approximately" by $W \log_2 (1 + P_o/N_oW)$. The term "approximately" used here will, of course, be given a precise meaning below.

In Section II, Shannon's model and results are discussed, and in Section III our models and results are stated completely and discussed. Our proofs follow in Sections IV and V. A glossary is included at the end of the paper.

II. THE SHANNON MODEL

2.1 *The Time-Discrete Channel*

In order to fix ideas as well as to review some results which will be required subsequently, let us consider the following class of (time-discrete) channels: Every T seconds the input to the channel is a sequence of $n = [\alpha T]$ real numbers $\mathbf{x} = (x_1, x_2, \cdots, x_n)$, where $\alpha (0 < \alpha \leqq \infty)$ is a fixed parameter. Further, the input sequence must satisfy the "energy" constraint

$$E(\mathbf{x}) = \sum_{k=1}^{n} x_k^2 \leqq PT, \tag{1}$$

where $P > 0$ is another fixed parameter, and where $E(\mathbf{x})$ is, as indicated, the sum of the squares of the components of $\mathbf{x}$.

The channel output is also a real n-sequence $\mathbf{y} = (y_1, y_2, \cdots, y_n)$, where

$$y_k = x_k + z_k, \qquad k = 1,2, \cdots, n, \tag{2}$$

and the noise digits z_k $(k = 1,2, \cdots, n)$ are independent, normally distributed random variables with mean zero and variance N.

Let us assume that this channel is to be used in the communication system of Fig. 1. The output of the message source is a sequence of independent and equally likely binary digits which appear at the input of the coder at the rate of R_b digits (bits) per second. Every T seconds the coder input is one of $M = 2^{R_b T}$ binary sequences, each sequence being equally likely. Let us number the possible messages as $1,2, \cdots, M$. The coder contains a mapping of the message set $\{1,2, \cdots, M\}$ to a set (called a *code*) of M real n-sequences $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_M\}$ (called *code words*) satisfying (1). If message $i (i = 1,2, \cdots, M)$ is the coder input, then the coder output (and hence channel input) is the code word $\mathbf{x}_i$. Since it takes T seconds to transmit a code word, the system can process information continuously without a "backup" at the coder input. The transmission rate is R_b bits per second or $R = (\ln 2)R_b$ nats per second.

It is the task of the receiver (or decoder) to examine the received sequence $\mathbf{y}$, and determine which of the M code words was actually transmitted. Thus, we may think of the decoder as a rule which assigns to each possible received sequence $\mathbf{y}$, a code word $\mathbf{x}_i$. Let us denote by P_{ei} the probability that the decoder chooses the wrong code word given that $\mathbf{x}_i$ was transmitted. The over-all error probability is then

$$P_e = \frac{1}{M} \sum_{i=1}^{M} P_{ei}. \tag{3}$$

A transmission rate R (nats per second) is said to be *permissible* if for every $\lambda > 0$ one can find a T sufficiently large and a code with parameter T with $M = [e^{RT}]$ code words and $P_e \leqq \lambda$. With such a code, the system could process $R_b = R/\ln 2$ bits per second. We define the *channel capacity* C as the supremum of permissible rates. For the channel under discussion the channel capacity is given by the celebrated formula

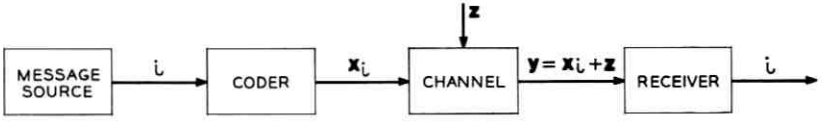


Fig. 1. — Time-discrete channel.

$$C = C_\alpha = \frac{\alpha}{2} \ln\left(1 + \frac{P}{\alpha N}\right). \tag{4}$$

In order to establish C as the capacity, one must prove two theorems. The first ("direct half") states that any $R < C$ is a permissible rate; that is, there exist codes with vanishingly small P_e as $T \to \infty$. The second theorem ("weak converse") states that no $R > C$ is a permissible rate; that is, for any sequence of codes with rate $R > C$, P_e is bounded away from zero. This has been done for the present channel for the case of a finite α by Shannon.[1,2,3] Let us observe that if we let $\alpha \to \infty$ in (4), we have $C_\alpha \xrightarrow{\alpha} P/2N$. The fact that $C_\infty = P/2N$ has been established by Ash.[4] The reader is referred to Ash [Ref. 5, Chapter 8] for a complete discussion of the above. The significance of the channel capacity then, is that it is the maximum rate for which arbitrarily high reliability is possible using signals in a certain class (i.e., those which satisfy (1)) with sufficiently long delay T.

2.2 *Application to the Band-Limited Gaussian Channel*

Shannon[1,2] has applied the above results to the communication system of Fig. 2. As above, the message source emits binary digits at the rate of R_b per second, and after T seconds, one of $M = 2^{R_b T}$ possible messages appears at the coder input. Corresponding to the ith message ($i = 1,2, \cdots, M$) the coder output is the function

$$x_i(t) = \sum_{k=1}^{n} x_{ik}\delta(t - k/2W), \tag{5a}$$

where $\delta(t)$ is the unit impulse, $n = [2WT]$, and the $\{x_{ik}\}_{k=1}^{n}$ satisfy

$$\sum_{k=1}^{n} x_{ik}^2 \leqq 2WP_eT, \qquad i = 1,2,\cdots,M. \tag{5b}$$

As for the time-discrete channel, the coder must contain a set of M real n-sequences. The channel input $s_i(t)$ is the result of passing $x_i(t)$ through an ideal low-pass filter with transfer function

$$H(\omega) = \begin{cases} \dfrac{1}{2W} & |\omega| \leqq 2\pi W, \\ 0 & |\omega| > 2\pi W, \end{cases} \tag{6}$$

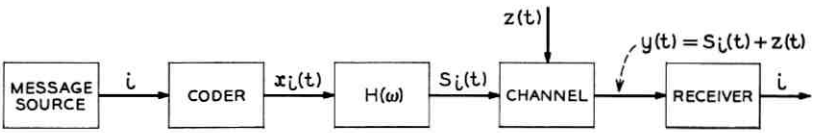


Fig. 2 — Shannon's time-continuous band-limited channel.

so that

$$s_i(t) = \sum_{k=1}^{n} x_{ik}\left[\frac{\sin 2\pi W(t - k/2W)}{2\pi W(t - k/2W)}\right]. \tag{7}$$

Thus, it takes T seconds to generate the filter input, and the system can process information at a rate of $R = (\ln 2)R_b$ nats per second without a "backup" at the coder input. Let us also remark that although the signal $s_i(t)$ is generated in T seconds, due to the physical unrealizability of $H(\omega)$, $s_i(t)$ is nonzero almost everywhere on $(-\infty,\infty)$. This leads to a fundamental difficulty which we shall discuss later.

Let $s(t)$ be the input to the channel due to a repeated application of the coding process (every T seconds). Then $s(t)$ is bandlimited to W cycles per second, and

$$\operatorname*{limit}_{T_o\to\infty} \frac{1}{T_o}\int_{-T_o/2}^{T_o/2} s^2(t)dt \leqq P_o. \tag{8}$$

Inequality (8) follows from (5b) and the orthogonality of

$$\frac{\sin 2\pi W(t - k/2W)}{2\pi W(t - k/2W)} \quad \text{and} \quad \frac{\sin 2\pi W(t - k'/2W)}{2\pi W(t - k'/2W)}$$

$(-\infty < k < k' < \infty)$ on the infinite interval $(-\infty,\infty)$. Thus, the channel input is a bandlimited signal with "average power" not exceeding P_o.

Again turning our attention to Fig. 2, the channel output is a function $y(t) = s(t) + z(t)$, where $z(t)$ is a sample from a Gaussian random process with spectral density

$$N(\omega) = \begin{cases} N_o/2 & |\omega| \leqq 2\pi W, \\ 0 & |\omega| > 2\pi W. \end{cases} \tag{9a}$$

The corresponding autocorrelation function of the noise is

$$R(\tau) = \mathcal{E}[z(t)z(t + \tau)] = N_o W\frac{\sin 2\pi W\tau}{2\pi W\tau}, \tag{9b}$$

where $\mathcal{E}$ denotes expectation.

Again it is the function of the receiver (or decoder) to examine $y(t)$ and determine what the input information was. Let us consider the signal $s_i(t)$ (7), which was generated during the interval $[0,T]$. The coefficients $\{x_{ik}\}_{k=1}^{n}$ are the values of $s_i(t)$ at the "sampling instants" $t = k/2W$, $k = 1,2,\cdots,n$. Since the noise is also bandlimited, the received signal $y(t)$ is bandlimited and may be completely characterized by its values at the sampling instants $y_k = y(k/2W)$, $k = 0, \pm 1, \pm 2, \cdots$. Clearly

$$y_k = x_{ik} + z_k, \qquad k = 1,2,\cdots,n, \tag{10}$$

where $z_k = z(k/2W)$ is the value of the noise $z(t)$ at the sampling instant $t = k/2W$. Since $s_i(k/2W) = 0$, for $k < 1$ and $k > n$, the only useful samples of y are $\{y_k\}_{k=1}^{n}$. Further it follows directly from (9b) that the z_k are independent, normally distributed random variables with mean zero and variance N_oW. Thus, it suffices to consider the input and output as n-sequences $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{in})$ and $\mathbf{y} = (y_1, \cdots, y_n)$ $(n = 2WT)$ related by (10). Let us remark here, that the code words corresponding to previous and successive intervals will not cause any interference with the code word corresponding to the interval $[0,T]$, since these other code words are zero at the sampling instants.

Inequality (5b) and (10) permit us to apply the results for the time-discrete Gaussian channel discussed above with parameters $\alpha = 2W$, $P = 2WP_o$, and $N = N_oW$. We conclude that this communication system (in Fig. 2) is capable of processing information at any rate R less than

$$C = W \ln \left(1 + \frac{P_o}{N_oW}\right), \tag{11}$$

with vanishingly small error probability as T becomes large. Since the channel inputs are bandlimited to W cycles per second, and by (8) have average power not exceeding P_o, it is generally believed that the *capacity* (taken as the maximum "error-free rate") of a channel which admits only bandlimited signals with average power P_o is given by (11). In fact, it has only been shown that it is possible to do at least as well as C (using the system of Fig. 2), and no converse has been proven. This is the first difficulty with the Shannon model which we shall attempt to remedy.

Further, there are other difficulties inherent in the use of this model. We are taking "capacity" to be a (maximum) transmission *rate*, but what is the rate for the system of Fig. 2? We have said merely that the *coder* can process information at a rate of R nats per second. However, because of the physical unrealizability of $H(\omega)$, we must discard all temporal notions about the channel input $s_i(t)$ as well as the output $y(t)$. The notion of *rate*, therefore, has only a limited meaning. In fact, since the received signal $y(t)$ is an entire function, it is perfectly predictable for all time from observations over a finite interval. Thus the receiver, by observing $y(t)$ in a tiny interval, could extrapolate $y(t)$ for all time and obtain sample values at an arbitrarily high rate. This anomaly is the second difficulty with the Shannon model.

It is the purpose of this paper to present a model for the time-con-

tinuous band-limited Gaussian channel for which the capacity (defined as the maximum "error-free rate") is given by (11). This will necessitate proving a "direct half" and "converse" to a coding theorem. Further, the model should avoid the second difficulty mentioned above. We shall obtain results of the following form:

Let $a(T,W,P_o)$ be a class of functions which are "approximately bandlimited to W cycles per second and approximately time-limited to T seconds", and which have total "energy" not exceeding P_oT. The noise is taken to be stationary and Gaussian with spectral density given (or "approximately" given) by (9a). Then the channel capacity, defined as the maximum rate for which arbitrarily high reliability is possible (using signals from a) as T becomes large, is given "approximately" by $W \ln (1 + P_o/N_oW)$. The term "approximately" used here will, of course, be given a precise meaning below.

III. SUMMARY OF RESULTS

We shall propose four models for the channel and find the capacity of each. Each model is of the following form:

(*i*) Definition of a suitable class of allowable signal functions, $a(T,W,P_o)$, which are "approximately bandlimited to W cycles per second, approximately time-limited to T seconds", and with total energy not exceeding P_oT.

(*ii*) Definition of the noise — taken to be stationary additive Gaussian noise with spectral density $N(\omega)$, which is "approximately" given by (9a).

We shall take W and P_o to be fixed parameters. A *code* with parameter T is a set of M functions (called *code words*) in $a(T,W,P_o)$. The transmission rate R is defined by $R = (1/T) \ln M$, so that $M = e^{RT}$. A decoding scheme is a mapping of the space of possible received signals (code word plus a noise sample) onto the code. If code word i ($i = 1, 2, \cdots, M$) is transmitted, we take P_{ei} to be the conditional probability that the decoder chooses a code word other than i, and hence makes an error. Since all code words are equally likely to be transmitted, the over-all error probability P_e is given by (3), i.e.,

$$P_e = \frac{1}{M} \sum_{i=1}^{M} P_{ei} .$$

A transmission rate R is said to be *permissible*, if for every $\lambda > 0$ one can find a T sufficiently large and a code with $M = [e^{RT}]$ code words for which $P_e \leqq \lambda$. The *channel capacity* C is defined as the supremum of

permissible rates. We shall find the capacity corresponding to a number of different $a(T,W,P_o)$ and $N(\omega)$. This will, as for the time-discrete channel, necessitate proving two coding theorems — a "direct half" and a "weak converse".

Before beginning the summary we shall need the following definitions. Let $s(t)$, $-\infty < t < \infty$, be a real-valued square-integrable function and $S(\omega)$ be its Fourier transform. Let the norm of $s(t)$ be

$$\| s \| = \left[\int_{-\infty}^{\infty} s^2(t) dt \right]^{\frac{1}{2}}. \tag{12}$$

The frequency and time "concentration" of s are

$$K_B(s,2\pi W) = \frac{1}{2\pi} \int_{-2\pi W}^{2\pi W} | S(\omega) |^2 d\omega / \| s \|^2, \tag{13a}$$

and

$$K_D(s,T) = \int_{-T/2}^{T/2} s^2(t) dt / \| s \|^2, \tag{13b}$$

respectively. Further, let D_T be the "time-truncation" operator defined by

$$D_T s = \begin{cases} s(t) & |t| \leqq T/2, \\ 0 & |t| > T/2. \end{cases} \tag{14}$$

With these definitions in hand, we are able to state our results. In each case we shall define the channel model and then give the channel capacity. Although there are some difficulties inherent in these models, each model leads to a *mathematical theorem* which justifies Shannon's capacity formula.

Model 1: To begin with, let us take for the set a of "allowable" inputs. $a_1(T,W,P_o)$, the set of functions $s(t)$ satisfying

$$s(t) = 0, \quad |t| > T/2, \tag{15a}$$

$$\| s \|^2 \leqq P_o T, \tag{15b}$$

$$K_B(s,2\pi W) \geqq 1 - \eta \quad (0 < \eta < 1). \tag{15c}$$

Hence, our allowable signals are functions which are strictly time-limited and approximately band-limited. As $\eta \to 0$, the allowable signals become more perfectly bandlimited. The noise spectrum is taken to be

$$N(\omega) = \begin{cases} N_o/2 & |\omega| \leqq 2\pi W, \\ \nu N_o/2 & |\omega| > 2\pi W, \end{cases} \tag{16}$$

where $0 < \nu \leqq 1$. As $\nu \to 0$, (16) is in some sense "approximately" the same as (9a). The average noise power outside the band ($|\omega| > 2\pi W$), however, is infinite. In this case, Theorem 3 establishes

$$C = C_\eta = W \ln \left(1 + (1 - \eta) \frac{P_o}{N_o W}\right) + \eta \frac{P_o}{\nu N_o} \tag{17}$$

as the channel capacity. As $\eta \to 0$, the capacity approaches the classical formula $W \ln (1 + P_o/N_o W)$.

The principal difficulty with this model is the assumption of infinite average noise power, which is hardly a physically acceptable notion. Further, there are mathematical difficulties inherent in a spectral density given by (16) which implies a covariance containing an impulse function. Often the assumption of a spectrum in (16) can be justified by the fact that it can be approximated as closely as desired in the frequency range of interest by a spectrum with finite power. However, the following theorem, the proof of which is Appendix B, renders this justification meaningless in this case.

Theorem 5: Let $a(T,W,P_o)$ be as in (15) and let the noise be additive and Gaussian with spectral density $N(\omega)$, where

$$\int_{-\infty}^{\infty} N(\omega) d\omega < \infty.$$

Then the capacity $C_\eta = \infty$ regardless of how small η may be.

Intuitively, we may see that this is true by observing that, since the above integral exists, $N(\omega)$ must be arbitrarily small in some frequency range. Hence, by placing some signal energy into this frequency range, we can make the "signal-to-noise" ratio arbitrarily large, and therefore, the permissible rate of transmission arbitrarily high.

Accordingly, we shall assume for the remaining models that the noise is additive, Gaussian, with spectral density

$$N(\omega) = \begin{cases} N_o/2 & |\omega| \leqq 2\pi W, \\ 0 & |\omega| > 2\pi W. \end{cases} \tag{18}$$

This corresponds more closely with the usual formulation of a band-limited channel. It remains to find a suitable class of input signals, $a(T,W,P_o)$. We consider some possibilities.

Model 2: This model defines $a = a_2(T,W,P_o)$ as the set of functions $s(t)$

satisfying

$$S(\omega) = 0, \qquad |\omega| > 2\pi W, \tag{19a}$$

$$\| s \|^2 \leqq P_o T, \tag{19b}$$

$$K_D(s,T) \geqq 1 - \eta \quad (0 < \eta < 1). \tag{19c}$$

Thus, a_2 is a set of strictly band-limited, approximately time-limited functions. As $\eta \to 0$, the allowable signals become more perfectly time-limited. With the noise as defined in (18), Theorem 2 establishes

$$C = C_\eta = W \ln \left(1 + (1 - \eta) \frac{P_o}{N_o}\right) + \eta \frac{P_o}{N_o} \tag{20}$$

as the channel capacity. Again, as $\eta \to 0$, C_η approaches the classical formula $W \ln [1 + (P_o/N_oW)]$.

Model 2 is an intuitively plausible model for the band-limited channel, and Theorem 2 which establishes its capacity is a mathematically rigorous result which, in the limit, yields the desired capacity formula. There are, however, two difficulties inherent in this formulation. The first is that since the allowable signals $s(t)$ are band-limited, it is not possible to generate them in finite time. Thus the central idea of a transmission *rate* has, at best, a limited meaning. The Shannon model (Fig. 2) also suffers from this difficulty (see Section II). The other problem with this formulation is that if code words are transmitted sequentially, we will have an interference problem (i.e., the tails of successive signals will overlap), the resolution of which is not known at present. The following two models contain neither of these difficulties.

Model 3: This model avoids the difficulties of Model 2 by letting the code words be strictly time-limited and approximately band-limited. However, as we have seen in Theorem 5, the definition of approximately band-limited functions employed above (15) yields an infinite capacity. Thus we seek an alternate way of characterizing "approximately" band-limited or "slowly changing" functions. We proceed as follows. Let $x(t)$ be a function satisfying $x(t) = 0$, $|t| > T/2$, and $\| x \|^2 < \infty$. If $x = D_T\hat{x}$, where $\hat{x}$ is a strictly bandlimited function and D_T is defined by (14), we may define a "frequency concentration" of x by

$$K_B'(x,2\pi W) = \frac{\| x \|^2}{\| \hat{x} \|^2}. \tag{21}$$

If we cannot express x as $D_T\hat{x}$, we take $K_B' = 0$. For example, if $x(t)$ or any of its derivatives has even a small discontinuity then we cannot write $x = D_T\hat{x}$, so that $K_B'(x,2\pi W) = 0$ and x is not approximately

bandlimited in this sense. This is so no matter how large $K_B(x,2\pi W)$ may be. Conversely, it is shown in Appendix C that for any function x

$$K_B(x,2\pi W) \geqq 1 - 2\sqrt{\frac{1 - K_B'(x,2\pi W)}{K_B'(x,2\pi W)}}, \tag{22}$$

so that a K_B' close to unity implies a K_B close to unity. Thus, saying that a function x has a K_B' close to unity implies that x is "slowly changing" and that K_B is also close to unity.

We now choose that set $a = a_3(T,W,P_o)$ of allowable inputs as the set of functions $s(t)$ for which

$$s(t) = 0, \qquad |t| > T/2, \tag{23a}$$

$$\| s \|^2 \leqq P_o T, \tag{23b}$$

$$K_B'(s,2\pi W) \geqq 1 - \eta \quad (0 < \eta < 1). \tag{23c}$$

Thus a_3 is a set of strictly time-limited, and approximately band-limited functions. In this case, Theorem 4 establishes

$$C = C_\eta = W \ln\left(1 + \frac{P_o}{N_o W}\right) + \frac{\eta}{1 - \eta}\frac{P_o}{N_o} \tag{24}$$

as the channel capacity. Again $C_\eta \to W \ln [1 + (P_o/N_o W)]$ as $\eta \to 0$.

The significance of constraint (23c) is that it makes it impossible for the communicator to make any *use* of the high-frequency components which must of necessity be included in the signal (since it is time-limited). Model 3, therefore, provides a mathematically rigorous theorem which does not involve any complications concerning physical realizability, and yields the desired capacity.

Our final formulation is as follows:

Model 4: Let $a = a_4(T,W,P_o)$ be the set of strictly time-limited, approximately band-limited functions $s(t)$ which satisfy

$$s(t) = 0, \qquad |t| \geqq T/2, \tag{25a}$$

$$\| s \|^2 \leqq P_o T, \tag{25b}$$

$$K_B(s,2\pi W) \geqq 1 - \eta. \tag{25c}$$

Now Theorem 5 (stated above) tells us that if the noise were as in (18), then the capacity is infinite. In actuality one could not be sure that the noise was absolutely band-limited. In fact, whether or not the noise is

strictly band-limited is not verifiable in the laboratory. It is reasonable, therefore, to assume that the noise is given by $z(t) = z_1(t) + z_2(t)$, where $z_1(t)$ is a sample from a Gaussian random process with spectral density (18). For $z_2(t)$ we require only that

$$\int_{-T/2}^{T/2} z_2^2(t)dt \leqq \nu N_o W T, \tag{26}$$

where $\nu > 0$ is small. We place no other restrictions on the spectrum of z_2 or on its probability structure. Since the expected value of the energy of $z_1(t)$ in $[-T/2,T/2]$ is $N_o WT$, (26) implies that the energy of $z(t)$ is nearly all in $z_1(t)$ ($\nu \ll 1$). We shall assume that $z_2(t)$ may depend on the code and decoding rule used, on the code word transmitted, and the sample $z_1(t)$. We require our communication system to perform well no matter what $z_2(t)$ may be.

Let us say that a code (satisfying (25)) and a decoding rule have been chosen. Let us also assume that the rule for selecting $z_2(t)$ has been chosen. Let $P_e(z_2)$ be the resulting error probability. Then define

$$P_e = \max_{z_2} P_e(z_2), \tag{27}$$

where the maximization in (27) is over all rules for choosing $z_2(t)$ — with the code and decoding rule fixed. The channel capacity is the supremum of those rates for which P_e may be made to vanish as $T \to \infty$.

It can be shown (see Appendix D) that the capacity C is given by

$$C = C_{\eta,\nu} = W \ln\left(1 + \frac{P_o}{N_o W}\right) + \varepsilon(\eta,\nu), \tag{28}$$

where $\varepsilon(\eta,\nu) \to 0$ as $\eta,\nu \to 0$ provided $\nu/\eta > P_o/N_o W$, the signal-to-noise ratio. Since we may consider η and ν to be limits on the accuracy of our measuring equipment, the former on measuring the signal* and the latter on measuring the noise, it is reasonable to assume, as we did in (28), that η and ν go to zero at the same rate.

An alternate and mathematically equivalent formulation of Model 4 is as follows: Let the signals $s(t)$ be as in (25) and the noise $z(t)$ be as in (18). Now in reality one could not expect the decoder to be capable of infinitesimally accurate measurements. It is reasonable, therefore, to assume that there is an inherent uncertainty in all measurements made by the decoder, and to require that the communication system perform well despite this uncertainty. Specifically, we require that the decoding regions satisfy the following condition: If $y_1(t)$ is decoded as s_i, and $y_2(t)$ is decoded as $s_j (i \neq j)$, then

* *I.e.*, η represents a limit on the measurement of the frequency component of the signal outside the band.

$$\int_{-T/2}^{T/2} (y_1(t) - y_2(t))^2 \, dt \geqq 2\nu N_0 WT. \tag{29}$$

In other words, if a received signal $y(t)$ is close to the "border" between decoding regions, we cannot, because of the uncertainty in the accuracy of our measurements, be sure to which region $y(t)$ belongs. Condition (29) forces the decoder to give up on such a $y(t)$ and to announce an error. The capacity for this alternate model is also given by (28). Let us remark that here ν is again a measure of the accuracy of our measuring instruments, this time at the decoder, so that again it is reasonable to expect η and ν to tend to zero at the same rate.

IV. PRELIMINARIES TO PROOFS

4.1 *The Product of Time-Discrete Channels*

The *product* or parallel combination of r time-discrete Gaussian channels is defined as follows. Every T seconds the input to the channel is an r-tuple $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(r)})$, where

$$\mathbf{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \cdots, x_{n_i}^{(i)}) \qquad (i = 1, 2, \cdots, r)$$

is a real n_i-vector ($n_i = \alpha_i T$, α_i a fixed parameter). Each vector $\mathbf{x}^{(i)}$ satisfies the energy constraint

$$E[\mathbf{x}^{(i)}] = \sum_{k=1}^{n_i} [x_k^{(i)}]^2 \leqq P_i T, \qquad i = 1,2,\cdots, r, \tag{30}$$

where the $P_i > 0$ are fixed parameters. The channel output is also an r-tuple $(\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(r)})$, where the $\mathbf{y}^{(i)}$ are n_i-vectors given by

$$\mathbf{y}^{(i)} = \mathbf{x}^{(i)} + \mathbf{z}^{(i)}, \tag{31}$$

where the $\mathbf{z}^{(i)}$ are n_i-vectors whose coordinates are independent Gaussian random variables with mean zero and variance $N_i (i = 1, 2, \cdots, r)$. Further, the $\{\mathbf{z}^{(i)}\}_{i=1}^{r}$ are statistically independent. Codes, permissible rates of transmission, and channel capacity are defined as in Section I. The following is proved in Ref. 6.

Lemma A: The capacity C of the product of r time-discrete Gaussian channels, with parameters (α_i, P_i, N_i), $i = 1, 2, \cdots, r$, is given by the sum of the capacities of the component channels:

$$C = \sum_{i=1}^{r} \frac{\alpha_i}{2} \ln \left(1 + \frac{P_i}{\alpha_i N_i}\right). \tag{32}$$

Equation (32) also holds when one or more of the $\alpha_i = \infty$. In this case we read $x \ln [1 + (c/x)] \mid_{x=\infty}$ as c.

4.2 *The Jointly-Constrained Product Channel*

We define the jointly-constrained product of time-discrete channels exactly as the ordinary product with constraint (30) replaced by constraints of the following form:

Type 1: Let $r = 2$ and $N_1 = N_2 = N$ and instead of (30) we have

$$E(\mathbf{x}) = E(\mathbf{x}^{(1)}) + E(\mathbf{x}^{(2)}) \leqq PT. \tag{33a}$$

If $\alpha_1 \leqq \alpha_2$ we introduce an additional constraint on $\mathbf{x}^{(2)}$

$$E(\mathbf{x}^{(2)}) \leqq \hat{\eta} E(\mathbf{x}) \tag{33b}$$

where $\hat{\eta} (0 \leqq \hat{\eta} \leqq 1)$ is another fixed parameter. In other words, we have constrained the total energy of the two input vectors (33a), and introduced another constraint on the second input vector $\mathbf{x}^{(2)}$ requiring it to have no more than $\hat{\eta}$ of the total energy (33b). If $\alpha_2 \leqq \alpha_1$, we replace (33b) by a similar constraint on $\mathbf{x}^{(1)}$.

Type 2: Let $r = 3$, $N_1 = N_2$, and $N_1 \geqq N_3$. Further, let $\alpha_3 = \infty$. Instead of (30) we require that $\mathbf{x}$ satisfy

$$E(\mathbf{x}) = E(\mathbf{x}^{(1)}) + E(\mathbf{x}^{(2)}) + E(\mathbf{x}^{(3)}) \leqq PT, \tag{34a}$$

$$E(\mathbf{x}^{(3)}) \leqq \hat{\eta} E(\mathbf{x}). \tag{34b}$$

This is a special case of type 1 when $\alpha_2 = 0$, $N_1 = N_3$.

Type 3: Let $r = 2$, $N_1 = N_2 = N$, and $\alpha_2 = \infty$. Instead of (30) require $\mathbf{x}$ to satisfy

$$E(\mathbf{x}^{(1)}) \leqq PT, \tag{35a}$$

$$E(\mathbf{x}^{(2)}) \leqq \hat{\eta} E(\mathbf{x}). \tag{35b}$$

We now ask what is the capacity C of these channels? The answer is the following theorem which is proven in Appendix A.

Theorem 1: The capacity C of the jointly-constrained product channel as defined above is

Type 1 ($r = 2, N_1 = N_2 = N$):

$$C = C_1((1 - \beta)P) + C_2(\beta P), \tag{36}$$

where

$$\beta = \min\left(\hat{\eta}, \frac{\alpha_2}{\alpha_1 + \alpha_2}\right), \tag{37a}$$

and

$$C_i(x) = \frac{\alpha_i}{2} \ln\left(1 + \frac{x}{\alpha_i N}\right), \qquad i = 1,2. \tag{37b}$$

Again when $\alpha_i = \infty$, we interpret $x \ln [1 + (c/x)] \mid_{x=\infty} = c$. In particular, when $\alpha_2 = \infty$ $(\alpha_1 < \infty)$, $\beta = \hat{\eta}$, so that (36) implies that we can do no better than putting as much energy into Channel 2 as (33b) will permit.

Type 2 $(r = 3, N_1 = N_2 \geqq N_3, \alpha_3 = \infty)$:

$$C = \frac{\alpha_1}{2} \ln\left(1 + \frac{(1 - \hat{\eta})P}{(\alpha_1 + \alpha_2)N_1}\right) + \frac{\alpha_2}{2} \ln\left(1 + \frac{(1 - \hat{\eta})P}{(\alpha_1 + \alpha_2)N_1}\right) + \hat{\eta}\frac{P}{2N_3} \tag{38}$$

Type 3 $(r = 2, N_1 = N_2 = N, \alpha_2 = \infty)$:

$$C = \frac{\alpha_1}{2} \ln\left(1 + \frac{P}{\alpha_1 N}\right) + \frac{\hat{\eta}P}{(1 - \hat{\eta})2N}. \tag{39}$$

4.3 *Prolate Spheroidal Wave Functions*

The following material can be found in Ref. 7. Given any $W,T > 0$ we can find a countably infinite set of real functions $\{\psi_i(t)\}_{i=1}^{\infty}$, called *prolate spheroidal wave functions* (PSWF), and a set of real positive numbers

$$1 > \lambda_1 > \lambda_2 > \cdots \tag{40}$$

with the following properties:*

(*i*) The $\psi_i(t)$ are bandlimited to W cycles per second, orthonormal on the real line, and complete in the space of bandlimited functions of bandwidth W cycles per second.

(*ii*) The restrictions of the $\psi_i(t)$ to the interval $[-T/2, T/2]$ are orthogonal:

$$\int_{-T/2}^{T/2} \psi_i(t)\psi_j(t)dt = \begin{cases} \lambda_i & i = j, \\ 0 & i \neq j. \end{cases} \tag{41}$$

* Note that the first PSWF is $\psi_1(t)$. In Ref. 7, on the other hand, the first PSWF is $\psi_0(t)$.

The restrictions of the $\psi_i(t)$ are also complete in $\mathfrak{L}_2$ $[-T/2,T/2]$, the space of square integrable functions on $[-T/2,T/2]$.

(iii) For all t, the $\psi_i(t)$ satisfy the integral equation

$$\lambda_i\psi_i(t) = \int_{-T/2}^{T/2} \psi_i(s) \frac{\sin 2\pi W(t-s)}{\pi(t-s)}\, ds. \tag{42}$$

Thus the λ_i are the eigenvalues, and the ψ_i the eigenfunctions of the integral equation (42). It follows immediately from (42) that the time-limited functions $D_T\psi_i$ (see (14)) have frequency concentration (see (13a))

$$K_B(D_T\psi_i, 2\pi W) = \lambda_i, \qquad i = 1, 2, \cdots. \tag{43}$$

It can be shown that the λ_i and ψ_i depend upon W and T only through the product WT. Further,

(iv) For a fixed $\delta > 0$:

$$\lambda_{2WT(1-\delta)} \to 1 \quad \text{as} \quad WT \to \infty \tag{44a}$$

and

$$\lambda_{2WT(1+\delta)} \to 0 \quad \text{as} \quad WT \to \infty. \tag{44b}$$

Thus roughly speaking, for large WT, approximately $2WT$ of the λ_i are approximately unity, and the remainder are approximately zero.

4.4 *Karhunen-Loeve Expansion*

Let $z(t)$ be a Gaussian random process with spectral density $N(\omega)$ given by (18). Then, using the Karhunen-Loeve Theorem[8], we may write $z(t)$ as

$$z(t) = \sum_{k=1}^{\infty} z_k\psi_k(t), \qquad -\frac{T}{2} \leqq t \leqq \frac{T}{2}, \tag{45}$$

where the $\psi_k(t)$ are PSWF's, and the z_k are independent random variables which are normally distributed with mean zero and variance $N_o/2$. The sum in (45) converges to $z(t)$ with probability 1 for every t.

If $N(\omega)$ is given by (16), then we may formally represent $z(t)$ by

$$z(t) = \sum_{k=1}^{\infty} z_k \frac{\psi_k(t)}{\sqrt{\lambda_k}}, \qquad -\frac{T}{2} \leqq t \leqq \frac{T}{2}, \tag{46}$$

where the λ_k are the eigenvalues of the PSWF's (40), and the z_k are independent normally distributed random variables with mean zero and variance

$$\mathcal{E}(z_k^2) = \frac{N_o}{2}[\lambda_k(1-\nu)+\nu].$$

Thus from (44) roughly speaking, for large WT, approximately $2WT$ of the z_k have variance $N_o/2$, and the remainder variance $\nu N_o/2$.

V. PROOFS OF THE THEOREMS

The general ideal of the proofs in this section is as follows. All the time continuous input signals (i.e., members of $a(T,W,P_o)$ can be written in a Fourier series in PSWF's in which, roughly speaking, the first $2WT$ terms correspond to the part of the signal which is simultaneously approximately confined to the frequency band $|\omega| \leqq 2\pi W$ and to the time interval $|t| \leqq T/2$. The noise sample $z(t)$ may also be written in a Karhunen-Loeve expansion in PSWF's. The result is to reduce the time-continuous channel into a jointly-constrained product of time-discrete channel (discussed in Section 4.2). Channel 1 corresponds to the first $2WT$ PSWF's so that the parameter $\alpha_1 = 2W$. Channel 2 corresponds to the remaining PSWF's so that $\alpha_2 = \infty$. The energy requirement on the time continuous signal $\| s \|^2 \leqq PT$ yields a joint energy constraint for the product channels (as in (33a) for example), and the requirement that the energy outside the frequency band (or time-interval) be small yields a second energy constraint on the input to Channel 2 (as in (33b) for example). Application of Theorem 1 then yields the desired theorems. In the remainder of this section we shall make these ideas precise.

We begin by establishing the capacity of the channel defined by Model 2.

Theorem 2: Let the allowable signal set be $a_2(T,W,P_o)$, the set of functions $s(t)$ satisfying

$$S(\omega) = 0, \qquad |\omega| > 2\pi W, \tag{47a}$$

$$\| s \|^2 \leqq P_o T, \tag{47b}$$

$$K_D(s,T) \geqq 1 - \eta \; (0 < \eta < 1). \tag{47c}$$

The noise is a sample from a Gaussian random process with spectral density

$$N(\omega) = \begin{cases} N_o/2 & |\omega| \leqq 2\pi W, \\ 0 & |\omega| > 2\pi W. \end{cases} \tag{48}$$

Then the channel capacity is

$$C = C_\eta = W \ln\left(1 + (1-\eta)\frac{P_o}{N_o W}\right) + \eta \frac{P_o}{N_o}. \tag{49}$$

Proof:

(*i*) Direct Half: Let R be given satisfying

$$R < W \ln \left[1 + (1 - \eta) \frac{P_o}{N_o W} \right] + \eta \frac{P_o}{N_o}. \tag{50}$$

Since the right member of (50) is continuous in η and W, we may find a $\delta > 0$ and $\sigma > 0$ sufficiently small so that

$$R < W(1 - \delta) \ln \left[1 + (1 - \eta + \sigma) \frac{P_o}{N_o W (1 - \delta)} \right] + \frac{(\eta - \sigma) P_o}{N_o} \triangleq C^*.$$

We see from (36) that C^* is the capacity of a type 1 jointly constrained product channel with parameters

$$P = P_o, \quad N = N_o/2, \quad \hat{\eta} = \eta - \sigma, \quad \alpha_1 = 2W(1 - \delta), \quad \alpha_2 = \infty. \tag{51}$$

We now show how to construct codes for the time-continuous "channel' with rate R and with vanishingly small error probability (as $T \to \infty$). Let $\mathbf{x} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$ be an allowable input vector for the type 1 time-discrete product channel with parameters given by (51). Then the corresponding input for the time-continuous channel is

$$s(t) = \sum_{k=1}^{2W(1-\delta)T = \alpha_1 T} x_k^{(1)} \psi_k(t) + \sum_{k=1}^{\infty = \alpha_2 T} x_k^{(2)} \psi_{k+2W(1-\delta)T}(t) \tag{52}$$

where the $\{\psi_i(t)\}_{i=1}^{\infty}$ are the PSWF's (Section 4.3) with parameters W and T. We first verify that signals of the form of (52) are allowable inputs, i.e., belong to $a_2(T,W,P_o)$ and satisfy (47). That the $s(t)$ are bandlimited and satisfy (47a), follows from the fact that the PSWF's have this property (Section 4.3). Further, the energy of $s(t)$ satisfies

$$\| s \|^2 = \sum_{k=1}^{\alpha_1 T} [x_k^{(1)}]^2 + \sum_{k=1}^{\alpha_2 T} [x_k^{(2)}]^2 = E(\mathbf{x}) \leqq PT, \tag{53}$$

where use has been made of the orthonomality of the PSWF's on $(-\infty, -\infty)$ (Section 4.3(*i*)), and the joint energy constraint on $\mathbf{x}$ (33a). Thus $s(t)$ satisfies (47b). Finally, from the orthogonality of the PSWF's on $[-T/2, T/2]$ (41), and the monotonicity of the λ_k (40) we have

$$\begin{aligned} 1 - K_D(s,T) &= \frac{\| (1 - D_T)s \|^2}{\| s \|^2} \\ &= \sum_{k=1}^{2W(1-\delta)T} \frac{(1 - \lambda_k)[x_k^{(1)}]^2}{E(\mathbf{x})} + \sum_{k=1}^{\infty} \frac{(1 - \lambda_{2WT(1-\delta)+k})}{E(\mathbf{x})} [x_k^{(2)}]^2 \\ &\leqq [1 - \lambda_{2WT(1-\delta)}] \frac{E(\mathbf{x}^{(1)})}{E(\mathbf{x})} + \frac{E(\mathbf{x}^{(2)})}{E(\mathbf{x})}. \end{aligned} \tag{54}$$

Now since $\lambda_{2WT(1-\delta)} \to 1$ as $T \to \infty$ (44a), and $E(\mathbf{x}^{(1)})/E(\mathbf{x}) \leqq 1$, with T sufficiently large we have

$$[1 - \lambda_{2WT(1-\delta)}] \frac{E(\mathbf{x}^{(1)})}{E(\mathbf{x})} \leqq \sigma.$$

Since $E(\mathbf{x}^{(2)})$ must satisfy (33b) (with $\hat{\eta} = \eta - \sigma$), we have (with T sufficiently large)

$$1 - K_D(s,T) \leqq \sigma + \eta - \sigma = \eta, \tag{55}$$

so that $s(t)$ satisfies (47c). Thus $s(t)$ belongs to $a_2(T,W,P_o)$.

Now we may express the noise in a Karhunen-Loeve expansion as

$$z(t) = \sum_{k=1}^{\alpha_1 T} z_k^{(1)} \psi_k(t) + \sum_{k=1}^{\infty} z_k^{(2)} \psi_{\alpha_1 T+k}(t), \tag{56}$$

where again the ψ_k are PSWF's and the $\{z_k^{(i)}\}_{1 \leqq k < \infty}^{i=1,2}$ are independent normally distributed random variables with mean zero and variance $N = N_o/2$. The output signal $y(t) = s(t) + z(t)$ is

$$y(t) = \sum_{k=1}^{\alpha_1 T} y_i^{(1)} \psi_k(t) + \sum_{k=1}^{\infty} y_k^{(2)} \psi_{\alpha_1 T+k}(t), \tag{57}$$

where the $y_k^{(i)}$ are obtainable by integration from the signal $y(t)$. Further,

$$y_k^{(i)} = x_k^{(i)} + z_k^{(i)}, \tag{58}$$

so that we conclude that our time-continuous channel with signals constructed in this way is equivalent to the type 1 jointly-constrained product channel with parameters (51) and capacity C^* (see Appendix E). Since $R < C^*$, we may therefore construct codes with rate R for either channel with error probability $P_e \to 0$ as $T \to \infty$. This is the direct half of Theorem 2.

(ii) Weak Converse: Say we are given a sequence of codes for our time-continuous channel with parameters $\{T_i\}_{i=1}^{\infty}$, with code words belonging to $a_2(T_i, W, P_o)$ (as defined in (47)), with error probability $P_e^{(i)}$, and rate

$$R > W \ln\left(1 + (1 - \eta)\frac{P_o}{N_o W}\right) + \eta \frac{P_o}{N_o}. \tag{59}$$

We shall show that $P_e^{(i)}$ must be bounded away from zero so that the capacity C (the maximum permissible rate) cannot exceed the right member of (59).

Now as in the proof of the direct half we may (by (59)) find a $\delta > 0$ and $\sigma > 0$ sufficiently small so that

$$R > W(1+\delta)\ln\left[1+\left(1-\frac{\eta}{1-\sigma}\right)\frac{P_o}{N_oW(1+\delta)}\right] + \frac{\eta}{1-\sigma}\frac{P_o}{N_o} \triangleq C^*. \tag{60}$$

Again, as in the direct half, C^* is the capacity of the type 1 jointly-constrained product channel with parameters

$$P = P_o, \qquad N = N_o/2, \qquad \hat{\eta} = \frac{\eta}{1-\sigma}, \qquad \alpha_1 = 2W(1+\delta), \qquad \alpha_2 = \infty. \tag{61}$$

Now if $s(t)$ is a code word from the code with parameter T_i, (so that $s \,\varepsilon\, a_2(T_i, W, P_o)$), we may write $s(t)$ as a Fourier series in PSWF's (due to the completeness of the PSWF's on the space of band-limited functions) (Section 4.3),

$$s(t) = \sum_{k=1}^{2WT_i(1+\delta)} x_k^{(1)}\psi_k(t) + \sum_{k=1}^{\infty} x_k^{(2)}\psi_{k+2WT_i(1+\delta)}(t), \qquad -\infty < t < \infty. \tag{62}$$

Hence, to each code word $s(t)$ for the time-continuous channel, there corresponds a vector $\mathbf{x} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$ whose coordinates are the coefficients in the above Fourier series. We now show that $\mathbf{x}$ is an allowable input to the type 1 jointly-constrained product channel with parameters given by (61). From the orthonormality of the PSWF's on $(-\infty, \infty)$ we have from (62), $\| s \|^2 = E(\mathbf{x})$. Since $s(t) \,\varepsilon\, a_2(T_i, W, P_o)$, we have $E(\mathbf{x}) \leqq PT_i$, so that $\mathbf{x}$ satisfies (33a). Further, from the orthogonality of the PSWF's on $[-T/2, T/2]$ and the monotonicity of the λ_k we have

$$1 - K_D(s,T_i) = \frac{\| (1-D_{T_i})s \|^2}{\| s \|^2} = \sum_{k=1}^{2WT_i(1+\delta)} \frac{[x_k^{(1)}]^2(1-\lambda_k)}{E(\mathbf{x})} + \sum_{k=1}^{\infty} \frac{[x_k^{(2)}]^2(1-\lambda_{2WT_i(1+\delta)+k})}{E(\mathbf{x})} \geqq [1-\lambda_{2WT_i(1+\delta)}]\frac{E(\mathbf{x}^{(2)})}{E(\mathbf{x})}. \tag{64}$$

With T_i sufficiently large (from 44b) we may put $\lambda_{2WT_i(1+\delta)} \leqq \sigma$, and since $1 - K_D(s,T_i) \leqq \eta$,

$$E(\mathbf{x}^{(2)}) \leqq \frac{\eta}{1-\sigma} E(\mathbf{x}) = \hat{\eta}E(\mathbf{x}), \tag{65}$$

so that $\mathbf{x}^{(2)}$ satisfies (33b).

Finally, if we proceed as in the proof of the direct half of this theorem and express the noise in a Karhunen-Loeve expansion in PSWF's, we can conclude that for each code for this time-continuous channel we can obtain a code for the time-discrete jointly-constrained product channel with the same rate and error probability (see Appendix E). Since the rate R exceeds the capacity of the latter channel we conclude from the weak converse to Theorem 1 that the error probability is bounded away from zero. This completes the proof.

The following theorems establish the capacity of the channels defined by Models 1 and 3.

Theorem 3: (Model 1) Let the allowable signal set be $a_1(T,W,P_o)$ *the set of functions* $s(t)$ *satisfying*

$$s(t) = 0, \qquad |t| \geqq T/2, \tag{66a}$$

$$\| s \|^2 \leqq P_o T, \tag{66b}$$

$$K_B(s,2\pi W) \geqq 1 - \eta \quad (0 < \eta < 1). \tag{66c}$$

The noise is a sample from a Gaussian random process with spectral density

$$N(\omega) = \begin{cases} N_o/2 & |\omega| \leqq 2\pi W, \\ \nu N_o/2 \quad (\nu \leqq 1) & |\omega| > 2\pi W. \end{cases} \tag{67}$$

Then the channel capacity is

$$C = C_{\eta,\nu} = W \ln \left(1 + (1-\eta)\frac{P_o}{N_o W}\right) + \frac{\eta P_o}{\nu N_o}. \tag{68}$$

Theorem 4: (Model 3) Let the allowable signal set be $a_3(T,W,P_o)$ *the set of functions* $s(t)$ *satisfying*

$$s(t) = 0, \qquad |t| \geqq T/2, \tag{69a}$$

$$\| s \|^2 \leqq P_o T, \tag{69b}$$

$$K_B'(s,2\pi W) \geqq 1 - \eta \quad (0 < \eta < 1), \tag{69c}$$

where K_B' *is the frequency concentration defined by (21). The noise is as in Theorem 2 (48). Then the channel capacity is*

$$C = C_\eta = W \ln \left(1 + \frac{P_o}{N_o W}\right) + \frac{\eta}{1-\eta}\frac{P_o}{N_o}. \tag{70}$$

Proofs of Theorems 3, 4: Since the proofs of Theorems 3 and 4 parallel that of Theorem 2 (which was given in detail above) we shall confine ourselves to a few remarks which will enable the interested reader to fill in the details on his own.

Theorem 3: In the direct half we consider, as in the proof of Theorem 2, a jointly-constrained product channel. In this case it is a type-2 channel with parameters

$$\alpha_1 = 2W(1-\delta), \quad \alpha_2 = 0, \quad \alpha_3 = \infty, \quad P = P_o,$$
$$N_1 = \frac{N_o}{2}(1-\xi), \quad N_2 = \frac{\nu N_o}{2}, \quad \hat{\eta} = \eta - \sigma, \tag{71}$$

where $\xi,\delta,\sigma > 0$ are "small". In the present proof, this channel plays the role that the type-1 channel played in the proof of the direct half of Theorem 2. Since $\alpha_2 = 0$, we may write a channel input as $\mathbf{x} = (\mathbf{x}^{(1)},\mathbf{x}^{(3)})$. Corresponding to $\mathbf{x}$ we construct an input signal for our time-continuous channel as

$$s(t) = D_T\left[\sum_{k=1}^{2W(1-\delta)=\alpha_1 T} x_k^{(1)} \frac{\psi_k(t)}{\sqrt{\lambda_k}} + \sum_{k=1}^{\infty} x_k^{(3)} \frac{\psi_{k+W(1-\delta)T}}{\sqrt{\lambda_{k+2W(1-\delta)T}}}\right]. \tag{72}$$

where the ψ_k are PSWF's, the λ_k the associated eigenvalues (40), and D_T the time-truncation operator (14). Equation (72) replaces (52) in the proof of Theorem 2. It is easily verified that signals of the form (72) belong to $a_1(T, W, P_o)$ as defined by (66). If we write the noise in the expansion of (46) we can, as in Theorem 2, establish the equivalence of the time-discrete and time-continuous channels, and establish the direct-half of Theorem 3. The weak converse is proved in a similar manner, the jointly-constrained product channel employed here being of type-2 with parameters

$$\alpha_1 = 2W(1-\delta), \quad \alpha_2 = 4W\delta, \quad \alpha_3 = \infty, \quad P = P_o,$$
$$N_1 = \frac{N_o}{2}, \quad N_2 = \frac{N_o}{2}, \tag{73}$$
$$N_3 = (\nu - \xi)\frac{N_o}{2}, \quad \eta = \frac{\eta}{1-\sigma},$$

where again $\delta,\xi,\sigma > 0$ are "small".

Theorem 4: For the direct-half we consider a type-3 jointly-constrained product channel with parameters

$$\alpha_1 = 2W(1-\delta), \quad \alpha_2 = \infty, \quad P = P_o, \quad N = \frac{N_o}{2}, \quad \hat{\eta} = \eta - \sigma. \tag{74}$$

The signals are constructed from vectors $\mathbf{x}$ as in (72). For the converse we use a type-3 channel with parameters

$$\alpha_1 = 2W(1+\delta), \quad \alpha_2 = \infty, \quad P = P_o, \quad N = \frac{N_o}{2}, \quad \hat{\eta} = \frac{\eta}{1-\sigma}. \tag{75}$$

APPENDIX A

Proof of Theorem 1

We shall give a proof of Theorem 1 for the type-1 jointly-constrained product channel only. The proofs for types 2 and 3 are similar.

The proof as usual is in two parts.

A.1 *Direct Half*

We set $P_1 = (1-\beta)P$, $P_2 = \beta P$ and consider codes for the ordinary product channel (Section 4.1). If (30) is satisfied for all code words with these values of P_1 and P_2, then the joint constraint (33a) is also satisfied. Further since $\beta \leqq \hat{\eta}$, (33b) is also satisfied. Hence the direct half of Lemma A for the ordinary product channel implies that any rate less than $C_1(P_1) + C_2(P_2) = C_1((1-\beta)P) + C_2(\beta P)$ is permissible, and the direct-half of Theorem 1 follows.

A.2 *Converse*

Let us define $C^* = C_1((1-\beta)P) + C_2(\beta P)$. We must show that any rate $R > C^*$ is not permissible. Let us assume the contrary, i.e.; for some $R = C^* + \varepsilon (\varepsilon > 0)$, there exists a sequence of numbers $\{T_i\}_{i=1}^{\infty}$ where $T_i \to \infty$ as $i \to \infty$, and a corresponding sequence of codes for the jointly constrained product channel (satisfying (33a) and (33b), with parameters $\hat{\eta}$ and P); with the ith code ($i = 1,2, \cdots$) having parameter $T = T_i$ and e^{RT_i} code words, and error probability $P_e = P_e^{(i)}$ where $P_e^{(i)} \to 0$ as $i \to \infty$.

Since $C_1(x)$ is uniformly continuous on the closed interval $[O,P]$, let us choose an integer J_o (sufficiently large) so that

$$\left| C_1(x) - C_1\left(x - \frac{\hat{\eta}P}{J_o}\right) \right| < \frac{\varepsilon}{2}, \qquad 0 \leqq x \leqq \hat{\eta}P. \tag{76}$$

We now partition the ith code ($i = 1,2, \cdots$) into J_o classes $S_i(j)$ ($j = 1,2, \cdots, J_o$). A code word $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$ in the ith code will belong to the jth class $S_i(j)$, according as the energy of its second component satisfies

$$(j-1)\frac{\hat{\eta}PT}{J_o} < \sum_{k=1}^{n_2} [x_k^{(2)}]^2 \leqq \frac{j\hat{\eta}PT_i}{J_o}, \qquad j = 1,2, \cdots, J_o. \tag{77}$$

Since $\mathbf{x}^{(2)}$ satisfies (33b), each code word belongs to exactly one class.

(To be precise, we assign code words for which the energy in $\mathbf{x}^{(2)}$ is zero to class $S_i(1)$.)

For each $i (i = 1,2, \cdots)$, let S_i^* be the subcode of the ith code (with parameter $T = T_i$) consisting of the class $S_i(j) (j = 1,2, \cdots, J_o)$ containing the most members. Since S_i^* is the largest class in a partition of a code with e^{RT_i} code words into J_o classes, the number of code words in $S_i^* \geqq e^{RT_i}/J_o$, so that the corresponding transmission rate for S_i^* is

$$R^* \geqq R - \frac{1}{T_i} \ln J_o . \tag{78}$$

Further, since S_1^* is a subcode of the ith code (which has error probability $P_e^{(i)}$), the error probability of S_i^* is not more than $P_e^{(i)}$.

Since there are a finite number (J_o) of classes in the partition of the ith code $(i = 1,2, \cdots)$, there must be at least one j_o $(1 \leqq j_o \leqq J_o)$ such that for an infinite number of i, the largest partition S_i^* is the j_oth partition $S_i(j_o)$. Let $(i_1, i_2, \cdots)$ be the subsequence of i's for which $S_i^* = S_i(j_o)$. Thus the $\{S_{i_t}^*\}_{t=1}^{\infty}$ are a sequence of codes with rate R^* satisfying (78), and error probability not more than $P_e^{(i_t)}$, where $P_e^{(i_t)} \to 0$ as $t \to \infty$. Further, if a code word $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \varepsilon S_{i_t}^*$, it belongs to the class $S_{i_t}(j_o)$, so that from (77) the energy of the second component satisfies

$$E(\mathbf{x}^{(2)}) = \sum_{k=1}^{n_2} [x_k^{(2)}]^2 \leqq \frac{j_o \hat{\eta} P T_{i_t}}{J_o} , \tag{79}$$

and from (77) and (33a), the energy of the first component satisfies

$$E(\mathbf{x}^{(1)}) = \sum_{k=1}^{n_1} [x_k^{(1)}]^2 \leqq \left[1 - \frac{(j_o - 1)\hat{\eta}}{J_o}\right] P T_{i_t} . \tag{80}$$

We conclude that $\{S_{i_t}^*\}_{t=1}^{\infty}$ is a sequence of codes which satisfy the constraints for the ordinary product channel (30) with parameters

$$P_1 = [1 - \{(j_o - 1)/J_o\}\hat{\eta}]\, P \quad \text{and} \quad P_2 = (j_o \hat{\eta}/J_o)\, P.$$

Since the error probability for $S_{i_t}^*$, $P_e^{(i_t)} \to 0$ as $t \to \infty$, we conclude that the rate R^* is a permissible rate for the ordinary product channel. By the converse half of Lemma A we have that R^* does not exceed the capacity of this product channel, i.e.,

$$R^* \leqq C_1\left(\left(1 - \frac{(j_o - 1)}{J_o}\hat{\eta}\right)P\right) + C_2\left(\frac{j_o \hat{\eta}}{J_o} P\right), \tag{81}$$

where $C_i(x) (i = 1,2)$ is defined by (37b). Applying (76) to (81) we obtain

$$R^* \leqq C_1((1-\delta)P) + C_2(\delta P) + \frac{\varepsilon}{2}, \tag{82}$$

where $\delta = j_o\hat{\eta}/J_o$. Now it follows immediately (by differentiation) from the definition of $C_1(x)$ and $C_2(x)$ that if $\alpha_2 \geqq \alpha_1$, $f(\delta) \triangleq C_1((1-\delta)x) + C_2(\delta x)$ is an increasing function for δ for $\delta < \alpha_2/(\alpha_1 + \alpha_2)$, and $f(\delta)$ is a decreasing function of δ for $\delta > \alpha_2(\alpha_1 + \alpha_2)$. We conclude that since $\delta = (j_o/J_o)\hat{\eta} \leqq \hat{\eta}$,

$$C_1((1-\delta)P) + C_2(\delta P) \leqq C_1((1-\beta)P) + C_2(\beta P) = C^*, \tag{83}$$

where $\beta = \min(\hat{\eta}, \alpha_2/(\alpha_1 + \alpha_2))$. Combining (78), (82) and (83), we obtain

$$R \leqq C^* + \frac{\varepsilon}{2} + \frac{1}{T_{i_t}} \ln J_o. \tag{84}$$

If we let $t \to \infty$, then $T_{i_t} \to \infty$ and have from (84)

$$R \leqq C^* + \frac{\varepsilon}{2}.$$

But $R = C^* + \varepsilon$, and the contradiction establishes the weak converse to Theorem 1.

APPENDIX B

Proof of Theorem 5

Theorem 5: Let $\mathfrak{a}(T,W,P_o)$ *be the set of all* $s(t)$ *satisfying*

$$(i) \qquad s(t) = 0, \qquad |t| > T/2, \tag{85a}$$

$$(ii) \qquad \|s\|^2 \leqq P_o T, \tag{85b}$$

$$(iii) \qquad K_B(s,2\pi W) \geqq 1 - \eta \quad (0 < \eta < 1). \tag{85c}$$

Let the Gaussian noise be additive with spectral density $N(\omega)$ *where*

$$\int_{-\infty}^{\infty} N(\omega)d\omega = \bar{N} < \infty \tag{86}$$

Then $C_\eta = \infty$ *(all* η*).*

Proof: Let $R > 0$ and $\varepsilon > 0$, and $\eta (0 < \eta \leqq 1)$ be specified and fixed. We shall construct a code satisfying (85) with $M = e^{RT}$ code words with error probability $P_e \leqq \varepsilon$.

To begin with let us choose T sufficiently large so that

$$\frac{1}{\sqrt{4\pi RT}} \leqq \varepsilon, \tag{87a}$$

$$\lambda_1 \geqq 1 - \frac{\eta}{2}, \tag{87b}$$

where λ_1 is the first PSWF eigenvalue (40). With T fixed we now construct the code.

Let us expand the noise in a series of PSWF's

$$z(t) = \sum_{k=1}^{\infty} z_k \frac{\psi_k(t)}{\sqrt{\lambda_k}}, \qquad -\frac{T}{2} \leqq t \leqq \frac{T}{2}, \tag{88}$$

where

$$z_k = \int_{-T/2}^{T/2} z(t) \frac{\psi_k(t)}{\sqrt{\lambda_k}}\, dt, \tag{89}$$

and the $\{z_k\}_{k=1}^{\infty}$ are Gaussian random variables with mean zero, but not necessarily independent.

Now from (86) we have

$$\mathcal{E} \int_{-T/2}^{T/2} z^2(t)dt = \bar{N}T, \tag{90}$$

where "$\mathcal{E}$" denotes expectation. From the orthogonality of the PSWF's (41) we have from (88)

$$\bar{N}T = \mathcal{E} \int_{-T/2}^{T/2} z^2(t)dt = \sum_{k=1}^{\infty} \mathcal{E}(z_k^2). \tag{91}$$

Thus we can find an integer K sufficiently large so that

$$\mathcal{E}(z_{K+i}^2) \leqq \frac{\eta P_o}{16R}, \qquad i = 1,2, \cdots, M. \tag{92}$$

With K so chosen, let the M code words be

$$s_i(t) = D_T\left[\sqrt{P_oT\left(1 - \frac{\eta}{2}\right)}\frac{\psi_1(t)}{\sqrt{\lambda_1}} + \sqrt{\frac{\eta}{2}P_oT}\frac{\psi_{K+i}(t)}{\sqrt{\lambda_{K+i}}}\right], \qquad i = 1,2, \cdots, M \tag{93}$$

Let us first verify that $s_i(t)$, as given by (93), satisfies (85). Equation (85a) follows from the definition of D_T (14). From the orthogonality of PSWF's (41) we have

$$\| s_i \|^2 = \left(1 - \frac{\eta}{2}\right) P_oT + \frac{\eta}{2} P_oT = P_oT \tag{94}$$

so that (85b) is satisfied. Finally,

$$\begin{aligned} K_B(s_i, 2\pi W) &= \frac{P_o T\left(1-\frac{\eta}{2}\right)\lambda_1 + \frac{\eta}{2} P_o T \lambda_{K+i}}{\| s_i \|^2} \\ &\geqq \frac{P_o T}{P_o T}\left(1-\frac{\eta}{2}\right)\lambda_1 \geqq \left(1-\frac{\eta}{2}\right)\left(1-\frac{\eta}{2}\right) \\ &\geqq 1-\eta, \end{aligned} \tag{95}$$

where the next to last inequality follows from (87b). Thus $s_i(t) \varepsilon$ $\mathfrak{a}(T,W,P_o)$. It remains to show that $P_e \leqq \varepsilon$.

We can write the received signal $y(t)$ in a Fourier series in PSWF's

$$y(t) = S_i(t) + z(t) = \sum_{k=1}^{\infty} y_k \frac{\psi_k(t)}{\sqrt{\lambda_k}}, \tag{96}$$

where the y_k are recoverable from $y(t)$ by integration. Say that the receiver disregards all the y_k except $y_{K+1}, y_{K+2}, \cdots, y_{K+M}$. We may write

$$y_{K+j} = \begin{cases} z_{K+j} + \sqrt{\frac{\eta}{2} P_o T}, & j = i, \\ z_{K+j}, & j \neq i, \end{cases} \tag{97}$$
$$(j = 1,2, \cdots, M).$$

If y_{K+i} is the maximum of the $\{y_{K+j}\}_{j=1}^{M}$, the receiver decodes $y(t)$ as $s_i(t)$. Thus if code word i is transmitted, the error probability is

$$\begin{aligned} P_{ei} &= P_r \bigcup_{j \neq i} \left[z_{K+j} > z_{K+i} + \sqrt{\frac{\eta}{2} P_o T} \right] \\ &\leqq M P_r \left[z_{K+j} - z_{K+i} > \sqrt{\frac{\eta}{2} P_o T} \right]. \end{aligned} \tag{98}$$

Now $z_{K+j} - z_{K+i}$ is Gaussian, with mean zero and variance

$$\begin{aligned} \mathcal{E}((z_{K+j} - z_{K+i})^2) &\leqq [E(z_{K+j}^2)^{\frac{1}{2}} + E(z_{K+i}^2)^{\frac{1}{2}}]^2 \\ &\leqq \frac{\eta P_o}{4R}, \end{aligned} \tag{99}$$

where the last inequality follows from (92). Thus (98) becomes

$$P_e \leqq M \operatorname{erf}\ (-\sqrt{2RT}), \tag{100}$$

where

$$\text{erf } (x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

is the cumulative error function. Since erf $(-x) \leqq e^{-x^2/2}/(\sqrt{2\pi x})$, (100) yields with the help of (87a)

$$P_e \leqq \frac{e^{RT} e^{-RT}}{\sqrt{4\pi RT}} = \frac{1}{\sqrt{4\pi RT}} \leqq \varepsilon. \tag{101}$$

Thus the theorem is proven.

APPENDIX C

In this appendix we verify inequality (22)

$$K_B(x,2\pi W) \geqq 1 - 2\sqrt{\frac{1 - K_B'(x,2\pi W)}{K_B'(x,2\pi W)}}, \tag{102}$$

where K_B is defined by (13a) and K_B' by (21). Let $f(t)$ be a function with Fourier transform $F(\omega)$, and define the operator B by

$$g = Bf, \tag{103a}$$

where

$$g(t) = \frac{1}{2\pi} \int_{-2\pi W}^{2\pi W} F(\omega) e^{i\omega t} d\omega. \tag{103b}$$

Thus Bf is the result of passing f through an ideal low-pass filter with bandpass W cycles per second. Then

$$K_B(f,2\pi W) = \frac{\| Bf \|^2}{\| f \|^2}. \tag{104}$$

Say that $x(t) = 0, |t| \leqq T/2$ and $\| x \|^2 < \infty$. We assume that we may write $x = D_T \hat{x}$, where $\hat{x}$ is bandlimited to W cycles per second. (If we cannot then $K_B'(x,2\pi W) = 0$, and (102) follows immediately.) Let us write

$$\hat{x}(t) = x(t) + y(t), \tag{105}$$

where $y(t) = 0, |t| \leqq T/2$. Then

$$\| \hat{x} \|^2 = \| x \|^2 + \| y \|^2, \tag{106}$$

and from the definition of K_B',

$$K_B'(x,2\pi W) = \frac{\| x \|^2}{\| \hat{x} \|^2}. \tag{107}$$

Hence, from (107) and (106),

$$\frac{\| y \|^2}{\| x \|^2} = \frac{1 - K_B'(x,2\pi W)}{K_B'(x,2\pi W)}. \tag{108}$$

Now, since $\hat{x}$ is bandlimited, $B\hat{x} = \hat{x}$ and we have

$$\begin{aligned} \| \hat{x} \|^2 &= \| B\hat{x} \|^2 = \| Bx + By \|^2 \leqq [\| Bx \| + \| By \|]^2 \\ &\leqq [\| Bx \| + \| y \|]^2 = \| Bx \|^2 + \| y \|^2 + 2 \| Bx \| \| y \|. \end{aligned} \tag{109}$$

Combining (106) and (109) we have

$$\| x \|^2 + \| y \|^2 = \| \hat{x} \|^2 \leqq \| Bx \|^2 + \| y \|^2 + 2 \| Bx \| \| y \|, \tag{110}$$

so that (from (104))

$$K_B(x,2\pi W) = \frac{\| Bx \|^2}{\| x \|^2} \geqq 1 - 2 \frac{\| Bx \| \| y \|}{\| x \|^2} \geqq 1 - 2 \frac{\| y \|}{\| x \|}. \tag{111}$$

Finally, from (108) and (111) we have

$$K_B(x,2\pi W) \geqq 1 - 2 \sqrt{\frac{1 - K_B'(x,2\pi W)}{K_B'(x,2\pi W)}}. \tag{112}$$

This is inequality (102).

APPENDIX D

The Capacity of Model 4

To establish the capacity of the channel defined by Model 4 we must, as always, prove a direct-half and converse. In this appendix we give an outline of the proof of the direct-half, and a remark about the proof of the converse.

D.1 *Direct-Half*

Let $R < W \ln [1 + (P_o/N_oW)]$ be given. We show here that for ν sufficiently small we may construct codes for Model 4 with rate R and with vanishing error probability (as $T \to \infty$). By the continuity of the "ln" function we may find a $\delta > 0$, $a > 0$ sufficiently small so that

$$R < W(1 - \delta) \ln \left[1 + \frac{P_o(1 - a)}{N_oW(1 - \delta)} \right] = C^*. \tag{113}$$

We observe that C^* is the capacity of a single time-discrete channel

(Section 2.1) with parameters

$$P = P_o(1 - a), \qquad N = N_o/2, \qquad \alpha = 2W(1 - \delta). \quad (114)$$

Since $R < C^*$, we can find a code $\mathcal{C} = \{\mathbf{x}_i\}_{i=1}^{M}$ for this time-discrete channel (so that $E(\mathbf{x}_i) \leqq P_o(1 - a)T$) with $M = e^{RT}$ code words, and with error probability given that $\mathbf{x}_i$ is transmitted ($i = 1,2, \cdots, M$) (using the minimum distance decoder)

$$\begin{aligned} P_{ei} &= \Pr \bigcup_{j \neq i} [d_E(\mathbf{x}_i, \mathbf{y}) > d_E(\mathbf{x}_j, \mathbf{y})] \\ &= \Pr \bigcup_{j \neq i} \left[\| \mathbf{z}^{ij} \| > \frac{d_{ij}}{2} \right] = e^{-\beta T + o(T)}, \end{aligned} \quad (115)$$

where $\mathbf{y}$ is the received vector, $d_E(\mathbf{u},\mathbf{v})$ is the Euclidean distance between n-vectors $\mathbf{u}$ and $\mathbf{v}$, $d_{ij} = d_E(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{z}^{ij}$ is the projection of the noise vector $\mathbf{z}$ on the line passing through code words $\mathbf{x}_i$ and $\mathbf{x}_j$, and $\| \mathbf{u} \| = [E(\mathbf{u})]^{\frac{1}{2}}$ is the square root of the sum of the squares of the components of $\mathbf{u}$. The exponent β has been estimated by Shannon.[3] Since $\| \mathbf{z}^{ij} \|$ is a Gaussian random variable with mean zero and variance $N_o/2$ we may lower bound P_{ei} by

$$P_{ei} \geqq \Pr \left[\| \mathbf{z}^{ij} \| > \frac{d_{ij}}{2} \right] = \operatorname{erf} \left(- \frac{d_{ij}}{\sqrt{2N_o}} \right), \quad (116)$$

$$(j = 1,2, \cdots, M \qquad j \neq i)$$

where

$$\operatorname{erf} x = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$

is the cumulative error function. Since for large x,

$$\operatorname{erf} (-x) \approx \left(\frac{1}{\sqrt{2\pi}x} \right) e^{-x^2/2},$$

(115) and (116) yield for large T

$$d_{ij}^2 \geqq 4\beta N_o T, \qquad i,j = 1,2, \cdots, M \qquad i \neq j. \quad (117)$$

From the code $\mathcal{C}$, let us construct a new code $\hat{\mathcal{C}} = \{\hat{\mathbf{x}}_i\}_{i=1}^{M}$, where

$$\hat{\mathbf{x}}_i = \frac{1}{1 - a} \mathbf{x}_i, \qquad i = 1,2, \cdots, M. \quad (118)$$

Thus the members $\hat{\mathbf{x}}_i$ of $\hat{\mathcal{C}}$ satisfy

$$E(\hat{\mathbf{x}}_i) \leqq P_o T. \quad (119)$$

Let us now assume that there are two noises in the channel, i.e., the

noise vector $\mathbf{z} = \mathbf{z}_1 + \mathbf{z}_2$. The first noise $\mathbf{z}_1$ is the usual spherical Gaussian noise (with variance $N_o/2$), and the second $\mathbf{z}_2$ is an unknown n-vector ($n = \alpha T = 2W(1 - \delta)T$) for which we require only

$$E(\mathbf{z}_2) \leqq \nu N_o W(1 - \delta)T = \nu \frac{N_o}{2} n. \tag{120}$$

We place no other restrictions on the probability structure of $\mathbf{z}_2$. The vector $\mathbf{z}_2$ may depend on the code $\hat{\mathfrak{C}}$, the code word transmitted and the value of $\mathbf{z}_1$. The noise vector $\mathbf{z}_1$ corresponds to the noise function $z_1(t)$ in Model 4, and the noise vector $\mathbf{z}_2$ corresponds to $z_2(t)$ in Model 4. If we use $\hat{\mathfrak{C}}$ on the time-discrete channel with this noise and use the minimum distance decoder, we have an error probability given that $\hat{\mathbf{x}}_i$ is transmitted

$$\begin{aligned} \hat{P}_{ei} &= \Pr \bigcup_{j \neq i} \left[d_E^{\,2}(\hat{\mathbf{x}}_i, \hat{\mathbf{y}}) > d_E(\hat{\mathbf{x}}_j, \hat{\mathbf{y}}) \right] \\ &= \Pr \bigcup_{j \neq i} \left[\| (\mathbf{z}_1 + \mathbf{z}_2)^{ij} \| > \frac{\hat{d}_{ij}}{2} \right] \end{aligned} \tag{121}$$

where

$$\hat{d}_{ij} = d_E(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) = \frac{d_{ij}}{(1 - a)}.$$

Now since " $\| \ \|$ " is a norm

$$\| (\mathbf{z}_1 + \mathbf{z}_2)^{ij} \| \leqq \| \mathbf{z}_1^{\,ij} \| + \| \mathbf{z}_2^{\,ij} \| \leqq \| \mathbf{z}_1^{\,ij} \| + \sqrt{\nu N_o W(1 - \delta)T}. \tag{122}$$

Thus the event

$$\begin{aligned} &\left[\| (\mathbf{z}_1 + \mathbf{z}_2)^{ij} \| > \frac{\hat{d}_{ij}}{2} \right] \\ &\qquad \subseteq \left[\| \mathbf{z}_1^{\,ij} \| > \frac{d_{ij}}{2(1 - a)} - \sqrt{\nu N_o W(1 - \delta)T} \right], \end{aligned} \tag{123}$$

where "$\subseteq$" denotes set inclusion. Now we would like to say that the right member of (123)

$$\left[\| \mathbf{z}_1^{\,ij} \| > \frac{d_{ij}}{(1 - a)2} - \sqrt{\nu N_o W(1 - \delta)T} \right] \subseteq \left[\| \mathbf{z}_1^{\,ij} \| > \frac{d_{ij}}{2} \right]. \tag{124}$$

If this is so, then $\hat{P}_{ei} \leqq P_{ei} \to 0$ as $T \to \infty$. In fact (124) is satisfied if

$$\frac{d_{ij}}{2} \leqq \frac{d_{ij}}{2(1 - a)} - \sqrt{\nu N_o W(1 - \delta)T}, \tag{125}$$

or

$$\nu \leqq \frac{d_{ij}}{\sqrt{4N_oW(1-\delta)T}}\left(\frac{a}{1-a}\right). \tag{126}$$

Now from (117), $d_{ij} \geqq \sqrt{4\beta N_oT}$ so that if

$$\nu \leqq \sqrt{\frac{\beta}{w(1-\delta)}}\left(\frac{a}{1-a}\right), \tag{127}$$

(126) is satisfied. Hence $\hat{P}_{ei} \xrightarrow{T} 0$.

If we now make the same correspondence between the time-continuous channel and the time-discrete channel which was made in the proof of Theorem 3, we deduce the existence of codes for Model 4 {with rate $R < W \ln [1 + (P_o/N_oW)]$} with $P_e \to 0$ as $T \to \infty$ (provided ν is sufficiently small — the choice of ν depending on W, P_o/N_o, and R). Note that this construction was done for any η. Thus we have shown in effect that the capacity of Model 4 is

$$C = C_{\eta,\nu} \geqq W \ln\left(1 + \frac{P_o}{N_oW}\right) + \varepsilon_1(\nu), \tag{128}$$

where $\varepsilon_1(\nu) \to 0$ as $\nu \to 0$ independent of η.

D.2 *Converse*

The proof of the converse also parallels the proofs of the converse halves of Theorems 2, 3, and 4. However, since the noise may depend on the entire code and decoding scheme used (which is not the usual assumption of information theory coding theorems), it is necessary to go back and re-prove Theorem 1 (which in turn depends on Lemma A) for this new situation. Although this task is not a terribly difficult one it is rather tedious and we shall side step this chore here. It will suffice to state the version of Lemma A which is required here and to leave the rest of the proof to the interested reader.

Lemma A′: Let us say that we are given time-discrete channel as defined in Section I (with parameters α,P) where the noise vector is $\mathbf{z} = \mathbf{z}_1 + \mathbf{z}_2$ where $\mathbf{z}_1$ is the usual spherical Gaussian noise with variance N and $\mathbf{z}_2$ is an unknown vector for which require only

$$E(\mathbf{z}_2) \leqq \xi T. \tag{129}$$

We place no other restriction on the probability structure of $\mathbf{z}_2$. The noise vector $\mathbf{z}_2$ may depend on the entire code and decoding scheme, the code word transmitted and the value of $\mathbf{z}_1$. We define the error probability P_e as we did in (26) for Model 4 and do likewise for the capacity. Let $C(\alpha,P,N,\xi)$ be the capacity of this channel.

Now consider the product of r time-discrete channels as in Section 4.1 with parameters (α_i, P_i, N_i) $i = 1,2, \cdots, r$. *Here too, we assume a second noise vector*

$$\mathbf{z}_2 = (\mathbf{z}_2^{(1)}, \mathbf{z}_2^{(2)}, \cdots, \mathbf{z}_2^{(r)}), \tag{130}$$

which is unknown but must satisfy

$$\sum_{i=1}^{r} E(\mathbf{z}_2^{(i)}) \leqq \xi T, \tag{131}$$

and as above may depend on the entire code and decoding scheme, the code word transmitted, and the values of the spherical Gaussian noises.

Lemma A′ states that the capacity C^* *of this channel satisfies*

$$C^* \leqq \sum_{i=1}^{r} C(\alpha_i, P_i, N_i, \gamma_i \xi), \tag{132a}$$

where

$$\sum_{i=1}^{r} \gamma_i = 1. \tag{132b}$$

APPENDIX E

Equivalence of Time-Discrete and Time-Continuous Models

In this appendix, we give some details on the validity of the equivalence of the time-discrete and time-continuous channel models which is the key to the proofs of our capacity theorems.

To begin with, let us consider the direct-half of our theorems. In these proofs we deduce the existence of time-continuous coding and decoding schemes from the existence of time-discrete coding and decoding schemes. To be specific let us consider the proof of the direct half of Theorem 2. We may omit the reference to the Karhunen-Loeve expansion (5.10) and consider the received signal $y(t) = s_i(t) + z(t)$. Now it follows from Loeve (Ref. 9, p. 472, A) that

$$\mathcal{E} \int_{-T/2}^{T/2} z^2(t)dt = \int_{-T/2}^{T/2} R(0)dt = N_0 W T < \infty, \tag{133}$$

so that with probability 1, $z(t)$ and, therefore, $y(t)$ is square-integrable. It then follows that the integrals

$$y_k^{(1)} = \frac{1}{\sqrt{\lambda_k}} \int_{-T/2}^{T/2} y(t)\psi_k(t)dt \quad \text{and}$$

$$y_k^{(2)} = \frac{1}{\sqrt{\lambda_{\alpha_1 T+k}}} \int_{-T/2}^{T/2} y(t)\psi_{\alpha_1 T+k}(t)dt \tag{134}$$

(where $\psi_k(t)$ and the λ_k are the kth PSWF and eigenvalue, respectively) exist for all k with probability 1. Further, it follows directly on substituting $y(t) = s_i(t) + z(t)$ into (134) that

$$y_k^{(i)} = s_{ik} + z_k^{(i)}, \qquad i = 1,2, \tag{135}$$

where the $z_k^{(i)}$ are independent normally distributed random variables with mean zero and variance $N_0/2$. Thus, the decoder for the time-continuous code may obtain the $y_k^{(i)}$ from the $y(t)$ and make use of the decoding scheme for the time-discrete code and obtain the same error probability. Hence, the direct-half of this and the subsequent theorems is valid.

Let us now consider the converse half of our theorems. In each of these proofs we assume that for a fixed rate R exceeding capacity, we are given a sequence of codes for the time-continuous channel with error probability P_e. We must show that P_e is bounded away from zero. To do this we deduce the existence of a corresponding sequence of codes with rate R and error probability P_e for a time-discrete channel with the same capacity as the time-continuous channel. Since we can invoke a converse for this time-discrete channel (Theorem 1), we then conclude that P_e is bounded away from zero. We will now show how to make this correspondence precise. Again let us refer specifically to the proof of Theorem 2, the others following similarly.

Let $\{s_i(t)\}_{i=1}^{M}$ be the code for the time-continuous channel, and $\mathbf{x} = (\mathbf{x}^{(1)},\mathbf{x}^{(2)})$ be the corresponding input to the time-discrete (product) channel. Further, we may write the noise signal $z(t)$ and the received signal $y(t)$ in Fourier series in PSWF's where, as above, all the coordinates are finite with probability 1. We then let $\mathbf{z} = (\mathbf{z}^{(1)},\mathbf{z}^{(2)})$ and $\mathbf{y} = (\mathbf{y}^{(1)},\mathbf{y}^{(2)})$ be the vectors whose coordinates are the coefficients in these expansions. We can easily show that

$$\mathbf{y} = \mathbf{x} + \mathbf{z}, \tag{136}$$

where the coordinates of $\mathbf{z}$ are independent random variables with mean zero and variance $N_0/2$. Thus, we have established the correspondence of the time-continuous and time-discrete channels and codes. We must now show that the time-continuous and time-discrete codes have the same error probability. In other words, we mush show that there exists a decoding scheme for the $\mathbf{y}$ which has the same error probability as the decoding scheme for the continuous received signal $y(t)$. We proceed as follows:

Let $\mathfrak{B}$ be the usual (Kolmogorov) σ-algebra on $\mathfrak{L}_2[-\bar{T}/2,\bar{T}/2]$, i.e., $\mathfrak{B}$ is the σ-algebra generated by the "intervals" of the form

$$\{y(t): y(t_1) \leqq \rho_1, y(t_2) \leqq \rho_2, \cdots, y(t_n) \leqq \rho_n\}.$$

Corresponding to the code for the time-continuous channel $\{s_i(t)\}_{i=1}^{M}$, we define M probability measures $P_1, P_2, \cdots, P_M$ on $\mathfrak{B}$ as follows. If $B \in \mathfrak{B}$, then

$$P_i(B) = \text{Prob}\,[(s_i(t) + z(t)) \in \mathfrak{B}], \tag{137}$$

where the probability in (137) is computed for $z(t)$, a noise sample function. A decoding-coding rule for this code is a set of M disjoint $\Lambda_i \in \mathfrak{B}$ $(i = 1,2, \cdots, M)$, called decoding regions. The error probability given that $s_i(t)$ is transmitted is

$$P_{ei} = 1 - P_i(\Lambda_i). \tag{138}$$

Now let $\hat{\mathfrak{B}} \subseteq \mathfrak{B}$ be the sub-σ-algebra on $\mathfrak{L}_2[-T/2,T/2]$, consisting of those sets determined by the coefficients of a representation of a function in PSWF's. That is, if $y(t) \in \mathfrak{L}_2(-T/2,T/2)$, let

$$y_k^{(1)} = \frac{1}{\sqrt{\lambda_k}} \int_{-T/2}^{T/2} y(t)\psi_k(t)dt \quad \text{and} \quad y_k^{(2)} = \frac{1}{\sqrt{\lambda_k}} \int_{-T/2}^{T/2} y(t)\psi_{\alpha_1 T+k}(t)dt.$$

Then $\hat{\mathfrak{B}}$ is the σ-algebra generated by intervals of the form

$$\{y(t): y_{k_1}^{(1)} \leqq \rho_1^{(1)}, y_{k_2}^{(1)} \leqq \rho_n^{(1)}, \cdots, y_{k_m} \leqq \rho_m^{(1)},$$
$$y_{j_1}^{(2)} \leqq \rho_1^{(2)}, y_{j_2}^{(2)} \leqq \rho_2^{(2)}, \cdots, y_{j_n} \leqq \rho_n^{(2)}\}.$$

A decoding rule for a time-discrete code with M code words is a set of M disjoint $\hat{\Lambda}_i \in \hat{\mathfrak{B}}$ $(i = 1,2, \cdots, M)$ (decoding regions), and the error probability given that vector $\mathbf{x}_i$ ($\mathbf{x}_i$ is the representation of $s_i(t)$ in PSWF's) is transmitted is

$$\hat{P}_{ei} = 1 - P_i(\hat{\Lambda}_i).$$

Kadota [Ref. 10, Appendix D] has shown that for each $\Lambda_i \in \mathfrak{B}$, there exists a $\Lambda_i \in \hat{\mathfrak{B}}$ such that

$$P(\Lambda_i \Delta \hat{\Lambda}_i) = 0,$$

where Δ denotes "symmetric difference". Thus, if $\{\Lambda_i\}_{i=1}^{M}$ are the decoding regions for a time-continuous code we can find a set $\{\hat{\Lambda}_i \in \hat{\mathfrak{B}}\}_{i=1}^{M}$ of decoding regions for the corresponding time-discrete code such that the error probabilities $P_{ei} = \hat{P}_{ei}$.

We conclude that the error probability for the time-discrete code equals the error probability for the time-continuous code, and the converse is valid.

GLOSSARY

The following symbols are used throughout the paper:

M = the number of members of a code.
T = time required to transmit a code word.
R = $(1/T) \ln M$ = transmission rate in nats per second.
C = channel capacity = maximum "error free" rate.
P_{ei} = probability that the receiver makes an incorrect decoding decision when code word i is transmitted ($i = 1,2, \cdots, M$).
P_e = $(1/M) \sum_{i=1}^{M} P_{ei}$ = over-all error probability.
$\mathcal{E}(X)$ = expected value of the random variable X.
ψ_k, λ_k = kth prolate spheroidal wave function (PSWF) and eigenvalue respectively ($k = 1,2, \cdots$).

The following symbols are used in connection with time-discrete or time-continuous channels:

Time-Discrete Channels:

$\mathbf{x},\mathbf{y},\mathbf{z}$ = input, output, and noise vectors, respectively.
$n = \alpha T$ = dimension of above vectors, so that α is the rate at which the channel passes real numbers.
$E(\mathbf{x})$ = sum of the squares of the coordinates of the vector $\mathbf{x}$.
P = parameter constraining $E(\mathbf{x})$ ($\mathbf{x}$ is channel input).
N = variance of the normally distributed noise.
r = number of components in the product (or parallel combination) of channels.
$\mathbf{x}^{(i)}$, $\mathbf{y}^{(i)}$, $\mathbf{z}^{(i)}$ = input, output, and noise vectors, respectively for the ith component of a product of channels ($i = 1,2, \cdots, r$).
n_i, α_i, P_i, N_i = parameters n,α,P,N, respectively, for the ith component of a product of channels ($i = 1,2, \cdots, r$).
$\hat{\eta}$ = parameter constraining the relative values of $E(\mathbf{x}^{(i)})$ in the product of channels.

Time-Continuous Channels:

$s(t),y(t),z(t)$ = input, output, and noise signals, respectively.
$S(\omega)$ = Fourier transform of $s(t)$.

$$\| s \|^2 = \int_{-\infty}^{+\infty} s^2(t)dt = \text{"energy" of } s(t).$$

$$K_B(s,2\pi W) = \frac{1}{2\pi} \int_{-2\pi W}^{2\pi W} |S(\omega)|^2 \, d\omega / \| s \|^2$$

= (energy) concentration in frequency band 0-W cps.

$K_D(s, T) = \int_{-T/2}^{T/2} s^2(t)dt / \| s \|^2$ = (energy) concentration in time interval $[-(T/2), (T/2)]$.

$K_B'(s,2\pi W)$ = an alternate measure of frequency concentration defined by (21).

D_T = operator which truncates a signal outside the time interval $[-(T/2), (T/2)]$ (see (14)).

$\mathfrak{L}_2[-T/2,T/2]$ = the space of square integrable functions defined on $[-T/2,T/2]$.

W = bandwidth of channel.

P_o = average "power" of input signals.

N_o = one-sided spectral density of noise $z(t)$.

a = $a_i(T,W,P_o)$ = set of allowable channel input signals (for Model i, i = 1,2,3,4). These signals are approximately time-limited to T secs, approximately band-limited to W cps, and have energy not exceeding P_oT.

η = parameter which measures the extent to which signals in a are not strictly time or bandlimited.

ν = parameter which measures the extent to which the noise spectral density is not zero for $|\omega| > 2\pi W$.

ACKNOWLEDGMENT

I wish to thank D. Slepian for many stimulating discussions and helpful suggestions.

REFERENCES

1. Shannon, C. E., A Mathematical Theory of Communication, B.S.T.J., *27*, July and October, 1948, pp. 379–423, 623–656.
2. Shannon, C. E., Communication in the Presence of Noise, Proc. IRE, *37*, January, 1949, pp. 10–21.
3. Shannon, C. E., Probability of Error for Optimal Codes in the Gaussian Channel, B.S.T.J., *38*, May, 1959, pp. 611–656.
4. Ash, R. B., Capacity and Error Bounds for a Time-Continuous Gaussian Channel, Information and Control, *6*, March, 1963, pp. 14–27.
5. Ash, R. B., *Information Theory*, John Wiley and Sons, New York, 1965.
6. Wyner, A. D., The Capacity of the Product of Channels, submitted to Information and Control.
7. Slepian, D., Landau, H., and Pollak, H., Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty I, II, B.S.T.J., *40*, January, 1961, pp. 43–84.
8. Davenport, W. and Root, W., *Random Signals and Noise*, McGraw-Hill Book Company, Inc., New York, 1958.
9. Loeve, M., *Probability Theory*, D. Van Nostrand, Princeton, New Jersey, 1955.
10. Kadota, T. T., Optimum Reception of Binary Gaussian Signals, B.S.T.J., *43*, November, 1964, pp. 2767–2810.

An Insertion Loss, Phase and Delay Measuring Set for Characterizing Transistors and Two-Port Networks Between 0.25 and 4.2 gc

By D. LEED

(Manuscript received November 29, 1965)

A new insertion loss, phase and delay measurement tool has been developed for characterizing gigacycle bandwidth transistors and general two-port networks on a small signal basis over a frequency range from 0.25 to 4.2 gc. Maximum inaccuracies are 0.1 db, 0.6 degree (over a 40-db loss range), and 0.5 nanosecond (over a 20-db loss range). Above 2.0 gc, the errors may double.

The particular parameters selected for measurement are closely related to the scattering coefficients of the device under test, evaluated with respect to a 50-ohm impedance level. When measuring transistors, measurement data are corrected for the residuals of jig and bias fixtures. Transformation from the measured parameters to other sets (e.g., h, y, or z matrices) is routine.

In order to minimize "instrument zero-line" and eliminate errors from circuit drift, a rapid sampling technique sequentially compares the unknown with a high-frequency reference. Measurement accuracy is held substantially independent of test signal frequency by heterodyning the measurement information to a fixed IF, where detection is performed by "IF substitution", using adjustable standards of loss, phase, and delay. Substantial use of automatic control circuitry contributes to an easy and facile interface between machine and operator.

This paper discusses the operation and design of the test set and its use as a tool in characterizing transistors.

I. INTRODUCTION

A new measuring instrument has been developed for making insertion loss, phase, and envelope delay measurements between 0.25 and

4.2 gc. The development was stimulated by measurement requirements growing out of recent advances in gigacycle transistor technology and semiconductor amplification for high bit rate PCM systems. The new instrument extends to 4.2 gc many of the operational features embodied in lower frequency instruments previously reported.[1,2,3]

This new test set is intended for characterizing transistors and general two-port networks, either passive or active, in a coaxial mode between 50-ohm terminations. By use of appropriate transducers, measurement may be extended to noncoaxially terminated unknowns. Of particular interest is the measurement of transistors using a special jig designed to provide a smooth electrical transition between the coaxial geometry of the test set ports and the pig tail lead geometry of the transistor. Waveguide networks are measured with the help of well-matched coax-to-waveguide transducers.

Since the need for reliable transistor characterization over the UHF band was an especially strong stimulus to the development of the measurement set, Section II deals with the transistor measurement problem and how it was solved. This section discusses the merits of the selected plan of measurement, the design of the required jigs and fixtures, the abstraction of transistor parameters from the measurement data, and examples of the results obtained.

Section III deals with the performance of the test set and the basic principles of measurement which are employed.

Sections IV and V describe the measurement circuit in progressively finer detail, starting from a block diagram description. Design questions are discussed, together with approaches used in solution. The goals which motivated the design are summarized.

Among the prime goals was measurement of transistors under "small-signal" excitation. The instrument design was to promote a congenial interface between operator and machine, the aim being to simpify the measurement procedure. This serves both to speed up measurements and reduce operator error.

Validation of accuracy and estimates of the residual inaccuracy are covered in Section VI.

Equipment design features of particular interest are noted in Section VII.

II. TRANSISTOR CHARACTERIZATION

Transistor characterization with the new test set makes use of the following basic data:

(*i*) Insertion loss and phase of the transistor between nominal 50-

ohm generator and load terminations for both the forward and reverse direction of transmission.

(*ii*) Insertion loss and phase which result when first the input terminals and then the output terminals of the transistor are bridged across the transmission path formed by connecting the 50-ohm generator directly to the 50-ohm load. During these bridging measurements, the terminal pair *not* connected to the bridging plane is terminated with 50 ohms.

These four measurement parameters lead naturally to an "exterior" characterization of the transistor in terms of its scattering (s) parameters. Data under (*i*) produce directly s_{21}^{-1} and s_{12}^{-1}; measurements under (*ii*) relate to s_{11} and s_{22} through elementary bilinear transformations. The scattering parameters can be routinely transformed to any of the other usual 2 terminal-pair descriptions (h's, y's, or z's) or, with more work, interior descriptions (e.g., equivalent circuits) may be deduced from the measurement data.

At high frequencies, this method of measurement has several advantages over techniques which attempt to measure directly the h, y, or z parameters. The direct measurement of these parameters calls for the projection of ac shorts and opens to the ports of the transistor through the arms of an intervening jig. It is difficult to reflect these singular values of impedance to the transistor terminals because of the impedance distortions introduced by the parasite residuals of the jig paths and dc biasing arrangements. Moreover, the necessity for using resonant lines to achieve the shorts or opens makes for a high degree of frequency sensitivity and may cause the transistor to oscillate. All of these difficulties are overcome by measuring the transistor between resistive impedances of moderate (and realizable) magnitude. At an impedance level of 50 ohms, stray reactances that would ordinarily make the realization of very large or very small impedances impracticable up to 4000 mc can be compensated so that they produce only modest reflections.

In the previously reported work on transistor measurements up to 250 mc,[1] much effort was expended to develop very low reflection jigs and biasing apparatus. In the case of UHF measurements, where development problems are more severe, it has proved fruitful to pursue an alternative course. Only a modest effort was expended on the design of jigs and biasing fixtures, but these devices were carefully characterized in a series of intensive measurements, and an analytic program was worked out to correct the transistor measurement data for impedance and transmission defects of the measurement hardware.

This approach was attractive for a number of reasons. First, the difficulties of developing jigs and bias fixtures having reflection coefficients under 0.02 up to 3000 mc appeared extremely formidable when compared with the labor of characterizing this hardware and accounting analytically for its defects. Secondly, there is little advantage to be gained in synthesizing a totally reflection-free environment since, even under ideal conditions, the computational work in transforming the measurement data to other characterization sets is sufficiently laborious to warrant the use of a digital computer. It is straightforward to include in the computer program the compensation for the transmission and impedance defects of all auxiliary test devices. And finally, since the measurement set views the transistor through the comparatively long path contributed by the biasing fixtures and jigs, corrections for path length must be introduced even in the absence of reflection. The "transformer" effect of the jig paths during the bridging measurements, for example, alters the apparent impedance of the transistor as sensed by the test set, and must be accounted for.

2.1 *Transistor Jig and Bias Fixture*

The jig for mounting the transistor under test was designed to adapt between the coaxial measurement ports of the test set and the pig tail lead geometry of "TO-18" encapsulated transistors. The objective was to bring a 50-ohm measurement plane right up to the base of the transistor header. In view of the plan to account analytically for hardware defects, a modest reflection target of 0.1 was set up for purity of the nominal 50-ohm termination looking toward source or load from the base plane of the header over the frequency range from 0.25 to 3.0 gc. The transistor lead wires were not to be bent during insertion into the jig, and the jig was designed to allow the metal case of the transistor to be firmly grounded.

The basic principle in achieving a smooth transition consists of forming 50-ohm coax lines around the transistor lead wires inserted into the jig. Fig. 1 shows the actual jig worked out using this principle. Each of the two cylindrical holes in the top of the jig forms a 50-ohm transmission line with one of the terminal wires of the transistor. The lead wires penetrate a brief distance before plugging into a short metal cylinder which functions as the inner conductor segment of a larger bore, 50-ohm line. The cylinder is drilled with a long hole for storing up to $\frac{1}{2}$ inch of lead wire. Discontinuity capacitance caused by the diameter difference between the two sections of 50-ohm line has been compensated by an appropriate setback of the step in the inner conduc-

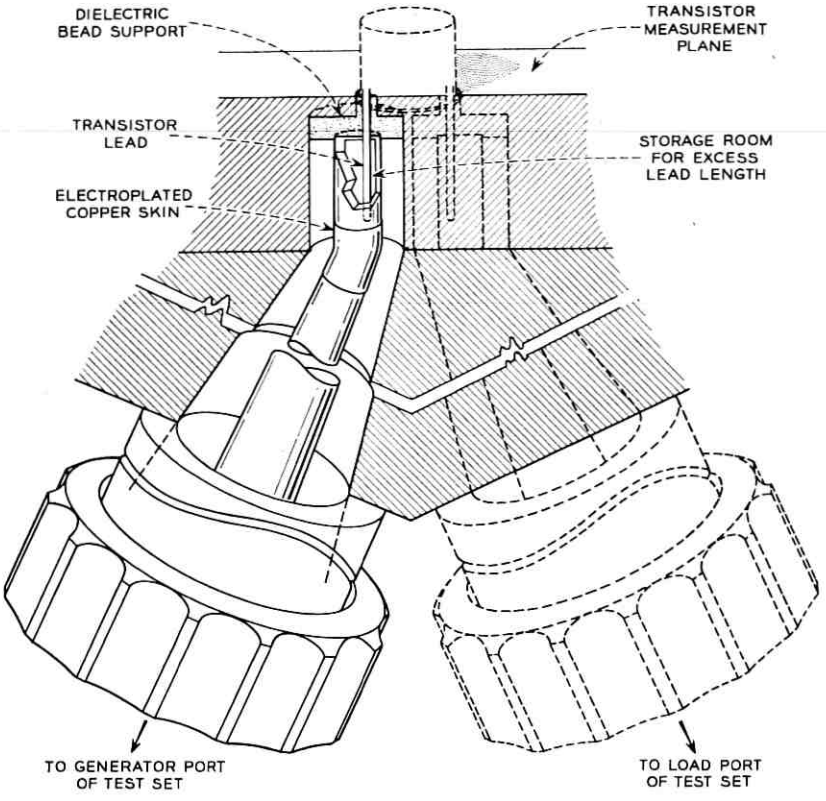


Fig. 1 — Transistor test jig for UHF measurements.

tors.[4] Conically tapered inner and outer conductors below the head piece of the jig slowly bring up the coaxial diameters to match those of a "Dezifix B" fitting. The tapered inner conductor is joined to the short segment of hollow inner conductor by means of a thin electroplated copper skin.

Dc biasing currents and voltage must be introduced to the transistor through the two coaxial paths of the jig. This is accomplished with the aid of biasing fixtures connected to the Dezifix ports. Each biasing fixture consists of a section of 50-ohm transmission line having an inner conductor interrupted with a series capacitor. Dc activating signals are fed in through a high-impedance tap connected to the segment of inner conductor adjacent to the jig. The construction of the bias fixture is suggested in Fig. 2.

Characterization measurements reported in Section 2.4 indicate a

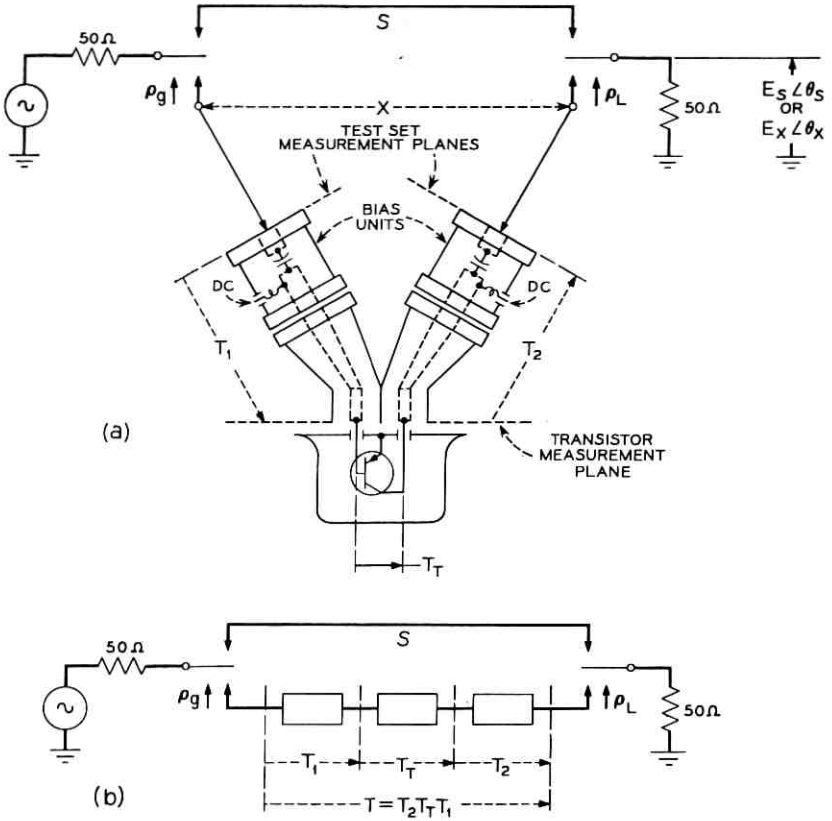


Fig. 2 — (a) Insertion loss, 20 $\log_{10} E_s/E_x$, and insertion phase, $\measuredangle \theta_s - \measuredangle \theta_x$ yield information for computation of scattering parameters S_{21} and S_{12} of transistor; (b) analytical model.

net reflection of less than 0.12 for the path between the small bore (transistor) port of the jig and the output port of the biasing fixture. This applies up to 2.5 gc.

The photo in Fig. 3 shows the jig-bias fixture assembly. Clamping plates for grounding the low-potential transistor lead are visible in the photo.

2.2 *Transistor Measurement*

It was noted previously that a minimum of two transmission measurements and two bridging measurements completely characterize the unknown being tested. For the most precise work it is actually necessary

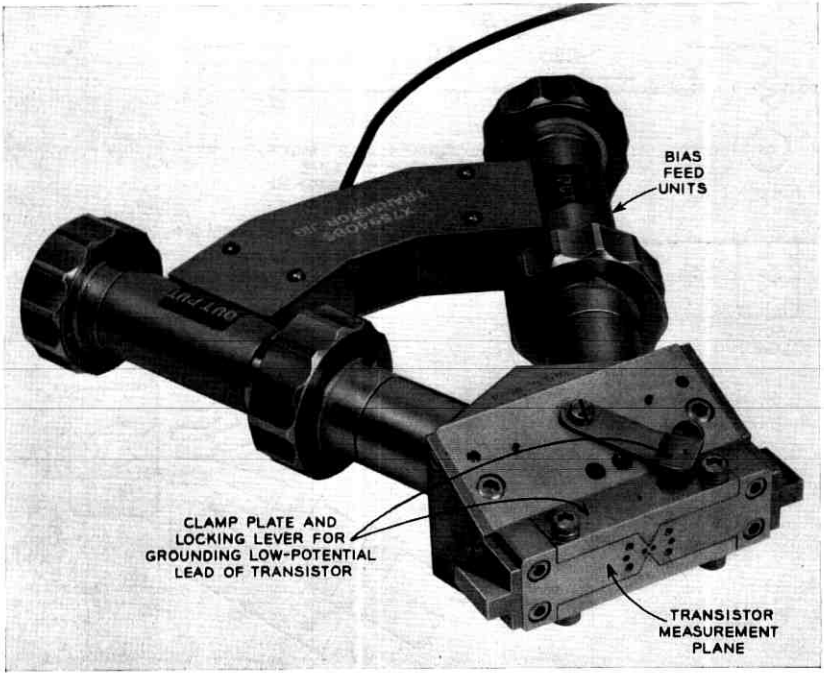


Fig. 3 — Jig and bias unit showing clamp mechanism for joining low potential lead of transistor to measurement plane.

to make additional "calibrating" measurements in order to take account of the electrical length of the jig and to compensate for the small but significant reflections presented at the coaxial test ports of the measuring set. The following procedure is required.

(*i*) The insertion loss and phase in both directions of transmission is measured, as suggested in Fig. 2. The measured losses and phases apply to the tandem connection of transistor, jig, and bias units. To obtain the transistor's loss and phase, the contributions of the jig and bias units must subsequently be subtracted out.

Next, the bridging measurements are made; the details are shown in Fig. 4. Several consecutive steps are involved, as noted below.

(*ii*) The input impedance of the transistor, as viewed through the input bias unit and jig path is connected to one of the ports of a coaxial "trombone". The other port of the trombone bridges the nominal 50-ohm transmission path in the test set. By "play-

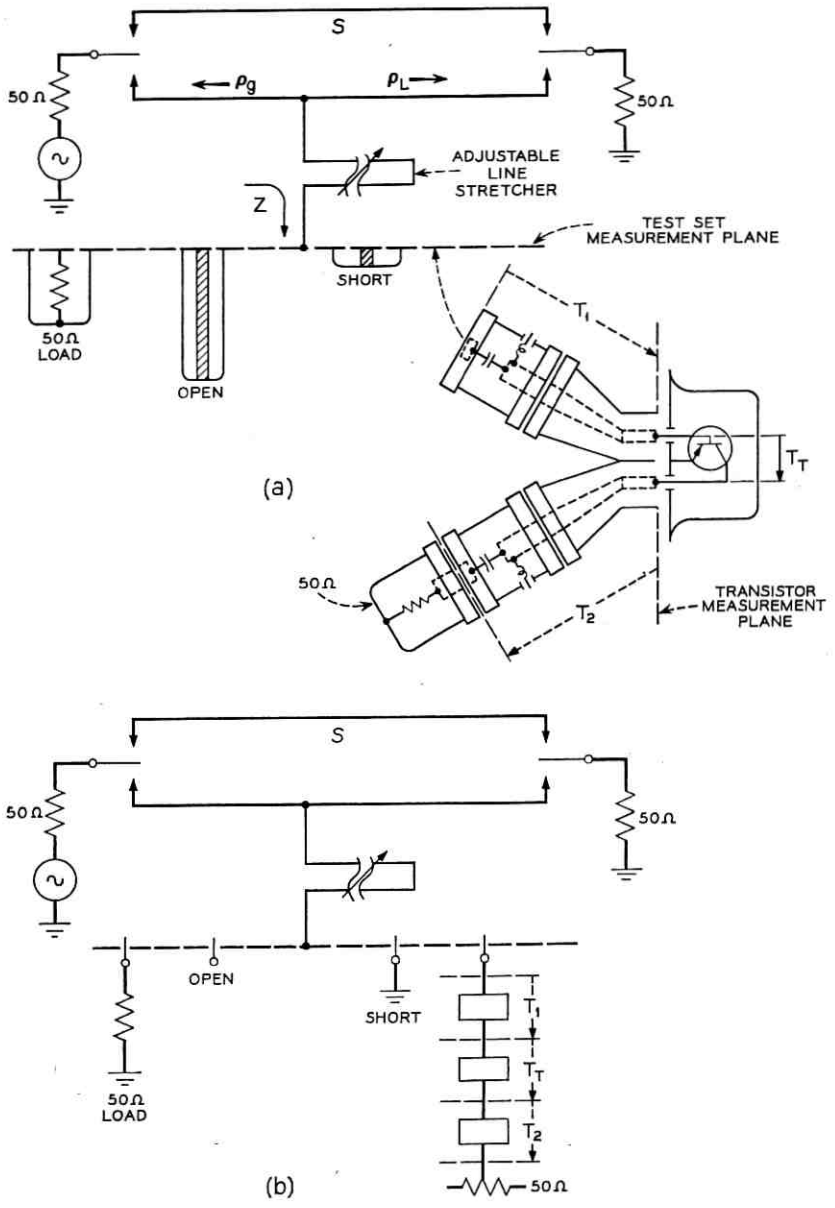


Fig. 4 — (a) Bridging loss and phase measurements to determine S_{11} and S_{22} of transistor; (b) analytical model.

ing" the trombone, a point is quickly found at which the indicated insertion loss is maximum. The loss and phase at this point are recorded; the line is now left undisturbed during the subsequent step (*iii*).

The purpose of the adjustable line is to overcome potentially adverse effects due to the transformation of the transistor input impedance by the electrical length of the input bias unit and jig path. With the physical lengths actually involved, quarter-wave inversion of small impedances to high may occur at frequencies as low as 300 mc. High impedances, if directly connected to the bridging junction in the 50-ohm transmission path, produce only small insertion losses and, as a result, accuracy suffers. The interposition of the adjustable line circumvents this difficulty since a length may always be found which reinverts high impedances back to low.

(*iii*) A known interrelation exists[1] between the impedance, Z, connected to the test set measurement plane at the accessible port of the line stretcher (Fig. 4), and the insertion ratio, $e^{-\theta}$, which this connection produces:

$$e^{-\theta} = \frac{A + BZ}{1 + CZ}. \tag{1}$$

The A, B, and C are constants which depend only on the line stretcher and the test set network behind it. At each frequency of measurement, the three constants may be evaluated by measuring loss and phase for three known values of Z. By combining the three calibrating measurements with the initial measurement on the unknown, the impedance of the unknown may be evaluated. As may be seen in Fig. 4, the calibrating measurements are made by observing the loss and phase caused by connecting in succession a coaxial short, a coaxial open, and a matched termination to the test measurement plane.

(*iv*) Steps (*ii*) and (*iii*) are repeated with the jig turned around to present the output impedance of the transistor to the test set.

To avoid inaccuracies, the standard reflections must be known with exactness, and they must be attached to the test set in precisely the plane occupied by the unknown in step (*i*). The required coincidence between all measurement planes has been assured through the use of the connector type (Dezifix B) in which electrical and physical junction planes are identical.

The 50-ohm standard exhibits a reflection smaller than 0.01 up to

4000 mc. The short-circuit standard is realized with a simple shorting plate and the open circuit is a 90° long section of shorted 50-ohm line of precisely controlled length and transverse dimensions. During the calibrating measurement with the open circuit, the test signal frequency is set to the appropriate nominal with an accuracy of at least 0.1 per cent.

2.3 *Reduction of Measurement Data to Transistor Parameters*

All of the measurements in Section 2.2 are made with respect to terminal planes remote from the transistor. Hence, in order to obtain the contribution of the transistor parameters, it is necessary to subtract out the contribution of the path through the jig and bias fixtures.

If, as indicated in Figs. 2 and 4, the circuit properties of the jig paths and the transistor are expressed individually in terms of their transmission, or "cascade wave" matrices,[5] then the over-all matrix, T, between the ports of the aggregate unknown is

$$T = T_2 \, T_T \, T_1 , \tag{2}$$

where the matrices T_2, T_T, and T_1 refer, respectively, to the output jig path, transistor, and input jig path.

T is defined from the measurements outlined in Section 2.2. T_1 and T_2 are known from the characterization work discussed in Section 2.4. Hence, T_T is completely defined in terms of known matrices, and may be deduced from (2) by the elementary inversion

$$T_T = T_2^{-1} \, T \, T_1^{-1}. \tag{3}$$

T_T may be converted to any of the other characterization sets, e.g., s, h, y, or z parameters.

The actual processing of data involves the following operations, all of which are run on a digital computer.

(*i*) Using relationships developed in previous work,[1] the scattering coefficients, s_{11} and s_{22}, of the aggregate unknown are computed from the bridging measurement data of steps (*ii*), (*iii*), and (*iv*) in Section 2.2. The impedance reference for the S matrix is 50 ohms.

(*ii*) From the transmission data for the aggregate unknown in step (*i*) of Section 2.2, it is possible to compute directly s_{21} and s_{12}.[1] The program includes a necessary correction for the residue of mistermination caused by the fact that the test set reflection coefficients, ρ_G and ρ_L, are not zero. These reflections have

been measured at a large number of frequencies in the 0.25 – 4.2-gc range, and are known.

(*iii*) The complete scattering set,

$$\begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix}$$

obtained from (*i*) and (*ii*) above is transformed to the corresponding matrix, T.

(*iv*) T is inserted into (3) together with the known characterizations T_2 and T_1 and the coefficients of the wave matrix of the transistor, T_T, are computed.

(*v*) T_T from step (*iv*) is transformed to other matrix sets.

A typical end result of this process of measurement and data reduction is illustrated in Fig. 5. The figure shows examples of two of the frequency characteristics of a UHF transistor derived from the measurement and data assimilation routines just outlined.

2.4 *Characterization of Jig and Bias Units*

Serious errors occur in the value of the deduced transistor parameters if the values taken for the matrices T_1 and T_2 do not faithfully represent the actual networks. Hence, considerable care was exercised in the measurement program for obtaining their characterizations.

Since T_1 and T_2 represent networks having very small losses, they were conveniently characterized by a variant of the "Weissfloch" technique. This procedure is based on the well-known transformer law

$$\rho_{in} = s_{11} + \frac{s_{12}^2}{1 - s_{22}\rho_L} \cdot \rho_L \tag{4}$$

relating the reflection looking into a passive two port to its scattering parameters and load side reflection. For example, to determine T_1 at one of the assigned frequencies, reflection coefficient measurements were made at the large bore port on the bias unit with known reflection standards inserted into the small bore port of the jig. Three such measurements, using three different small bore standards, are sufficient to define T_1. The sign indeterminancy in s_{12} is resolved on the basis of consistency with the electrical length as estimated from the physical dimensions of the actual model.

An example of a small bore reflection standard used in these measurements is shown in Fig. 6. The standard consists of a Teflon filled 50-ohm

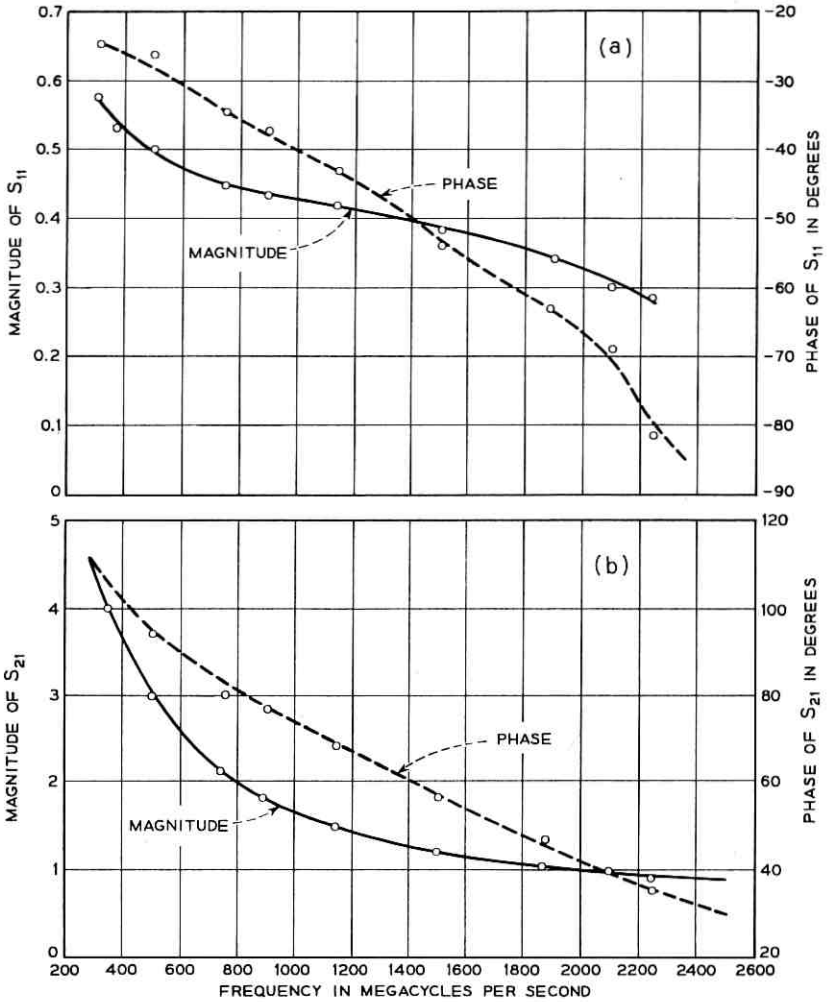


Fig. 5 — Scattering parameters for an L2254 transistor with respect to 50-ohm impedance; grounded emitter, $I_E = +5$ milliamperes, $V_{CE} = -5$ Volts; (a) magnitude and angle of S_{11}, (b) magnitude and angle of S_{21}.

coax line of very small, but also very carefully controlled interior dimensions, shorted at the far end.

The physical parameters were in sufficient control to define the angle of the reflection coefficient standards with an accuracy of about one degree up to 3000 mc. The round trip ohmic loss, and its variation with frequency, were included in the evaluation of the precise reflection magnitudes.

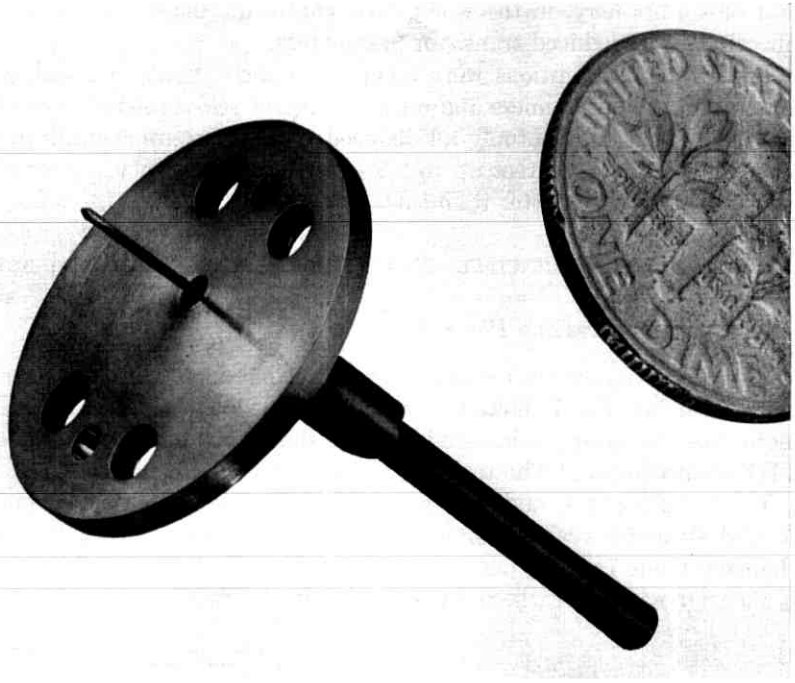


Fig. 6 — Example of one of the 0.054″ bore coaxial reflection standards used in characterization of transistor jig and bias fixture.

A very precise slotted line was used for measuring the values of ρ_{in} in (4). The line was initially checked for accuracy using the "sliding null" technique.[6]

2.5 *Accuracy of Transistor Characterization*

The two chief sources of transistor characterization error are the residual inaccuracies in the characterization of the jig and bias fixture paths and test set errors in measuring loss and phase. The error contributions from both of these sources depend heavily on the characteristics of the transistor, hence a categorical "error statement", in the usual sense, is not possible. Nonetheless, studies have been made which show that loss and phase measurement errors of 0.1 db and 0.6 degree (the maximum expected) may, in the worst conceivable case, produce errors of 0.8 db or 4 degrees in the determination of the s parameters of a transistor having the characteristics displayed in Fig. 5. The error from the estimated defect in the characterization of the jigs

and bias units may, in the worst case, contribute 0.5 db and 3 degree uncertainty in deduced transistor parameters.

Pessimistic assumptions were involved in these worst case analyses. Based on data smoothness and on a number of self-consistency checks, it appears that substantially all deduced parameters are accurate to at least 0.5 db and 3 degrees up to 2.5 gc, with the majority of measurements considerably better than this.

III. MEASUREMENT PRINCIPLES AND PERFORMANCE OF THE TEST SET

3.1 *Principle of Loss and Phase Measurement*

The basic quantities measured are insertion loss, phase shift, and envelope delay. Fig. 7 shows the principle of the loss and phase measurement; the operation for loss and phase is similar to that of the previous VHF measuring set.[1] The technique is that of "IF substitution".

Vibrating relays s_1 and s_2 sequentially interpose the unknown and a coaxial strap between a nominal 50-ohm source and load. The interchange is made $7\frac{1}{2}$ times per second. The signal (E_s or E_x) emerging from s_2 passes to a receiver where it is shifted down in frequency to a constant

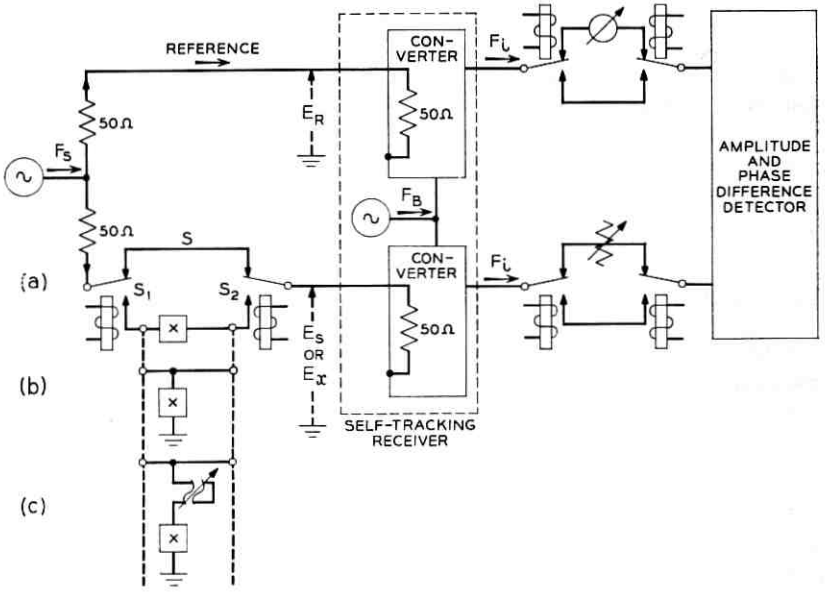


Fig. 7 — (a) Basic insertion loss and phase measuring arrangement embodying rapid comparison and frequency translation to fixed IF, (b) impedance measurement by bridging, (c) inversion of high impedances by adjustable transformer.

IF. A second pair of relays inserted in tandem with the output of the receiver switches an adjustable loss standard in synchronism with s_1 and s_2. The operator adjusts the loss standard so that the detector, which is capable of recognizing when the two sequentially applied inputs are of equal amplitude, delivers a null indication to a loss display meter. At the null point, the attenuator loss equals that of the unknown since the receiver is highly linear.

The difference of phase angle between E_s and E_x is the insertion phase of the unknown. This is measured in a manner exactly analogous to loss, using, this time, an adjustable phase shifter as the null-balanced standard.

3.2 *Principle of Delay Measurement*

Delay is measured by observing the phase shift experienced by a modulation envelope in its transit through the unknown.[7] The basic arrangement, which is illustrated in Fig. 8, employs a relatively simple AM modulator to generate the delay test signal. It is not necessary for the modulation envelope to be low in harmonic content. As in the previous case of loss and phase measurement, a pair of RF comparison switches sequentially completes measurement paths through the unknown and the standard at a $7\frac{1}{2}$-cps rate. The envelope delay data is contained in the phase difference between the fundamental components of the modulation envelopes borne by the signals E_s and E_x appearing sequentially at the input to the self-tracking receiver.

By conversion in the receiver, the modulation envelopes are super-

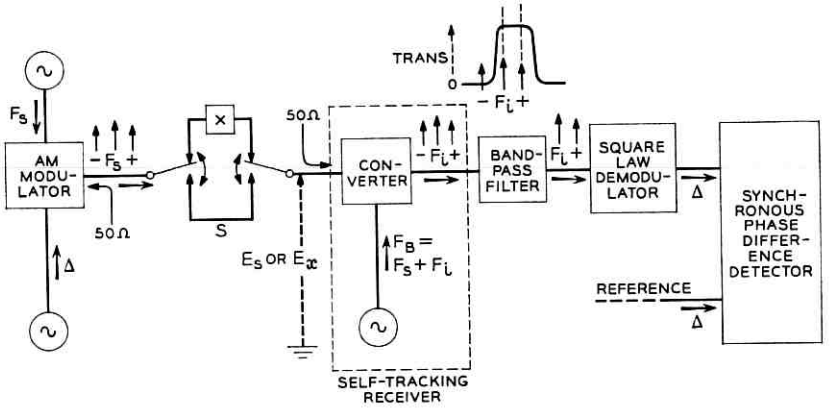


Fig. 8 — Principle of delay measurement showing reduction of three-tone envelope spectrum to twin-tone before demodulation.

imposed on the intermediate frequency, F_i, which is held rigidly constant by automatic frequency control of the beating oscillator. The tight control of the IF makes it possible to depart, at this point, from the usually employed arrangements which pass the IF output from the receiver directly to an envelope demodulator. Instead, advantage is taken of the constancy of the IF to eliminate all but two tones of the modulation spectrum, preliminary to detection. A steep-sided bandpass filter, interposed between the receiver and the demodulator, strips away all tones except the IF carrier and the upper modulation sideband. Since only two input tones are applied to the demodulator, the phase of the beat frequency product is totally independent of the amplitudes of the beating signals. In the conventional arrangement in which a stripping filter is absent, the phase of the baseband signal depends on the relative amplitudes and phases of all of the tones present in the RF spectrum. Hence, significant delay errors may result when the unknown exhibits loss and phase distortion across the frequency interval spanned by the array of side tones.

Of almost equal importance is the reduction in the size of delay measuring "aperture" accompanying the use of the stripping filter. Reducing the width of the modulation spectrum by a factor of two before demodulation serves to improve the ability to resolve finer-grained envelope delay detail by about the same factor.

The delay information is contained in the phase difference, $\theta_s - \theta_x$, between the sequentially appearing signals out of the demodulator. This phase difference is detected in the manner previously described in Section 3.1.

The unknown's envelope delay, τ, is closely given by

$$\tau = \frac{\theta_s - \theta_x}{\Delta} \text{ seconds,}$$

where the angle difference in the numerator and the modulation envelope rate in the denominator are both expressed in radians. By following through the frequency transpositions in Fig. 8, it is clear that the frequency at which τ applies may reasonably be taken to be the mean between the RF carrier and the adjacent lower side band. Δ is equal to (2π) $(5.55)10^6$ radians per second in the present set, hence, an increment of 0.5 nanosecond in the unknown's delay gives rise to one degree of envelope phase shift.

3.3 *Measurement Accuracy and Ranges*

Signal Source: 0.25 to 4.2 gc in four bands; ± 1 per cent scale calibration accuracy. The sources may be set to specific frequencies with a

tolerance of about 5 parts in 10^4 using high accuracy commercial counters for frequency measurement. Band changes are made automatically.

Insertion Loss Range: 50-db loss to 10-db gain. (Gains exceeding 10 db may be measured by preceding the unknown with a loss pad.) Accuracy: ±0.1 db from 10-db gain to 40-db loss; ±0.3 db from 40-db to 50-db loss. The accuracies just cited apply up to 2.0 gc. Above 2.0 gc the errors may increase by a factor of two.

Insertion Phase Range: 360 degrees. Accuracy: ±0.6 degree between 10-db gain and 40-db loss up to 2.0 gc; ±1.0 degree between 40-db and 50-db loss. Above 2.0 gc the errors may increase by a factor of two.

Delay Measurement (Limited to measurement on linear networks between 10-db gain and 20-db loss) Range: 180 nanoseconds. Accuracy: ±0.5 nanosecond up to 4 gc. Aperture width: 5.55 mc.

Source and Load Terminations for Coaxial Unknowns: Nominal 50 ohms; reflection coefficient magnitudes of source and load decay from 0.01 at 1.0 gc to 0.1 at 4.2 gc. Errors due to the residue of misermination have been treated in earlier work,[1,11] and are summarized in Section 6.1.

Source Power Applied to Unknown: Excitation level is automatically varied in accordance with unknown's loss so as to keep the input power at the lowest possible level consistent with satisfying signal-to-noise ratio requirements at the loss and phase detectors in the test set.

Down to losses of 19.9 db, the available source power is kept below −40 dbm. The power increases to −30 dbm for losses between 20 db and 29.9 db, to −20 dbm between 30 db and 39.9 db and to −10 dbm between 40 db and 50 db.

IV. OVER-ALL CIRCUIT DESIGN

4.1 *Objectives*

Experience has amply established that the development of components and devices for new communication systems entails a large volume of measurement. In view of this, it would be most unwise to achieve a high degree of measurement accuracy at the expense of awkward and laborious measurement procedures. Hence, in the present case, the aim has been to design for a facile interface between operator and machine without undue sacrifice of measurement accuracy.

A number of features were introduced to minimize the amount of operator labor required to make measurements. Automatic control circuitry has been liberally employed to tune oscillators, to set levels

and operating points of detectors, and to stabilize and render insensitive to test signal frequency the deflection sensitivity on the display meters. A considerable circuit complexity results from this high degree of automatic operation, but this is not obvious to the operator who benefits from the simplified measuring procedure. Where feasible, machine logic and mechanism have been substituted for operator decision and motor activity.

One of the most vital objectives was to provide a measuring facility with the capability to comprehensively characterize either passive or active linear two-ports. This was accomplished by the inclusion of bridging loss in addition to the insertion loss measurement features.

Transistors were to be characterized under light excitation conditions, in order to permit study of small signal parameter variation with shift of dc operating point. In the majority of measurements, less than 0.1 microwatt of power is absorbed by the transistor.

Measurement range, accuracy, and resolution were to be essentially independent of test signal frequency.

Solid-state design was to be used in all possible instances.

4.2 *Over-all Circuit Operation — Loss Measurement*

To achieve accuracy substantially independent of test signal frequency over the 4 octave range between 0.25 and 4.2 gc, it is necessary to heterodyne the measurement information to a fixed intermediate frequency where detection may be performed with the aid of precisely calibrated loss and phase shift standards. In addition, to minimize "instrument zero-line" and to eliminate error from drifts of phase shift and level within the set, it is desirable to sequentially compare the unknown with a high-frequency reference.[1] The rapid comparison and heterodyne design aspects of the test set were noted in the previous discussion of Figs. 7 and 8. A more complete development of these features is now presented in Fig. 9.

Referring to Fig. 9, loss may be measured with the instrument set up either in the basic phase or delay modes. Assuming selection of the phase mode, the AM modulator preceding the vibrating RF comparison switches is heavily dc biased for small transmission loss. Mode switches S_3, S_4, S_5, S_6, and S_7, program the instrument for phase or delay measurements. These switches are shown in the operating position for phase measurements.

A self-tracking double-conversion receiver translates the measurement information to a first IF of 60 mc, followed by a second conversion to 1.11 mc. Insertion loss is then detected with the aid of a differential

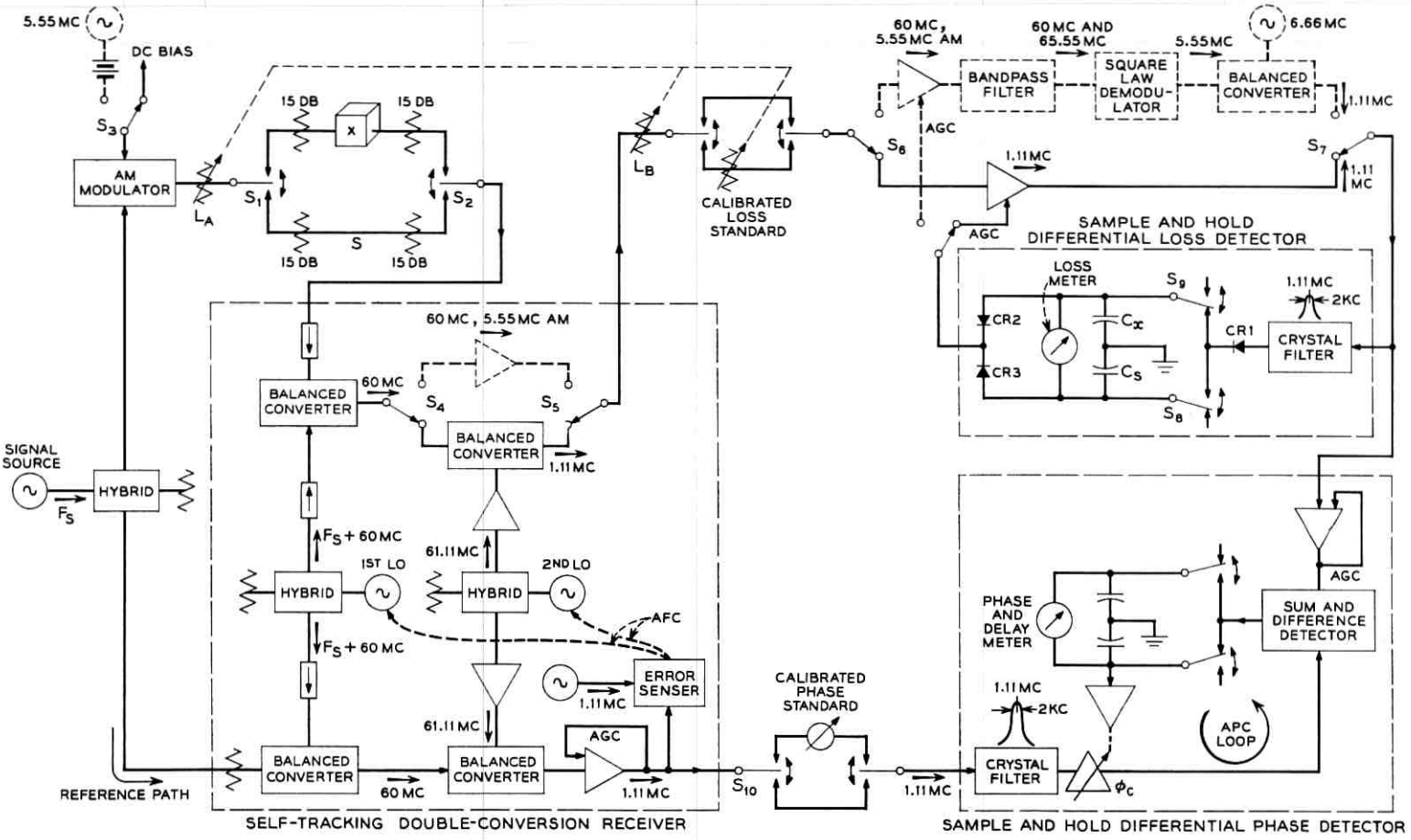


Fig. 9 — Block diagram of test set showing rapid comparison, IF substitution technique of measurement.

"sample and hold" detector involving the rectifier CR_1, the vibrating switches S_8 and S_9 and the storage capacitors C_x and C_s. The relays within the detector and those around the calibrated loss standard vibrate in phase synchronism with the $7\frac{1}{2}$-cycle rate of the RF comparison relays, S_1 and S_2. A null on the loss meter indicates when the attenuation standard, at 1.11 mc, has been adjusted for loss equality with the unknown. The measuring procedure is that of "IF substitution" and yields answers which are valid, for small losses, within the degree of linearity of the tracking receiver and for large losses, within limitations of noise performance.

Approximately equal detection sensitivity for various values of unknown loss is attained automatically by ganging common path attenuation, L_A and L_B, to the loss standard in such a way that the level of signals applied to the loss detector, at the null balance point, is maintained approximately constant.

In measuring at the higher losses, a greater portion of the common-path attenuation is assigned to the attenuator section located at intermediate frequency. Although this does have the somewhat undesirable effect of increasing the signal drive on transistors when measuring at high loss, it prevents the input level to the first converter from dropping dangerously close to noise. The way in which the pattern of attenuation is worked out assures sufficient signal for an S/N ratio of at least 25 db at the detection point during 50-db loss measurements. Error due to noise, under these conditions, is less than 0.05 db.[8]

The operation of the attenuators L_A and L_B does not completely insure constant level operation at the loss detector because of the residual frequency characteristic of the converters and the natural droop in loss of the test set cables with increasing frequency. These effects are prevented from altering the detection sensitivity by the presence of an over-all AGC circuit preceding the input to the loss detector.

A significant feature of this AGC circuit is the use of the steering diodes CR_2 and CR_3 across the storage capacitors for the purpose of rendering the AGC responsive only to the larger of the two stored signals. The resultant AGC action maintains the larger signal stiffly at some standardized level. This arrangement provides both for a constant deflection factor at the loss meter and a symmetrical calibration.

When making gain measurements, the loss standard is inserted in series with the unknown, by transposing it from the S to the X Channel. Necessary changes in common attenuation ganging are made at the same time.

4.3 *Over-all Circuit Operation — Phase Measurement*

Phase measurement, like the loss measurement, is performed by the technique of IF substitution. The essential IF measurement apparatus, as seen in Fig. 9, consists of the calibrated phase standard, the switches to introduce it sequentially in synchronism with the other comparison relays of the test set, and a "sum and difference" phase discriminator feeding the sample and hold circuitry. The capacitors in the sample and hold circuit store the output voltages obtained from the discriminator during the successive dwell periods of the comparison relays. When the two stored voltages are equal, as indicated by a null on the phase meter, the phase shift through the calibrated standard equals the insertion phase of the device under test, since the phase of the common path circuitry makes no contribution whatsoever to the phase meter indication.

The null balancing procedure using the phase standard assures only that the phase difference between the two inputs to the phase detector is the same for both dwell states of the comparison relays. Unless some special circuit provision is made, the absolute phase difference between the detector inputs could, at the null point, range over wide limits because of frequency sensitive asymmetry between the two transmission paths energized from the test signal source. Moreover, the two halves of the receiver, though nominally similar, fail to track perfectly with respect to conversion phase. Hence, in order to fix the operating point and make the provision of a calibrated meter possible, an APC loop was necessary to regulate the value of the common path phase shifter, φ_c, (see Fig. 9) so as to maintain the inputs to the detector in a state of phase quadrature during the period of S path closure. The ninety degree phase difference represents center of range on the transducer characteristic of a sum and difference type of detector.

In addition to operating point control, which is an absolute necessity for the viability of the scale calibration, it is most desirable to maintain constant scale factor. To achieve this, the signal amplitudes applied to the phase detector must be held at fixed values, independent of either the test signal frequency or the dwell state of the comparison relays. This poses no particular problem in the reference path since the operation of the comparison relays around the phase standard does not result in level shifts. An elementary AGC in the receiver is sufficient to remove any reference level variation traceable to frequency characteristic of converters or frequency dependence of signal source amplitudes.

On the other hand, the other input to the phase detector may fluctuate in amplitude at the 7½-cycle comparison rate, depending upon the loss standard setting. It is clear that the AGC circuit introduced in the phase detector to wipe out this cyclic level difference must not introduce spurious phase changes in the course of leveling. By the use of a system of gain control, which is relatively free of reactance shifts with control voltage, the level to phase conversion has been held to 0.1 degree for 10-db input level changes.

4.4 *Over-all Circuit Operation — Delay Measurement*

The elements in dashed outline in Fig. 9 are automatically switched in when the operator selects the delay mode of measurement. The operation closely follows the simplified description given previously in Section 3.2.

A balanced converter, excited from a 6.66-mc local oscillator source, is introduced beyond the square-law demodulator to translate the 5.55-mc envelope signal to the 1.11-mc IF. There is no question of frequency incoherence between the IF signal emanating from this converter and the signal of the same *nominal* frequency emitted at the reference path output of the receiver (Fig. 9), since the 6.66-mc local oscillator tone is formed by modulating the 1.11-mc reference in the receiver with the 5.55 mc producing the AM modulation envelope.

Delay detection is performed with the same measurement apparatus used for phase.

During delay measurement the loss standard is shifted to the 60-mc IF. This prevents any delay error from residual level-to-phase conversion in the square-law detector, since the initial operation of loss balancing equalizes the levels at the input to the detector during the two dwell states of the comparison relays. This would, of course, be only an illusory advantage if the delay through the loss standard varied with the loss setting.

In the course of delay measurements, the over-all AGC for control of loss meter scale factor is transferred to a 60-mc amplifier preceding the square-law demodulator.

4.5 *Principal Features of the Measurement System*

The measurement circuit combines features of rapid comparison and null balancing of standards with heterodyne detection. The advantages inherent in such measurement arrangements have been noted previously.[1]

Briefly, the comparison of s with x is so rapid that the measurement results are unaffected by slow wanders of source level, drifts of conversion gain or phase in the receiver, or by drifts of operating point and sensitivity in the loss and phase detector. This arrangement also obviates error from residual dependence of source level or receiver characteristics on frequency, since the source, receiver, and detector are common to the channels being compared. To prevent errors in the comparison of the unknown with the standard, the transmission paths through the comparison switches must be well matched to the nominal termination level, and the crosstalk from the open to the transmitting path must be small.

It is most desirable that the two paths through the switches transmit equally. This is not mandatory from an accuracy viewpoint, since, by paying the penalty of added labor and inconvenience, an initial "zero run" may be taken to acquire the asymmetry data on the comparison switches. Small "zero-line" residues prove to be inescapable.

The heterodyne technique has several conspicuous advantages. First of all, locating the loss and phase standards at the intermediate frequency avoids the problem of developing broadband standards.

Thermal noise power is reduced before detection by 2-kc wide crystal filters, thereby reducing measurement error due to system noise.

Loss of the device under test may be made up by single frequency gain at the IF.

And finally, the heterodyne system introduces the advantages of selective detection. It provides immunity from errors due to harmonics and stray signals in the output of the unknown.

While the attributes just listed are basic, certain other design features are quite important from the standpoint of conserving operator effort and expediting measurement. In particular, the self-tracking property of the receiver not only relieves the operator of the burden of adjusting local oscillators, but also permits the use of a narrow (2-kc wide) final IF by virtue of the tuning precision inherent in the automatic control.

The provision of calibrated loss, phase and delay scales in combination with slow scan of the signal frequency is useful in network adjustments, e.g., tuning for maximum or minimum responses or estimating limits of response variation over given frequency bands.

V. SUBSYSTEMS OF THE MEASUREMENT SET

The "front end" apparatus in Fig. 9 must span the 0.25- to 4.2-gc range of measurement frequency. This includes the signal and first local oscillator sources and such transmission components as the microwave

converters, hybrids, isolators, and AM modulators. None of this apparatus can be designed to cover the full frequency range without considerable sacrifice in important performance attributes. Hence, the test signal range has been subdivided into 4 bands of octave width, with separate front end apparatus provided for by each band. This represents no complication from the standpoint of the operator, since a system of remotely operated coaxial relays automatically switches the front end apparatus when passing from one band to another.

The test set circuits beyond the output of the first converters (Fig. 9) are common to all four bands.

Four signal sources cover the 0.25- to 4.2-gc range. Butterfly-type oscillators are used below 1.0 gc; klystron sources are used between 1.0 and 4.2 gc. Source oscillators with sliding contact mechanisms for changing frequency were avoided in order to guard against the occurrence of transient level or frequency "hits" on the AFC circuitry in the receiver.

The RF and IF comparison switching, sampling, and storage in this set, while unusually extensive in terms of numbers of relays required, is quite conventional and follows the design principles elucidated in previous work.[1] With the exception of the two comparison relays operating at the test signal frequency, all of the $7\frac{1}{2}$-cycle relays use mercury wetted contacts in order to realize the advantages of low contact resistance, transmission symmetry in the two dwell states and long life. Many of the relays are compound, i.e., complexes of several relays are formed for the purpose of achieving low crosstalk by shorting the nominally open transmission path. Both coaxial and noncoaxial contact designs are employed according to need.

The RF comparison at the $7\frac{1}{2}$-cycle rate is performed with a pair of solenoid-operated coaxial relays. These relays have dry contacts, but the contact pressure is very great and a wiping action occurs during the "make" phase. Leakage through the nominally "open" path is 44 db down on transmission through the closed path at 4.2 gc. Since switching is done at both input and output, error from this cause is less than 0.1 db at 40-db loss level.

With the exception of the signal and first local oscillators, all of the circuitry is solid state.

5.1 *Remote Tuning of Signal Sources*

Good practice in high-loss measurements calls for tight shielding of the signal sources in order to prevent any leakage of source power to susceptible low-level points in the test circuit. Hence, in the present

instance, all four test signal sources are housed in an RF tight cabinet together with associated leveling, control, and coaxial relay band-change circuitry.

The sources are remotely tuned by the use of the tachometer stabilized 60-cycle servo loop shown in Fig. 10. The tach feedback is adjusted to dynamically damp the loop rather heavily in order to limit the maximum rate of test signal frequency change to 100 mc per second. This represents the maximum rate of change of source frequency which the AFC circuitry in the receiver was designed to follow. The dead zone

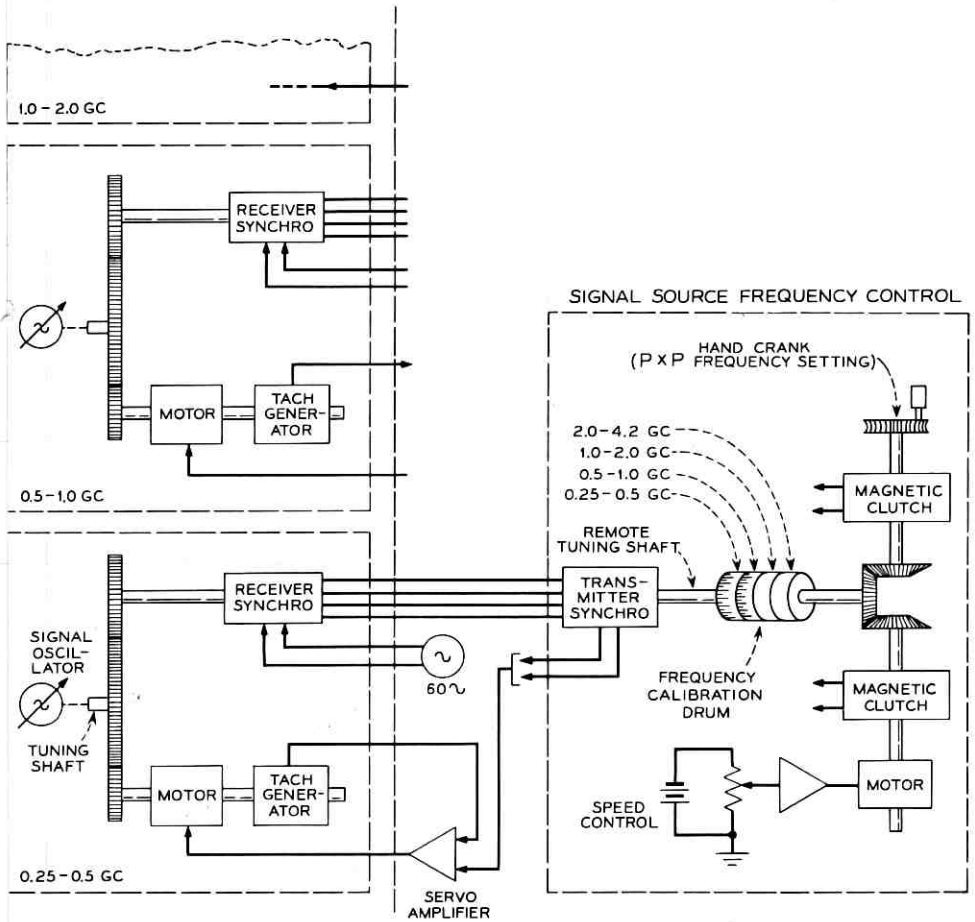


Fig. 10 — Remote tuning of test signal sources.

around the servo null is sufficiently narrow to permit remote frequency settability to about 5 parts in 10^4.

The transmit synchro is automatically switched to the appropriate receiver synchro when changing between bands. The servo amplifier is also shared among all four bands.

Either motor scan or point-by-point tuning is possible. As seen from Fig. 10, the mechanism for accomplishing this consists of an arrangement of electric clutches and differential to permit the operator to turn the frequency calibration drum dial either by hand or by an adjustable speed motor.

5.2 *Self-Tracking Receiver*

The receiver must faithfully transpose measurement data to 1.11 mc in the case of phase measurement and to 60 mc in the case of delay measurement. In Fig. 11, showing a block diagram of the receiver, the following aspects of design must be considered.

(*i*) The linear dynamic range, i.e., the range for applied signals bounded by receiver compression at one end and receiver noise at the other.

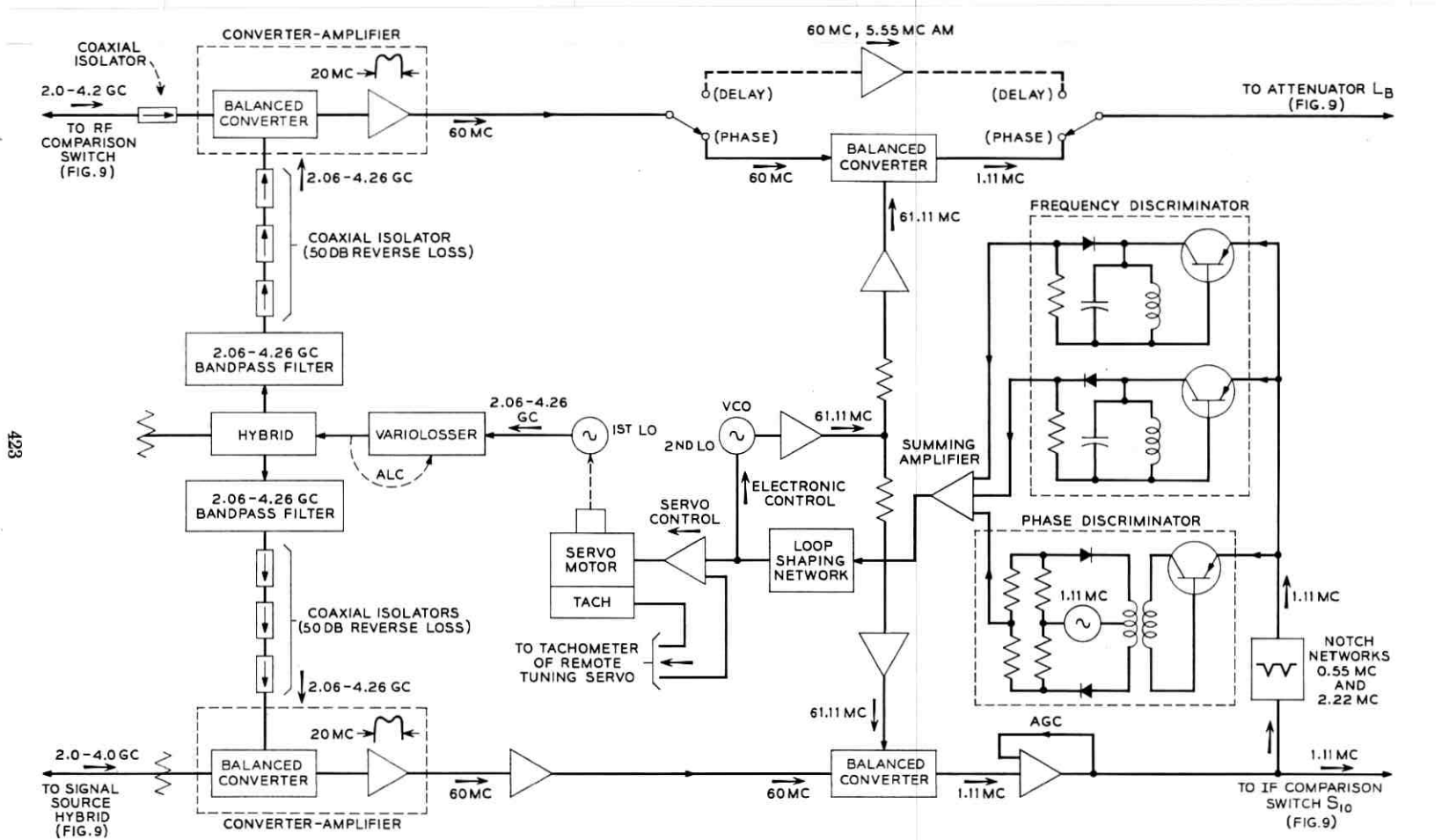
(*ii*) Suppression of crosstalk and pick up between the two similar halves of the receiver.

(*iii*) Execution of an AFC circuit capable of holding the second IF centered within the nominal 2-kc receiver bandwidth in the presence of signal frequency scan up to rates of 100 mc/second.

5.21 *Conversion to First IF*

To be compatible with accuracy objectives, the first converter must have a linear dynamic range of at least 40 db over which errors due either to compression or to noise must be smaller than 0.03 db. It was possible to meet this requirement using coaxial, balanced mixers covering octave bandwidths. The particular mixers employed have about 8-db conversion loss and 10-db noise figure. An integrally mounted 60-mc preamplifier of 20-mc bandwidth provides 25 db of gain.

Referring now to Fig. 11, potentially harmful leakage paths may be identified starting from the test signal frequency input to the first converter in one path and terminating in the local oscillator port of the converter in the other path. Test signal which is transmitted over these leakage paths beats with the local oscillator tones to produce traces of spurious 60 mc. In order to avoid the generation of such signals, which may produce errors when measuring at the highest loss levels, the


CONVERTER-AMPLIFIER
COAXIAL ISOLATOR
20 MC
2.0-4.2 GC
BALANCED CONVERTER
TO RF COMPARISON SWITCH (FIG. 9)
60 MC
60 MC, 5.55 MC AM
(DELAY)
(DELAY)
(PHASE)
(PHASE)
TO ATTENUATOR L_B (FIG. 9)
BALANCED CONVERTER
60 MC
1.11 MC
2.06-4.26 GC
COAXIAL ISOLATOR (50 DB REVERSE LOSS)
61.11 MC
FREQUENCY DISCRIMINATOR
2.06-4.26 GC BANDPASS FILTER
HYBRID
VARIOLOSSER
2.06-4.26 GC
IST LO
VCO
2ND LO
61.11 MC
SUMMING AMPLIFIER
ALC
ELECTRONIC CONTROL
SERVO CONTROL
PHASE DISCRIMINATOR
SERVO MOTOR
TACH
LOOP SHAPING NETWORK
2.06-4.26 GC BANDPASS FILTER
COAXIAL ISOLATORS (50 DB REVERSE LOSS)
1.11 MC
TO TACHOMETER OF REMOTE TUNING SERVO
NOTCH NETWORKS 0.55 MC AND 2.22 MC
2.06-4.26 GC
61.11 MC
AGC
20 MC
2.0-4.0 GC
BALANCED CONVERTER
BALANCED CONVERTER
1.11 MC
TO SIGNAL SOURCE HYBRID (FIG. 9)
60 MC
60 MC
TO IF COMPARISON SWITCH S_{10} (FIG. 9)
CONVERTER-AMPLIFIER

coupling between signal frequency and local oscillator terminals of the converters should be small, the transmission paths from the LO source to the converters must exhibit a high ratio of forward to reverse transmission and the conversion efficiency for local and signal oscillator tone applied to a common input of either converter must be low.

All of these factors have been controlled such that errors from signal frequency crosstalk between the two halves of the receiver are held to less than 0.2 db even when measuring at 50-db loss levels.

The isolators in Fig. 11 present high reverse loss over only the octave for which they are optimized. Consequently, in the case of the two bands above 1.0 gc, it is necessary to introduce bandpass filters in the local oscillator paths for the purpose of suppressing interpath transmission of higher order modulation products emitted at the local oscillator ports of the converters.

The unilateralization of the LO paths for the two bands below 1.0 gc is provided by transistorized buffer amplifiers having a 40-db ratio of reverse loss to forward gain. Each amplifier design covers an octave. The high reverse loss is obtained by cascading two grounded-emitter stages. Measurements show that this is the preferred connection to obtain maximum reverse loss in the 0.25- to 1.0-gc region.

5.22 *Conversion To Second IF*

Since the second converter is fed directly from the output of the first, it must meet similar requirements on dynamic range. These requirements, which are noted in Section 5.21, are met without difficulty by the use of ring modulators. Unwanted transmission of 60-mc power between the two converters is prevented by transistor buffer amplifiers inserted in the second local oscillator transmission paths. Direct path loss between the LO ports of the second converters is greater than 100 db for 60-mc signals and sufficient at other frequencies to block the transmission of any disturbing tones created by higher-order modulation.

5.23 *Automatic Frequency Control of Local Oscillators*

The heterodyne technique of phase and delay measurement calls for a first IF to serve as a subcarrier for the AM modulation envelope and a second IF of considerably lower frequency for operation of the null balance standards. In view of the decision to narrow the receiver bandwidth to 2 kc, it would be quite impractical to consider manual tuning of the local oscillators. Moreover, such a course would have been in conflict with the objective of simplifying measurement procedure.

Tight control of the IF is advantageous for several reasons. It permits narrow IF bandwidths, thereby enhancing S/N before detection. The absence of IF flutter permits a great reduction in the tolerance imposed on phase tracking between opposing tuned amplifiers in the 60-mc and 1-mc channels of the set. And lastly, it eliminates any measurement errors from the residual frequency sensitivity of the phase standard calibration.

To achieve the desired precision in control of the intermediate frequencies, the system of frequency and phase discriminators shown in Fig. 11 senses the error of the second IF and delivers a corrective voltage that actuates two modes of control. The first of these is an electromechanical servo which achieves a coarse correction by motor-tuning the frequency of the first LO for minimum IF error. This is supplemented by an all electronic frequency control acting around the second local oscillator. With a hold in range of ± 10 mc and a very crisp response rate, the electronic loop eliminates the residue of static and dynamic errors left by the operation of the mechanical loop.

Butterfly oscillators serve as first LO sources up to 1.0 gc. Klystron oscillators are used between 1.0 and 4.2 gc. The second LO is a voltage-controlled oscillator (VCO) of special design.

The combined use of frequency and phase sensitive transducers endows the control circuit with certain useful attributes which are not present when one or the other of the transducers is used alone.[10] Frequency discriminator control, for example, tolerates static frequency error but polarizes the LO's with respect to sideband of operation. Control by sum and difference phase detector eliminates static error but permits stable, closed loop operation with the LO's in either an upper or lower sidetone relation to the signal frequencies. The combination of the two retains the zero-error property of the phase control and the polarizing property of the frequency control. It is extremely important to establish particular LO sideband sense, since the sign of the measured insertion phase shift depends on the sideband polarity.

It has also been demonstrated that the introduction of frequency discriminator control extends the pull-in range of the loop.

Another feature of interest is the use of rate feedback from the remote tuning servos around the RF test signal sources shown in Fig. 10. This is instrumented by inserting the output of the tachometer generator from the remote tuning loop in series with the tachometer attached to the first LO servo motor. Since the tachometer polarities are connected in series opposition, the control effect is to induce the first LO motor to maintain a rate correspondence with the speed of the motor which tunes

the signal oscillator. This is beneficial in reducing the dynamic stress on the AFC circuit during periods of signal frequency scan.

The dynamical attributes of the servo are essentially those of a conventional, second-order regulator circuit employing tachometer stabilization.[9]

The design of the electronic control loop may be investigated in some greater detail with the aid of Fig. 12, which presents a simplified analytic picture of the operation. For purpose of first-order analysis, a simple time delay factor, T, accounts for the loop phase contributed by the "Q" of resonant circuits in the discriminator. In the actual circuit, this approximation is valid for loop frequencies up to about 50 kc. The collapse of the approximation above this frequency is not really significant, since gain crossover occurs at approximately 40 kc.

The asymptote structures in Fig. 12 show the dominant elements of the design. The essential idea, of course, is that the loop gain must be brought to zero before the parasitic corners in the transducer characteristics cause oscillation. At the lower loop frequencies, the magnitude

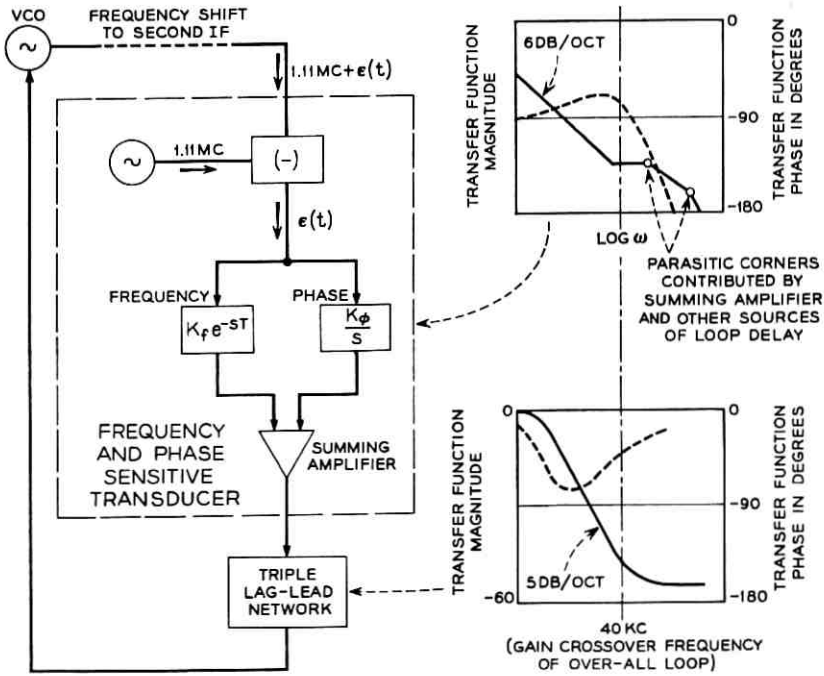


Fig. 12 — Simplified analytic model of AFC circuit.

asymptote of the aggregate frequency and phase sensitive transducer crudely approximates that of a "leading" corner. An upturn of the phase curve in the vicinity of the break may be advantageously used to obtain help on phase margin at gain cross-over. Parasitic corners above the low-frequency break, combined with the assumption of constant time delay in the frequency discriminator response, cause a fairly rapid crossover of the phase.

The loop shaping is done with a cascade of three lag-lead networks whose characteristic combines with that of the error transducer to produce an 11 db per octave over-all gain slope at crossover.

The realized loop gain and phase were measured by the procedure of applying a sinusoidal stress and observing the error residue. A procedure of this sort is necessary since a phase sensitive loop may not be opened without destroying its operation. The results, which are plotted in Fig. 13, show satisfactory agreement with the computed characteristics.

It is of interest to observe that for a given value of the discriminator

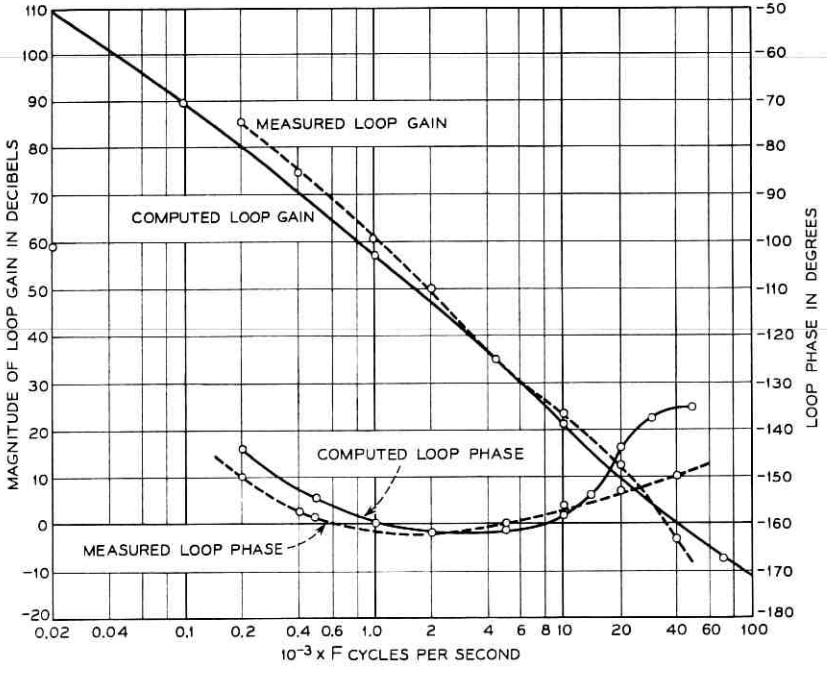


Fig. 13 — Gain and phase characteristic of AFC loop; measured and computed.

delay time, T, increasing the magnitude of the transducer gain, K_f, increases the phase up-lift in the vicinity of the low-frequency corner in the over-all characteristic of the two discriminators. Also, small values of T permit the design of wider band loops. T and K_f are not independently disposable in the actual circuit, since both are related to the network Q's.

The actual transducer parameters used in the loop are:

$$\begin{aligned} K_f &= 0.35\ (10)^{-5} \text{ volt sec/radian} \\ T &= 2.9\ (10)^{-6} \text{ sec} \\ K\varphi &= 4/\pi \text{ volts/radian} \\ K_0 &= 10^7\pi \text{ rad/volt sec} \quad \text{(VCO sensitivity).} \end{aligned}$$

The VCO is a varactor-tuned oscillator, shown schematically in Fig. 14. The operation is based on negative resistance at the base of the oscillating transistor, Q_1. The impedance presented at the base, which is approximately the product of h_{fe} and the emitter circuit impedance, contains a negative resistance term due to the capacitative emitter load and the phase angle of h_{fe} in the vicinity of 60 mc. Oscillation is supported at a frequency determined partly by the capacitance in the emitter circuit but principally by L_1 in shunt with the capacitance provided by the back-biased varactors. The combination $R_1 - C_3$ limits the oscillation amplitude through self-bias. A predistortion network linearizes the frequency deviation characteristic. This is desirable from the standpoint of maintaining the incremental loop gain of the AFC circuit independent of operating point. Q_2 provides amplification and load buffering for the oscillator.

An automatically executed program for resynching the AFC loop is initiated on test set turn-on or when passing from one test signal band to another.

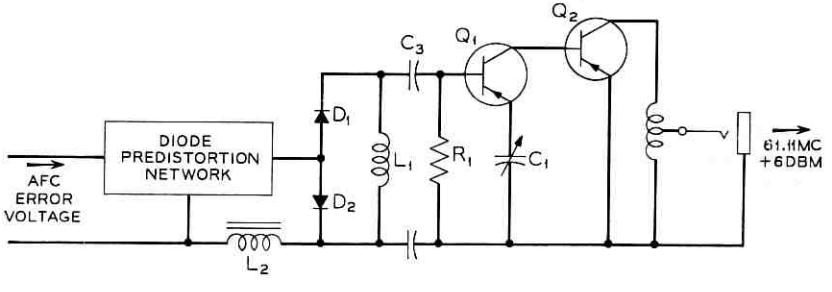


Fig. 14 — 5 mc/volt VCO with 61.11-mc center frequency. Predistortion network linearizes VCO control characteristic.

5.3 *Phase Standard*

The design targets for the phase standard call for a range of 360 degrees, a calibration accuracy of 0.1 degree maintained for long periods, freedom from warmup drift or calibration shift with ambient temperature and an output level independent of the phase shifter setting.

In view of the temperature dependence of critical transistor parameters, it was found desirable to make use of the rapid switching technique shown in Fig. 15. Transmission is alternated between a fixed phase path and the variable phase path at the $7\frac{1}{2}$-cycle repetition rate used universally in the test set. Since the switching rate is sufficiently rapid, drift of the common path amplifier characteristics is eliminated as a source of calibration error. In order to realize the maximum cancellation of drift, the input and output impedances of the fixed phase networks have been designed to approximate those of the opposing variable phase networks. This makes for symmetry of interaction between the amplifiers and the networks. Measurement of the advantage gained from the use of the switching technique shows that potential error due to amplifier drift has been reduced from approximately 0.4° per degree change of ambient to 0.05°.

The prime element of the phase standard is a continuously variable

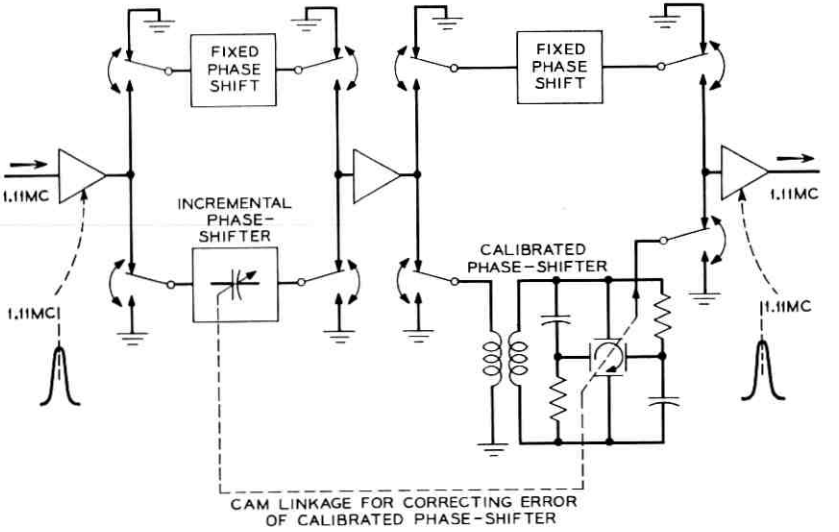


Fig. 15 — Block diagram of phase standard showing use of rapid switching to prevent calibration drift due to shift of amplifier parameters with temperature or age.

four-quadrant sine condenser of high quality and permanence. The linearity error is removed by use of an earlier noted technique.[1] A cam, whose profile is shaped to match the error curve, rotates with the condenser shaft. A follower arm resting on the cam periphery then adjusts an incremental phase shifter to compensate for the non-linearity error.

In addition to phase, the calibrated dials also bear delay scales (in different color than that used for phase); the conversion between delay and phase is on the basis of the equivalence between 0.5 nanosecond of delay and one degree of phase at the 1.11-mc IF. The dials may be slipped for setting up dummy "zeros".

The active circuitry in the common path follows standard design.

5.4 *Loss Standard*

The loss standard operates at 1.11 mc in the phase mode of measurement and at 60 mc in the delay mode. A calibration accuracy of 0.03 db is sought for loss differences up to 30 db. In addition, change of loss setting should not alter the insertion phase of the standard when measuring loss, or the insertion delay when delay is being measured, since phase and delay changes of the standards are indistinguishable from changes of the parameters of the device under test.

A decade attenuator of the dissipative type employing metal-film resistors performs satisfactorily in all respects. The loss standard consists of four such attenuators connected in tandem, spanning a total range of 49.99 db in 0.01-db steps.

Each of the decades is made up of a sequence of π pads mounted around a turret switch. The individual pads are semi-coaxial in design and this endows the loss standard with the bandwidth required for the dual operation at 1.11 mc and 60 mc. At the 1.11-mc level, phase shift is constant to 0.1 degree for any setting of the standard. The envelope delay at 60 mc changes less than 0.2 nanoseconds up to changes of 20 db in loss setting.

5.5 *Loss and Phase Detection*

The essentials of the loss detector operation, as illustrated in Fig. 9, have been discussed previously.

A crucial aspect of the operation of the loss detector relates to the timing of sampling relays, S_8 and S_9, which establish the charging paths to the storage capacitors, C_x and C_s. In connecting C_x and C_s during the sample periods, it is necessary to allow for the physical impossibility of perfectly synchronizing all the relays in the measuring

set with respect to instant of contact transfer and uniformity of dwell time. Moreover, short time transients are initiated at the change of state from x to s and vice versa. For these reasons, S_8 and S_9 are timed so as to delay the start of the sampling intervals until transients set up by the operation of preceding relays have decayed.

The component circuitry involved in the phase detection process, as illustrated in Fig. 16, consists of an AGC circuit to smooth level without accompanying phase change, a sum and difference discriminator with associated sample and hold circuitry at the output, and an APC loop preceding the reference input to the discriminator.

The AGC circuit in Fig. 16 introduces less than 0.1-degree change of phase in 1.11-mc transmission for a 10-db change of input level. Its operation is based on the use of an emitter-coupled transistor pair, Q_1 and Q_2, for gain control. The signal input current, which is applied to the common connection between the emitters, divides between Q_1 and Q_2 in accordance with the AGC error current feeding into the base of Q_2. Since the external emitter circuits of Q_1 and Q_2 share a large dc impedance, changes of dc operating point in Q_2 induce exactly opposite shifts of operating point in Q_1. Hence, the ac impedance (for small signals), seen looking into each of the emitters, shift in opposite directions with the result that the impedance presented to the signal input remains essentially constant over a wide range of gain control. The constancy of the load on the driving source is a prime factor in minimizing the level-to-phase conversion.

The grounded-base operation of Q_1 and Q_2, in so far as ac signal effects are concerned, is an additional aid in suppressing level-to-phase conversion. In this configuration, the parasitic parameters of the transistor shift least with operating point. Thus, by choosing a transistor of high f_α in relation to the operating frequency (1.11 mc), virtually all anomalous effects, including those ascribable to the phase angle of α, are eliminated.

Another aspect of interest arising in the process of phase detection concerns the operation of the APC loop. This loop is unusual with respect to its ability to absorb indefinitely large amounts of stress, without saturating. The operation is as follows: During the period when the S path is closed throughout the test set, signals e_s and e_r appear at the inputs to the phase discriminator in Fig. 16. For an extended interval during this period, the discriminator delivers charging current to capacitor C_2 through sampling relay S_2. Capacitor C_2 charges up to a dc voltage proportional to the deviation of the phase difference between e_s and e_r from ninety degrees. As external conditions change in a way

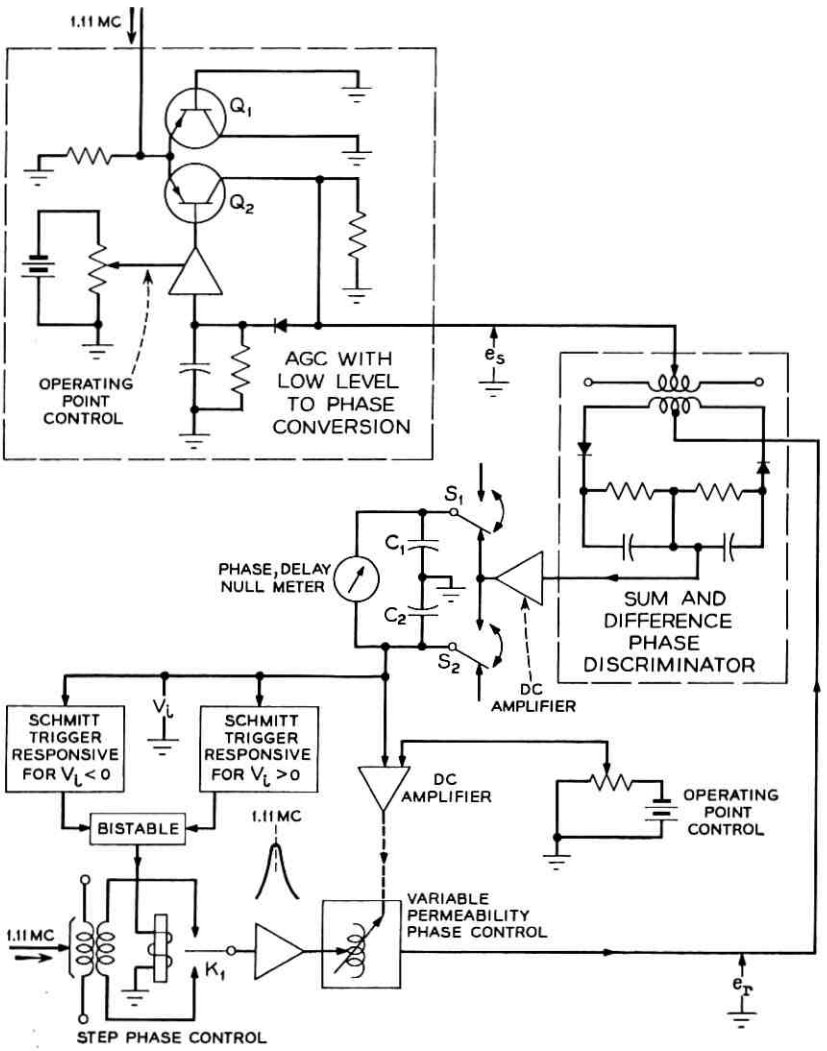


Fig. 16 — Phase detection including AGC circuit with insignificant level-to-phase conversion coefficient and APC circuit to fix operating point at phase discriminator.

which tends to increase the magnitude of this deviation, the APC connection through the voltage-sensitive phase shifters alters the phase of e_r to sustain near-quadrature between the discriminator inputs.

Two kinds of controllable phase networks are present. A continuous control, which makes use of permeability—current sensitivity to alter

inductance values in a cascade of "bridge" type phase shifters can introduce phase changes of up to 200 degrees of either sign. As these limits are approached in the course of introducing corrections, the voltage magnitude across capacitor C_2 increases sufficiently to trip one of the two Schmitt triggers. This occurs when the deviation from quadrature rises to 3 degrees. The particular trigger actuated depends directly on the sense of the loop stress, since this determines whether the departure of the C_2 voltage from null is positive or negative.

The firing of the Schmitt trigger precipitates the operation of relay K_1, which serves to alter the phase of e_r by 180 degrees. This has the effect of subtracting 180 degrees from the initially present stress, thereby removing the cause of loop strain and permitting the voltage across capacitor C_2 to return nearly to zero. This operation can be accomplished repeatedly so that effectively the loop can absorb indefinitely large amounts of stress. It is of interest that the operation of relay K_1 produces the stable effect just described only if another relay is employed to momentarily short the voltage across capacitor C_2 at the instant that relay K_1 changes dwell state. The erasure of "past history" establishes propitious boundary conditions from which the loop executes a stable operating point trajectory to the new state.

Both the step and continuous phase shifters are designed to alter phase, but not level; hence, the operation of the APC loop does not in any way alter the deflection sensitivity on the null meter.

VI. MEASUREMENT ACCURACY

Errors tend to be of two types. The first category includes errors attributable to residual imperfections of the test set. These result from such factors as calibration uncertainty of standards, the presence of low residual crosstalks and pickups, the minor influence of harmonic residues in various parts of the test set and small effects of noise. The error estimates, previously cited in Section 3.3, cover the sources just enumerated and apply equally to the measurement of transistors and coaxial unknowns.

Further errors, however, arise in the measurement of coaxial unknowns which depend upon interaction between the frequency characteristics of the unknown and certain of the attributes of the test set. Included here are errors due to the residual impedance misterminations around the unknown and "aperture" error in delay measurement.

6.1 *Mistermination Errors in Loss, Delay and Phase Measurement*[1,11]

The source and load impedances facing the unknown in this test set have finite return losses which drop to a minimum of 20 db at 4 gc.

When the transmission characteristics of the unknown are measured between these slightly imperfect impedances, the measured data is somewhat different than one would obtain if measurements were made between terminations of infinite return loss. The difference between the actual measurement and those which would be obtained between perfect impedances is the mistermination error.

An estimate has been made for the insertion loss, phase, and delay error due to mistermination. The starting point for such an estimate is the error, ε_f, in measurement of insertion ratio.[1] The magnitude of insertion ratio is the insertion loss expressed as the corresponding numerical ratio, and the angle of insertion ratio is insertion phase. The value of ε_f is given by

$$\varepsilon_f = \frac{\text{measured insertion ratio}}{\text{insertion ratio between 50 } \Omega \text{ impedances}}$$

$$= \frac{1 - s_{22}\rho_L - s_{11}\rho_G - \rho_G\rho_L(s_{12}s_{21} - s_{11}s_{22})}{1 - \rho_G\rho_L}. \tag{5}$$

The s coefficients are the scattering parameters of the unknown, and ρ_L and ρ_G are the reflection coefficients of source and load in the test set. The impedance reference here is 50 ohms. If the reasonable assumption is made that all of the reflection coefficients are small, i.e., $|s_{11}|$, $|s_{22}|$, $|\rho_G|$ and $|\rho_L| < 0.1$, then ε_f may be approximated by

$$\varepsilon_f = 1 - s_{11}\rho_G - s_{22}\rho_L - \rho_G\rho_L(s_{12}s_{21} - s_{11}s_{22} - 1). \tag{6}$$

If further, one deals with the worst case in which the round trip loss through the unknown is 0 db, i.e., $|s_{12}s_{21}| = 1$, then the largest possible error in loss or phase measurement would be closely equal to

$$|s_{11}\rho_G| + |s_{22}\rho_L| + 2|\rho_G\rho_L| \tag{7}$$

in nepers or radians.

Equation (6) is also useful in estimating upper bounds on delay measurement error due to mistermination. When one recalls that the measured envelope delay is equal to the increment of insertion phase shift across the 5.55-mc separation between RF carrier and adjacent sidetone divided by the radian interval between these frequencies, it is then apparent that the error in measuring insertion delay equals the difference of the errors in insertion phase at the two frequencies divided by the radian interval.

If the same assumptions with respect to reflection and transmission magnitude are made which apply to (7), and if the angles of the quantities in (6) are disposed to produce the maximum, oppositely sensed

errors in phase at the two frequencies defining the interval, then the maximum possible error in delay for a network having small loss and only modest reflections would be

$$2\,\frac{|\,s_{11}\rho_G\,| + |\,s_{22}\rho_L\,| + 2\,|\,\rho_G\rho_L\,|}{\Delta}\ \text{seconds}, \tag{8}$$

where it is assumed that $|\,s_{11}\,|$, $|\,s_{22}\,|$, $|\,\rho_G\,|$, and $|\,\rho_L\,|$ are the same at both frequencies. Δ is the radian frequency separation between the RF tones bounding the interval.

For example, consider the application of (7) and (8) at 2.0 gc in the present set where $|\,\rho_G\,|$ and $|\,\rho_L\,|$ are approximately 0.03. Under these circumstances, when measuring an unknown having $|\,s_{11}\,|$ and $|\,s_{22}\,|$ equal to 0.1, (7) and (8) show that the error in loss, phase, and delay measurement could approach 0.06 db, 0.4 degree, and 0.45 nanoseconds. Measured at 4 gc, the network just considered could be erroneously measured by as much as 0.3 db for loss, 2.3 degrees for phase, and 2.2 nanoseconds for delay in view of the increase in test set reflections to a level of 0.1.

The previous equations define absolute upper bounds on measurement error due to mistermination. More realistic error limits, particular to a given situation, may be obtained by applying (5) directly, when the requisite information is available. Advantage may also be taken of constraints imposed on the s parameters by virtue of passivity conditions in order to further bound mistermination errors.[11]

6.2 *Aperture Error in Delay Measurement*

The test set applies an amplitude-modulated wave to the unknown but only the carrier and one of the adjacent side tones ultimately beat together, after downward frequency translation, to form detected signal. The aperture errors are hence characteristic of those encountered in "two-tone" measurement sets, rather than the larger errors produced in a "three-tone" set.[7] The origin of the error is suggested in the Fig. 17. The test set indicates the slope of the secant line connecting the phase ordinates at ω_1 and ω_2. When higher-order curvature exists between ω_1 and ω_2, the slope of the secant line no longer is the same as the slope of the tangent, $\partial\theta/\partial\omega$, drawn at the mean frequency, $(\omega_1 + \omega_2)/2$, where the delay is considered to be evaluated.

If the phase curve of the unknown exhibits only algebraic variation over the frequency range encompassed by the measurements, elementary deductions are then possible with respect to errors in measurement of *delay distortion*, i.e., in measurement of the variation of delay across

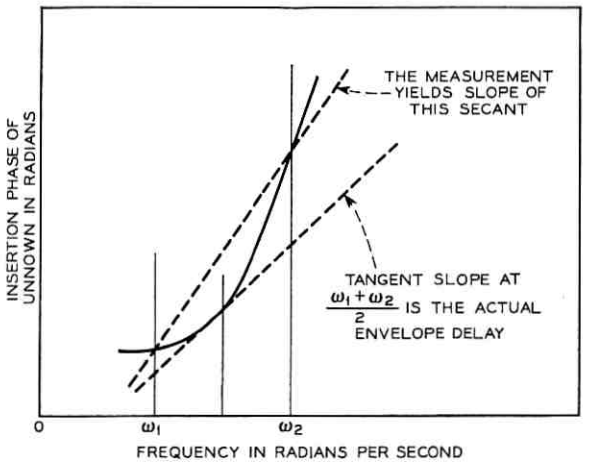


Fig. 17 — "Aperture error" in envelope delay measurement.

the test frequency range. When the phase curve is expressible as the polynominal

$$\Phi = a_0 + a_1\omega + a_2\omega^2 + \cdots a_n\omega^n, \tag{9}$$

it may then be demonstrated that no error results in delay distortion measurement for $n \leqq 3$. Hence, there is no aperture error when measuring a linear or a parabolic delay shape, or an additive combination of these shapes.

On the other hand, quite large errors can result in the measurement of ripples of delay superimposed on algebraic shapes. If the period of the ripple is Q and p is the separation between the two delay measuring tones, then the indicated value of the amplitude of a delay sinuousity having a peak-to-peak value, τ_0, would be erroneously indicated by

$$\tau_0' = \tau_0 \frac{\sin \frac{\pi p}{Q}}{\frac{\pi p}{Q}} \tag{10}$$

The magnitude of the error may be appreciated from the fact that the indicated value of τ_0 is in error by 7 per cent when the ratio of Q to p equals 5.

6.3 *Accuracy Validation*

Broadband standards of loss, phase, and delay, operable between 0.25 and 4.2 gc and having sufficient accuracy for use as a checking

standard, are not available. Hence, indirect techniques of validation had to be resorted to.

As a first step, the insertion loss and phase of each of two coaxial pads was measured at a large number of frequencies. The sum of these measurements was then compared with the over-all measurement of loss and phase through the tandem connection of the two pads. Study of the equation for ε_f in Section 6.1, as applied to the accuracy validation problem, shows that this comparison yields a valid measure of test set error, provided that the test set and pad reflections are sufficiently small. For the frequency range up to 2 gc, over which the test set terminations are <0.03, the use of pad standards with reflections of 0.03 would permit accuracy determinations, by the method just outlined, to a tolerance of 0.04 db and 0.25 degrees. Such pads are available, and when the "bootstrap" experiment was made using a pair of nominal 20-db units, it was found that the sum of individual measurement agreed with the over-all measurement to within 0.1 db and 0.3 degree up to 2 gc. The agreement proved to be about the same for a pair of 10-db pads.

Because of the increased reflection from both the test set and the pads above 2 gc, less favorable results were obtained in this region. The disagreement between arithmetic sum and measured sum cycled with test frequency, and this strongly suggested that the cause lay with reflection coefficient interactions. This was confirmed by building up to a given value of tandem loss using different combinations of pad values. The results varied with specific pad combinations even though the electrical lengths of all the pads were about equal.

In similar tests to evaluate the delay measurement performance, it was possible to correctly sum 2 cables of 20 nanosecond length to within 0.4 nanoseconds up to 4 gc.

VII. EQUIPMENT DESIGN

The test set apparatus is distributed among four bays. One of these bays provides a tightly shielded compartment for the test frequency signal sources. The remaining three bays are united to form a three-cabinet console. This arrangement is shown in Fig. 18.

The layout has been executed with operator convenience foremost in mind. Just above table level in the right-hand bay of the console are the programming knobs for selecting between loss or gain, and between phase or delay modes of measurement. An extensive array of coaxial switching automatically sets up the test set circuitry for measurement in any of the desired modes, on command from the programming knobs. Also occupying convenient locations in the right-hand bay are the

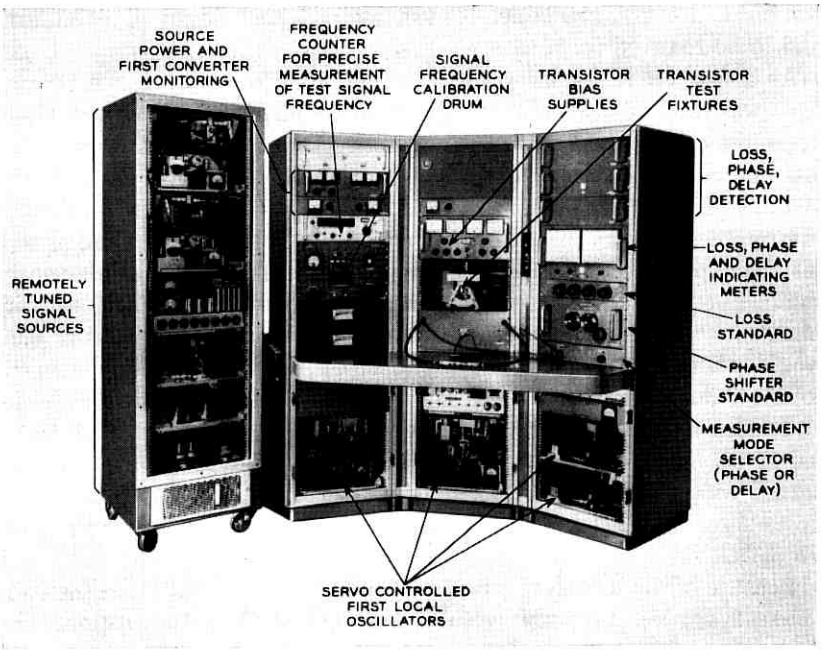


Fig. 18 — 0.25 to 4.2-gc transmission, phase and delay measuring set shown with front covers removed.

calibrated measurement standards and the associated null meters. The various scales on the meters are edge lit to unambiguously indicate the particular scale in use. Automatic ranging circuitry transfers operation to the more coarsely calibrated scales whenever the prevailing loss, gain, phase, or delay unbalances exceed the end limits of the most sensitive scales.

The left-hand bay contains the controls for selecting test signal frequency. The point-by-point and motor scan controls are visible in Fig. 18, just below, and to the left and right of the drum dial which bears the frequency calibrations for all four test signal bands. Monitor meters for indicating the power output of the sources and the dc currents flowing in the converter crystals are also provided in the left-hand bay. These are of no interest to the operator, but do play an important role in preventive maintenance. Each of the chassis modules is mounted on slides, thus making in-situ maintenance feasible.

Resting on the table, and emerging from the center bay, are the flexible cables for connection of coaxially terminated unknowns. Each of these cables is terminated in a 15-db impedance buffering pad. The

center bay also houses the built-in transistor measurement facilities. The jig and associated coaxial trombone mount on a retractable slide which the operator pulls out when making transistor measurements. Connections are made to the input and output test set cables through Type N-Dezifix B coaxial adapters.

VIII. ACKNOWLEDGMENTS

Mr. F. R. Dennis provided the test signal sources and remote tuning servo, as well as the first LO system and associated electro-mechanical servo control. He was also responsible for the original design of the control circuit for automatically switching operation of the set from one band to another.

Mr. W. G. Hammett was responsible for all the design effort involved in adapting the test set to the measurement of transistor characteristics. This included the design and characterization of all of the specially required coaxial apparatus as well as the analytic procedures for reducing measurement data to transistor parameters and estimating the errors involved. Mr. Hammett furnished the design for the VCO frequency control loop and related capture circuitry, phase standard, APC loop for the automatic regulation of the operating point at the phase detector, and the loss and phase detection circuitry.

Mr. O. Kummer contributed the over-all front-end design and specification of test set level patterns to achieve objectives of linearity, crosstalk, and S/N. This work included the provision of the converters and associated isolation circuitry. Mr. Kummer designed the RF comparison switching complex, the loss standard, all of the special circuitry involved in the measurement of envelope delay and the programming features for automatic set up of the instrument in either phase or delay measurement modes. Mr. Kummer was also responsible for over-all accuracy validation.

Mr. R. A. Berner made contributions to many areas of the set through fabrication and test of brassboard models and by supervising many of the details of construction.

As supervisor of the group which developed the measuring instrument, the author held the over-all project responsibility. The author's superior, Mr. S. Doba, Jr., contributed suggestions and comments which often led to improvements in the design.

REFERENCES

1. Leed, D. and Kummer, O., A Loss and Phase Set for Measuring Transistor Parameters and Two Port Networks between 5 and 250 mc, B.S.T.J., *40*, May, 1961, pp. 841–884.

2. Alsberg, D. A., Precise Sweep Frequency Method of Vector Impedance Measurement, Proc IRE, *39,* November, 1951, pp. 1393–1400.
3. Alsberg, D. A. and Leed, D., A Precise Direct Reading Phase and Transmission Measurement System for Video Frequencies, B.S.T.J., *28,* April, 1949, pp. 221–238.
4. Meinke, H. H. and Scheuber, A., Die Berechnung der Ubertrogungseigenschaften Zylin der Symmetriche Bauelemente Koaxialer Leitung aus dem Verhalten von Ebenen Elektrostatischen Feldern, Arch. Elektr. Ubertragung, Band 6, pp. 221–227, June, 1952.
5. Montgomery, C. G., Dicke, R. H., Purcell, E. M., *Principles of Microwave Circuits,* McGraw-Hill Book Company, Inc., New York, 1948, pp. 150–151.
6. Oliner, A. A., Calibration of Slotted Section for Precision Microwave Measurements, Review of Scientific Instruments, *25,* January, 1954, pp. 13–20.
7. Magnusson, R., Sensitive Group Delay Meters, Ericsson Technics, No. 1, 1957, pp. 110–141.
8. Goldman, S., *Frequency Analysis, Modulation and Noise,* McGraw-Hill Book Company, Inc., New York, 1948, p. 246, Eq. 114.
9. Toro, V. and Parker, S. R., *Principles of Control Systems Engineering,* McGraw-Hill Book Company, Inc., New York, 1960, pp. 132–139.
10. Leed, D., Automatic Frequency Control Circuit, U.S. Patent No. 2,610,297, December 14, 1948.
11. Youla, D. C. and Paterno, P. M., Realizable Limits of Error for Dissipationless Attenuators in Mismatched Systems, IEEE Trans., Microwave Theory and Techniques, *MTT-12,* May, 1964, pp. 289–299.

Computing the Spectrum of a Binary Group Code*

By M. M. BUCHNER, JR.

(Manuscript received December 10, 1965)

The weight distribution of the code vectors of a binary group code has been referred to as the spectrum of the code. This paper presents a technique for calculating the spectrum of such a code, the spectra of shortened codes obtainable from the code, and what are defined as the level weight structures of the code.

The method is conceptually straightforward and readily adaptable to digital computers. It involves operations no more complex than the addition of two $(n - k)$-tuples, the determination of the weight of certain $(n - k)$-tuples, and the ordinary addition of certain integers. Its computational complexities are independent of the code parameters. In principle, it may be used for any binary group code, but it is particularly useful for codes in which the number of parity check positions per code vector is rather small although the number of information positions may be large.

I. INTRODUCTION

The need for reliable data transmission systems has prompted the investigation of various coding techniques which attempt to detect and/or correct transmission errors. Because of the relative ease with which binary codes can be implemented, these codes have received special attention. It is with certain properties of these codes that this paper is concerned.

In general, the encoder receives a block of k binary symbols (called a message) from a message source from which it determines $(n - k)$ binary parity check symbols (called an ending). The message symbols and the ending symbols may be interleaved or transmitted sequentially thus forming a block of length n (called a code vector). Because any

* The material presented in this paper formed Appendix II of the dissertation "Coding for Numerical Data Transmission" submitted by the author to The Johns Hopkins University in conformity with the requirements for the degree Doctor of Philosophy.

code in which these symbols are interleaved is equivalent[1] to a code in which the message and ending are transmitted sequentially, attention may be restricted to the latter situation.

The elements 0 and 1 form a field. Two vectors (or n-tuples) whose components are these field elements may be added by adding modulo 2 the corresponding components of each vector. The symbol $\oplus$ will be used to denote this addition of vectors.

The set of all possible n-tuples forms a vector space V_n of dimension n over the field of two elements. A subset V is said to form a group code if the n-tuples in the subset form a group. Over the field of two elements, a set of vectors that forms a group is a subspace of V_n. Therefore, the vectors of a group code form a subspace of V_n.

The weight of a vector u is the number of nonzero components in u and is denoted by $w[u]$. The distance[1] between two code vectors u and v is $w[u \oplus v]$. Because the code vectors form a group, there exists a code vector $t = u \oplus v$. The distance between u and v is thus equal to $w[t]$.

Because of this relationship between code vector weights and distances between code vectors, it is useful in evaluating the error detecting and/or correcting capabilities of group codes to be able to determine the number of code vectors of each possible weight — i.e., from 0 to n. This information has been called the spectrum of a code and can in principle be obtained by calculating in detail each of the possible 2^k code vectors and then determining the weight of each of these code vectors. However, this method is not computationally feasible for values of k which are most often of interest.

MacWilliams[2] has determined a system of linear equations which relate the set of integers that forms the spectrum of a given code to the set of integers that forms the spectrum of its dual code. The method is particularly effective for codes in which the dimension of the dual code is relatively small so that the spectrum of the dual code is readily obtained.

The method presented herein enables the direct computation of the spectrum of a code without the actual formation of every code vector. The technique also gives both the spectrum of each of the possible shortened codes which may be obtained from the given code and the level weight structures of the given code. The level weight structures (which are defined later in this paper) have proved useful in the study of the effectiveness of error-correcting codes for numerical data transmission[3] and may indeed be of interest in other areas of code evaluation.

The method is conceptually quite simple, readily implemented on a digital computer, and does not depend upon the solution of any equa-

tions. In fact, the only operations involved are the component by component modulo 2 addition of $(n - k)$-tuples, the determination of the weight of certain $(n - k)$-tuples and the ordinary addition of certain integers.

II. COMPUTATIONAL TECHNIQUE

Let k denote the dimension of the code space V and let E_j $(1 \leqq j \leqq k)$ denote the k basis vectors of V. Take E_j in the usual systematic form

$$E_j = e_j \mid C_j \tag{1}$$

where the message e_j is the k-tuple with a 1 in position j and all other positions 0 and C_j is the $(n - k)$-tuple ending assigned to the message e_j. Note that if the code is specified by a parity check matrix[1] in the form

$$H = (h_1 h_2 \cdots h_k I_{n-k}) \tag{2}$$

where h_i $(1 \leqq i \leqq k)$ is the column of H in the ith position and I_{n-k} is the $(n - k) \times (n - k)$ identity matrix, then C_j is simply the transpose of h_j and E_j is readily obtainable.

The vectors $E_1, E_2, \cdots, E_l$ generate a subspace of V of dimension l which we shall denote as Γ_l. Γ_k is the code itself and Γ_l is the set of code vectors in which information positions $l + 1, l + 2, \cdots, k$ are 0. Γ_0 is defined as consisting exclusively of the all 0 code vector.

Let $\Lambda_l = \Gamma_l - \Gamma_{l-1} \cdot \Lambda_l$, which is called the l-level of the code, is the set of code vectors in which information positions $l + 1, l + 2, \cdots, k$ are 0 and information position l is 1. Any code vector in Λ_l is the sum of E_l and some vector in Γ_{l-1}.

The basic idea is to form for Γ_l an ending-weight matrix $S^{(l)}$. For convenience we shall deviate from usual practice and number the rows and columns of $S^{(l)}$ beginning with 0. The entry $s_{\alpha,\beta}{}^{(l)}$ in row α and column β of $S^{(l)}$ denotes the number of code vectors in Γ_l of weight α whose endings are the $(n - k)$-bit binary representation of β (denoted by $B(\beta)$). There must be $(n + 1)$ rows in $S^{(l)}$ to allow for all possible code vector weights and 2^{n-k} columns to allow for all of the possible $(n - k)$-bit endings. Therefore, $S^{(l)}$ is an $(n + 1) \times 2^{n-k}$ matrix.

The utility of this technique lies in the ease with which $S^{(l)}$ may be obtained from $S^{(l-1)}$. Suppose that $S^{(l-1)}$ is known. The code vectors of Λ_l are formed by adding E_l to the code vectors of Γ_{l-1}. However, the special form of E_l makes this operation equivalent to placing a 1 in information position l of each vector in Γ_{l-1} and, at the same time, adding C_l to the ending of each code vector in Γ_{l-1}.

Any code vector of weight α in Γ_{l-1} whose ending is $B(\beta)$ becomes a vector in Λ_l with ending $B(\beta) \oplus C_l$ and of weight γ where

$$\gamma = \alpha + 1 + w[B(\beta) \oplus C_l] - w[B(\beta)]. \tag{3}$$

For those values of α $(0 \leqq \alpha \leqq n)$ and γ $(0 \leqq \gamma \leqq n)$ for which (3) may be satisfied,

$$s_{\gamma, B^{-1}[B(\beta)\oplus C_l]}^{(l)} = s_{\gamma, B^{-1}[B(\beta)\oplus C_l]}^{(l-1)} + s_{\alpha,\beta}^{(l-1)} \tag{4}$$

where B^{-1} (the inverse of B) is the operator such that $\psi = B^{-1}B[\psi]$.

In general, it is not possible to satisfy (3) for every γ $(0 \leqq \gamma \leqq n)$. However, because all code vectors in Γ_{l-1} whose endings are $B(\beta)$ (i.e., the code vectors giving rise to the nonzero entries in column β of $S^{(l-1)}$) become code vectors in Λ_l of weight in the range 0 through n, all values of α corresponding to nonzero entries in column β of $S^{(l-1)}$ produce values of γ such that $0 \leqq \gamma \leqq n$. For these values of γ, (4) may be applied.

On the other hand, values of γ which would require values of α outside of the range $0 \leqq \alpha \leqq n$ in order to satisfy (3) are those values of γ for which it is impossible to have code vectors in Γ_{l-1} of weight α whose endings are $B(\beta)$. For these values of γ,

$$s_{\gamma, B^{-1}[B(\beta)\oplus C_l]}^{(l)} = s_{\gamma, B^{-1}[B(\beta)\oplus C_l]}^{(l-1)}. \tag{5}$$

The column numbers referred to in (4) and (5) are independent of α. Furthermore, as α increases, (4) and (5) simply refer to different elements in the same column. For this reason, these results may be expressed as column operations thus leading to a conceptually simple result.

Let $s_\beta^{(l)}$ denote column β in $S^{(l)}$. Define $\sigma^{(j)}$ to be a shifting operator which, when applied to $s_\beta^{(l)}$, shifts each element of $s_\beta^{(l)}$ by j positions filling in any resulting blank positions with zeros. For example, if

$$s_\beta^{(l)} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

then

$$\sigma^{(2)} \cdot s_\beta^{(l)} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

and

$$\sigma^{(-1)} \cdot s_\beta{}^{(l)} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

In terms of $\sigma^{(j)}$, the relationships expressed in (3), (4), and (5) may be conveniently expressed as

$$s_{B^{-1}[B(\beta)\oplus C_l]}{}^{(l)} = s_{B^{-1}[B(\beta)\oplus C_l]}{}^{(l-1)} + \sigma^{(w[B(\beta)\oplus C_l]-w[B(\beta)]+1)} \cdot s_\beta{}^{(l-1)}. \quad (6)$$

Clearly all columns of $S^{(l)}$ are obtained by successively applying (6) as β runs from 0 to $2^{n-k} - 1$.

It is important to notice the great simplicity of (6). In practice, it involves shifting one column of $S^{(l-1)}$ and then combining by simple addition the elements of this column with those of another column of $S^{(l-1)}$ to obtain a column of $S^{(l)}$. Determining the number of positions that $s_\beta{}^{(l-1)}$ should be shifted and the column with which $s_\beta{}^{(l-1)}$ should be combined is extremely easy. In particular, the operations in (6) are readily adapted to digital computer operations.

If $S^{(0)}$ is known, the remaining ending-weight matrices can be successively obtained. The only code vector in Γ_0 is the all 0 vector. Therefore,

$$s_{00}{}^{(0)} = 1$$

and

$$s_{\alpha,\beta}{}^{(0)} = 0$$

for all other values of α and β.

Now that the method for constructing the ending-weight matrices has been presented, the following will serve to indicate how the desired information is extracted.

(*i*) Spectrum of the code: The total number of code vectors of weight α in the code is

$$\sum_{\beta=0}^{2^{n-k}-1} s_{\alpha,\beta}{}^{(k)}. \quad (7)$$

The spectrum of the code is obtained from $S^{(k)}$ by using (7) for each value of $\alpha (0 \leqq \alpha \leqq n)$.

(*ii*) Spectra of shortened codes: Let k' denote the number of information positions in the shortened code — i.e., $k - k'$ information positions are deleted. Assume that the deleted posi-

tions are information positions $k' + 1, k' + 2, \cdots, k$. The total number of code vectors of weight α in the shortened code is

$$\sum_{\beta=0}^{2^{n-k}-1} S_{\alpha,\beta}^{(k')}. \tag{8}$$

The spectrum of the shortened code is obtained from $S^{(k')}$ by using (8) for each value of $\alpha (0 \leqq \alpha \leqq n - k + k')$.

(*iii*) Level weight structure: The set of code vectors Λ_l has been referred to as the l-level code vectors and the weight distribution of these code vectors as the l-level weight structure.[3] Note that the l-level weight structure is the difference between the spectrum of the shortened code consisting of l information positions and the spectrum of the shortened code consisting of $(l - 1)$ information positions.

Let $n_{l,\alpha}$ denote the number of code vectors of weight α on the l-level. The number of code vectors of weight α in Γ_{l-1} is

$$\sum_{\beta=0}^{2^{n-k}-1} S_{\alpha,\beta}^{(l-1)}.$$

Similarly, the number of code vectors of weight α in Γ_l is

$$\sum_{\beta=0}^{2^{n-k}-1} S_{\alpha,\beta}^{(l)}.$$

It follows that

$$n_{l,\alpha} = \sum_{\beta=0}^{2^{n-k}-1} S_{\alpha,\beta}^{(l)} - \sum_{\beta=0}^{2^{n-k}-1} S_{\alpha,\beta}^{(l-1)}. \tag{9}$$

III. CONCLUSIONS

The spectrum of any group code, the spectrum of any shortened code, and all level weight structures are obtainable in a straightforward manner by means of operations no more complex than the addition of two $(n - k)$-tuples (to determine the columns to combine), the computation of the weight of certain $(n - k)$-tuples, and the repeated addition of integers two at a time (to actually combine the columns). The number of computations does depend upon the parameters n and k but the method has the advantage that the complexity of the operations is invariant. Because the number of computations and the number of computer storage locations required for the ending-weight matrix are sensitive to changes in $(n - k)$ but rather insensitive to changes in k, the method is most effective for codes in which $(n - k)$ is moderate although k may be quite large.

As presented, the method treats each of the k ending-weight matrices in a similar manner by determining all of the $(n + 1)\cdot 2^{n-k}$ entries of each matrix. Computing time can be saved by realizing that the maximum possible weight of a vector in Γ_l is $l + n - k$ and, thus, that it is only necessary to compute the first $l + n - k$ rows of $S^{(l)}$ because the remaining rows contain zero entries exclusively. Additional programming sophistications, including processing only those columns of $S^{(l-1)}$ which contain nonzero entries in obtaining $S^{(l)}$ (particularly for the smaller values of l), improve the computing efficiency of the method.

This technique was originally developed for computing the level weight structures of certain codes. Thus, if the level weight structures and/or the spectra of the shortened codes are desired, this method offers a straightforward and effective means of obtaining such information. However, if all that is desired is the spectrum of the code, then under some conditions the method developed by MacWilliams[2] may be preferable from a computing time point of view although the conceptual simplicity of this method is still appealing. In any case, the relative advantages of the two methods should be considered before deciding which to use for a specific application.

The method has been used successfully to compute the level weight structures and the spectra of the (15,11), (31,26), and (63,57) Hamming perfect single error-correcting codes. In each case the information was obtained on an IBM 7094 digital computer in less than 0.01 hours.

IV. NUMERICAL EXAMPLE

The parity check matrix for a (7,4) Hamming code is

$$H = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}.$$

The basis vectors for this code are

$$\begin{aligned} E_1 &= 1000 \quad 111 \\ E_2 &= 0100 \quad 110 \\ E_3 &= 0010 \quad 101 \\ E_4 &= \underbrace{0001}\ \underbrace{011} \\ &\qquad \text{message} \quad \text{ending.} \end{aligned}$$

The sets of code vectors referred to as Γ_3 and Λ_4 are listed in Table I. The appropriate values of α and β are given next to each vector.

TABLE I

Γ_3	α	β	Λ_4	α	β
0000 000	0	0	0001 011	3	3
1000 111	4	7	1001 100	3	4
0100 110	3	6	0101 101	4	5
1100 001	3	1	1101 010	4	2
0010 101	3	5	0011 110	4	6
1010 010	3	2	1011 001	4	1
0110 011	4	3	0111 000	3	0
1110 100	4	4	1111 111	7	7

Tabulating this information yields $S^{(3)}$ and $S^{(4)}$.

$S^{(3)}:\alpha$ \ β

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

$S^{(4)}:\alpha$ \ β

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

We now turn to use the method herein developed to obtain $S^{(4)}$ from $S^{(3)}$. Specifically, we use (6) first with $\beta = 0$ and then successively increase β until $\beta = 7$.

When $\beta = 0$, $B^{-1}[B(0) \oplus C_4] = 3$. Thus, (6) reduces to

$$s_3^{(4)} = s_3^{(3)} + \sigma^{(3)} \cdot s_0^{(3)} = \begin{bmatrix}0\\0\\0\\0\\1\\0\\0\\0\end{bmatrix} + \begin{bmatrix}0\\0\\0\\1\\0\\0\\0\\0\end{bmatrix} = \begin{bmatrix}0\\0\\0\\1\\1\\0\\0\\0\end{bmatrix}$$

which is indeed correct.

Now let $\beta = 1$. Then $B^{-1}[B(1) \oplus C_4] = 2$ so (6) yields

$$s_2^{(4)} = s_2^{(3)} + \sigma^{(1)} \cdot s_1^{(3)} = \begin{bmatrix}0\\0\\0\\1\\0\\0\\0\\0\end{bmatrix} + \begin{bmatrix}0\\0\\0\\0\\1\\0\\0\\0\end{bmatrix} = \begin{bmatrix}0\\0\\0\\1\\1\\0\\0\\0\end{bmatrix}.$$

The remaining columns of $S^{(4)}$ are obtained in a like manner as β increases to 7.

The spectrum of the code is obtained by summing across the rows of $S^{(4)}$. The spectrum of the shortened code resulting from the deletion of the fourth information position is obtained by summing across the rows of $S^{(3)}$. The 4-level weight structure is the difference between these spectra. This information is tabulated in Table II.

TABLE II

Weight	Code Spectrum	Shortened Code Spectrum	4-Level Weight Structure
0	1	1	0
1	0	0	0
2	0	0	0
3	7	4	3
4	7	3	4
5	0	0	0
6	0	0	0
7	1	0	1

For an illustrative example, it was necessary to confine ourselves to a code in which k is small. However, it should be realized that the true utility of the method lies in the fact that it can, without modification or additional complexity, be used for codes in which k is quite large.

REFERENCES

1. Peterson, W. W., *Error Correcting Codes,* M.I.T. Press and John Wiley and Sons, 1961.
2. MacWilliams, J., A Theorem on the Distribution of Weights in a Systematic Code, B.S.T.J., *42,* January, 1963, pp. 79–94.
3. Buchner, M. M., Jr., Coding for Numerical Data Transmission, Ph.D. Dissertation, The Johns Hopkins University, Baltimore, Maryland, 1965.

The Statistical Effects of Random Variations in the Components of a Beam Waveguide

By WILLIAM H. STEIER

(Manuscript received November 22, 1965)

The effects of variations in the components of a beam waveguide are considered. These variations statistically cause the Gaussian beam spot size of the light propagating down the waveguide to grow and cause the beam center to oscillate about the waveguide axis with ever-increasing amplitude. Random variations in lens focal length and spacing and random lateral lens displacements are considered. It is also shown how random variations in focal length and spacing can be included in the published analyses for short random bends in the waveguide axis.

When the number of lenses is large, it is shown that the beam displacement and beam spot size both grow exponentially with distance.

As an example, a confocal waveguide with lenses spaced one meter apart and built to somewhat optimistic tolerances will require a beam redirector every 2.5 kilometers to prevent the beam oscillations from exceeding an rms value of 2 millimeters.

I. INTRODUCTION

A long sequence of spaced lenses is of considerable interest for optical communications. It is known that the diffraction losses in such an optical beam waveguide can be kept very small for moderate size lenses.[1,2] This means that if a transmission line is made of identical low-loss lenses, spaced identically along a straight line with each lens centered on this straight line, there is a mode of propagation which is low loss. However, if there are imperfections in the transmission line, the light beam will begin to wander from the axis or the beam size will grow and the beam will eventually strike the edge of the lens and be lost. Since the diffraction loss of the beam in a perfect line can be kept very small, it is the line imperfections, the line axis curvature, and the scattering and absorption at each lens which will primarily determine the

optical loss. Gas lenses have been considered for reducing the scattering and absorption losses.[3]

Rowe[4] and Hirano and Fukatsu[5] have shown how the beam position is affected by random lateral lens displacements. Berreman,[6] Marcuse,[7] and Unger[8] have considered correlated lateral lens displacements in the form of bends. All of these analyses have assumed perfect lenses and perfect spacing and have shown the growth of the beam displacement due to lateral lens displacements only. It is the purpose of this paper to show how the previously obtained results are altered when the lens focal lengths and lens spacings have random variations.

In this paper, we shall consider the statistical effect of random variations in lens focal length and lens spacing and random lateral lens displacements. We shall also consider random bends whose correlation length is much smaller than the total line length. It is shown how the various line imperfections couple to one another and cause the beam deviation from the axis to grow. The growth of the spot size of a Gaussian beam is also considered.

It is shown here that the random variations in f and L cause the rms expected value of beam displacement and the rms expected value of the beam spot size to grow exponentially with distance when the number of lenses is large. For bends, the variations in f and L cause an exponential increase in the average allowed bending radius of the guide when the number of lenses is large. In contrast, when f and L are perfect these effects grow more slowly with distance, and increase only as the square root of the number of lenses.

We shall use geometric optics since it is known that in the paraxial approximation the center of a Gaussian beam in a beam waveguide behaves as a ray.[9] The geometric optics analysis is extended to find the behavior of the beam spot size by replacing the Gaussian beam by its equivalent ray packet.[10]

II. GENERAL PROBLEM FORMULATION

We shall consider the problem in two dimensions only for simplicity. It has been shown that the three-dimensional problem can be split into two independent two-dimensional problems.[5] For aberration-free lenses, the motion of the beam in one transverse dimension is dependent only on the initial conditions and lens displacements in the same transverse direction.

Consider the transmission line shown in Fig. 1. We define r_n and r_n' as the position and slope of the ray just to the right of the nth lens. The

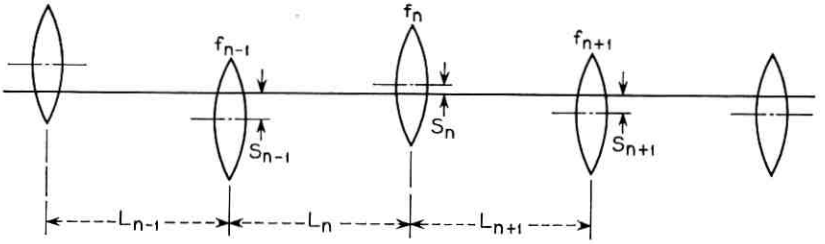


Fig. 1 — Beam waveguide notation.

ray position and slope are measured with respect to a straight line which is the nominal transmission line axis. The spacing between lenses is labeled L_n and the convergence of the lenses as C_n where

$$C_n = 1/f_n$$

and f_n is the focal length of the nth lens. The lateral distance between the center of the nth lens and the reference line is s_n . The displacement s_n is positive if the lens center is above the reference line.

Using this notation we can write

$$r_n = r_{n-1} + L_n r_{n-1}' \tag{1}$$

$$r_n' = -C_n r_{n-1} + (1-C_n L_n) r_{n-1}' + C_n s_n . \tag{2}$$

If we define

$$R_n = \begin{bmatrix} r_n \\ r_n' \end{bmatrix},$$

$$M_n = \begin{bmatrix} 1 & L_n \\ -C_n & 1 - L_n C_n \end{bmatrix},$$

and

$$V_n = \begin{bmatrix} 0 \\ C_n s_n \end{bmatrix}$$

we can write (1) and (2) in matrix form as

$$R_n = M_n R_{n-1} + V_n . \tag{3}$$

This relates the ray position and slope at the n plane to the ray position

and slope at the $n-1$ plane. We shall be interested in the rms expected value of the output beam displacement and hence in the square of the output beam slope and position. We shall, therefore, square the matrix (3). To do this, we take the Kronecker product[11] of each side of (3) with itself.

$$R_n \times R_n = (M_n \times M_n)(R_{n-1} \times R_{n-1}) + V_n \times V_n + (M_n R_{n-1}) \times V_n + V_n \times (M_n R_{n-1}). \tag{4}$$

Now take the expected value of (4)

$$\langle R_n \times R_n \rangle = \langle (M_n \times M_n)(R_{n-1} \times R_{n-1}) \rangle + \langle V_n \times V_n \rangle + \langle (M_n R_{n-1}) \times V_n \rangle + \langle V_n \times (M_n R_{n-1}) \rangle. \tag{5}$$

The expected value of a matrix is the matrix of the expected values.

We will now state our statistical assumptions. We assume small variations about L and C

$$L_n = L(1 + l_n)$$

$$C_n = C(1 + c_n).$$

$$\langle l_n \rangle = \langle c_n \rangle = \langle s_n \rangle = 0$$

$$\langle l_n l_k \rangle = \langle c_n c_k \rangle = 0 \qquad n \neq k$$

$$\langle l_n c_k \rangle = \langle l_n s_k \rangle = \langle c_n s_k \rangle = 0 \quad \text{all} \quad n$$

$$\langle c_n^2 \rangle = \sigma_c^2$$

$$\langle l_n^2 \rangle = \sigma_L^2$$

$$\langle s_n^2 \rangle = y^2.$$

Hence, l_n, c_n, and s_n are mutually independent random variables of zero mean. There is no correlation between the variations in spacing, the variations in focal length, and the variations in lateral displacement. The variations in L and C are completely random with no correlation between adjacent L's and adjacent C's. For the first sections of the paper we shall consider the lateral lens displacements to also be completely random with no correlation between adjacent displacements. In a later section it will be shown how correlation of the lens displacements in the form of random bends in the waveguide axis can be included.

III. A TRANSMISSION LINE WITH RANDOM LATERAL LENS DISPLACEMENTS

In this section we shall consider only random lateral lens displacements. Hence, we impose the additional statistical restriction that

$$\langle s_k s_n \rangle = 0 \qquad n \neq k.$$

The lateral displacement of any lens is unaffected by the lateral displacements of any other lens.

If we repeatedly substitute (3) into the last two matrices of (5), we see that they contain elements of the general form

$$\langle G(L_k\,,\ L_{k+1}\,, \cdots L_n\,,\ C_k\,,\ C_{k+1}\,, \cdots C_n) s_{k-1} s_n \rangle, \qquad k \leqq n$$

and

$$\langle F(L_1\,, L_2\,, \cdots L_n\,, C_1\,, C_2\,, \cdots C_n) R_0 s_n \rangle$$

where $G(\ )$ and $F(\ )$ are some functions. In view of the statistical assumptions, these can be written as

$$\langle G(L_k\,,\ L_{k+1} \cdots L_n\,,\ C_k\,,\ C_{k+1}\,, \cdots C_n) \rangle \langle s_{k-1} s_n \rangle, \qquad k \leqq n$$

and

$$\langle F(L_1\,, L_2\,, \cdots L_n\,, C_1\,, C_2\,, \cdots C_n) \rangle R_0 \langle s_n \rangle$$

which are zero since $\langle s_{k-1} s_n \rangle = 0$ for $k \leqq n$ and $\langle s_n \rangle = 0$. Hence, the last two matrices of (5) are both zero.

When there is some correlation between lateral lens displacements, i.e., a wavy transmission line axis, these two matrices are not zero. It is through these matrices that the correlation will enter.

Also, by using the above statistical assumptions M_n is independent of R_{n-1} and $\langle M_n \times M_n \rangle$ and $\langle V_n \times V_n \rangle$ are not functions of n. Equation (5) can, therefore, be written as

$$\langle R_n \times R_n \rangle = \langle M_n \times M_n \rangle^n R_0 \times R_0 + \sum_{k=0}^{n-1} \langle M_n \times M_n \rangle^k \langle V_n \times V_n \rangle \quad (6)$$

where R_0 is the matrix of the initial ray slope and position.

The Kronecker products in (6) are either 4×4 or 4×1 matrices. These can be reduced to 3×3 and 3×1 matrices by combining the two redundant terms.[11] For clarity and ease of computation we shall reduce the matrices and write them out explicitly below. We have assumed here that $\sigma_L^2 \ll 1$ and $\sigma_C^2 \ll 1$ hence, we have neglected $\sigma_L^2 \sigma_C^2$ compared to σ_L^2 or σ_C^2.

$$\langle R_n \times R_n \rangle = \begin{bmatrix} \langle r_n^2 \rangle \\ \langle r_n r_n' \rangle \\ \langle r_n'^2 \rangle \end{bmatrix}$$

$$R_0 \times R_0 = \begin{bmatrix} r_0^2 \\ r_0 r_0' \\ r_0'^2 \end{bmatrix}$$

$$\langle V_n \times V_n \rangle = \begin{bmatrix} 0 \\ 0 \\ C^2(1 + \sigma_C^2)y^2 \end{bmatrix}$$

$$\langle M_n \times M_n \rangle = \begin{bmatrix} 1 & 2L & L^2(1 + \sigma_L^2) \\ -C & 1 - 2LC & L - L^2C(1 + \sigma_L^2) \\ C^2(1 + \sigma_C^2) & -2C + 2LC^2 & 1 - 2LC + L^2C^2 \\ & (1 + \sigma_C^2) & (1 + \sigma_L^2 + \sigma_C^2) \end{bmatrix}.$$

3.1 *The Characteristic Roots of* $\langle M_n \times M_n \rangle$

To evaluate (6) will require the raising of $\langle M_n \times M_n \rangle$ to the kth power. To do this it will be necessary to find the characteristic roots of $\langle M_n \times M_n \rangle$. The characteristic roots of $\langle M_n \times M_n \rangle$ can be found from the equation

$$| \langle M_n \times M_n \rangle - I\lambda | = 0$$

where λ is the characteristic root and I is the unity matrix. This leads to the following cubic equation for λ:

$$\begin{aligned} \lambda^3 - \lambda^2[3 - 4LC + L^2C^2 + L^2C^2(\sigma_L^2 + \sigma_C^2)] \\ - \lambda[-3 + 4LC - L^2C^2 + L^2C^2(\sigma_L^2 + \sigma_C^2)] - 1 = 0. \end{aligned} \tag{7}$$

We assume in (7) that σ_L^2 and σ_C^2 are very small quantities, hence terms of higher power than 2 in σ_L and σ_C are neglected. Since σ_L^2 and σ_C^2 are assumed very small it is reasonable to assume that the roots

of (7) are very near the roots for the perfect transmission when $\sigma_L^2 = \sigma_C^2 = 0$. For the perfect transmission line, the roots are

$$\lambda = 1, \quad e^{2i\theta}, \quad e^{-2i\theta},$$

where $\theta = \cos^{-1}(1 - LC/2)$. We therefore write the three roots of (7) as

$$\lambda_1 = 1 + q_1$$
$$\lambda_2 = e^{2i\theta}(1 + q_2)$$
$$\lambda_3 = e^{-2i\theta}(1 + q_3)$$

where

$$|q_1|, |q_2|, |q_3| \ll 1.$$

For the case of $LC \neq 2$, i.e., a nonconfocal system, (7) gives

$$q_1 = \frac{2LC}{4 - LC}(\sigma_L^2 + \sigma_C^2)$$

$$q_2 - q_3 = \frac{-LC}{4 - LC}(\sigma_L^2 + \sigma_C^2).$$

If we define

$$a = \frac{2LC}{4 - LC}(\sigma_L^2 + \sigma_C^2)$$

then

$$\lambda_1 = 1 + a$$
$$\lambda_2 = e^{2i\theta}\left(1 - \frac{a}{2}\right)$$
$$\lambda_3 = e^{-2i\theta}\left(1 - \frac{a}{2}\right).$$

For the confocal case, $LC = 2$, the roots of (7) are

$$\lambda_1 = 1 + a$$
$$\lambda_2 = -1 + a$$
$$\lambda_3 = -1.$$

These two sets of solutions of (7) are valid so long as $a \ll 1$. This means $\sigma_C^2 \ll 1$, $\sigma_L^2 \ll 1$, and LC is not near 4.

3.2 *Sylvesters Theorem*

For raising the matrix $\langle M_n \times M_n \rangle$ to the kth power, it is helpful to use Sylvesters Theorem.[12] If λ_1, λ_2, and λ_3 are the characteristic roots of the matrix A then

$$A^k = \frac{(A - \lambda_2 I)(A - \lambda_3 I)}{(\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3)} \lambda_1{}^k + \frac{(A - \lambda_1 I)(A - \lambda_3 I)}{(\lambda_2 - \lambda_1)(\lambda_2 - \lambda_3)} \lambda_2{}^k + \frac{(A - \lambda_1 I)(A - \lambda_2 I)}{(\lambda_3 - \lambda_1)(\lambda_3 - \lambda_2)} \lambda_3{}^k$$

where I is the unity matrix.

In (6) we shall be interested in the case where n is a very large number, i.e., many lenses in the transmission line. The difference between the matrix $\langle M_n \times M_n \rangle$ and $M \times M$ where M is the perfect transmission line matrix is very small (terms of the order of $\sigma_C{}^2$ and $\sigma_L{}^2$). For the nonconfocal case, it will therefore only be necessary to consider the deviations from a perfect transmission line in $\lambda_1{}^k$, $\lambda_2{}^k$, and $\lambda_3{}^k$. In the coefficients of $\lambda_1{}^k$, $\lambda_2{}^k$, and $\lambda_3{}^k$ we can assume $\sigma_L{}^2 = \sigma_C{}^2 = 0$.

Hence, for the nonconfocal case, we can write

$$\langle M_n \times M_n \rangle^k = \frac{(M \times M - e^{2i\theta} I)(M \times M - e^{-2i\theta} I)}{(1 - e^{2i\theta})(1 - e^{-2i\theta})} e^{ka} + \frac{(M \times M - I)(M \times M - e^{-2i\theta} I)}{(e^{2i\theta} - 1)(e^{2i\theta} - e^{-2i\theta})} e^{2ik\theta} e^{-ka/2} + \frac{(M \times M - I)(M \times M - e^{2i\theta} I)}{(e^{2i\theta} - 1)(e^{-2i\theta} - e^{2i\theta})} e^{-2ik\theta} e^{-ka/2}. \quad (8)$$

Here we have used 1, $e^{2i\theta}$, and $e^{-2i\theta}$ for the roots of $M \times M$ and

$$\lambda_1{}^k = (1 + a)^k \approx e^{ka}$$

$$\lambda_2{}^k = e^{2ik\theta}\left(1 - \frac{a}{2}\right)^k \approx e^{2ik\theta} e^{-ka/2}$$

$$\lambda_3{}^k = e^{-2ik\theta}\left(1 - \frac{a}{2}\right)^k \approx e^{-2ik\theta} e^{-ka/2}.$$

In (6) we shall be interested in only the first element of $\langle R_n \times R_n \rangle$ which is $\langle r_n{}^2 \rangle$. To calculate this we will only need the first row of elements of $\langle M_n \times M_n \rangle^k$. These can be found from (8). The 1,1 element for

the nonconfocal case is

$$b_{11} = e^{ka}\left[\frac{4\sin^2\theta - 2LC\cos\theta}{4\sin^2\theta}\right] + e^{2ik\theta}e^{-ka/2}\left[\frac{-2LC\cos\theta}{(e^{2i\theta}-1)2j\sin 2\theta}\right] + e^{-2ik\theta}e^{-ka/2}\left[\frac{-2LC\cos\theta}{(e^{-2i\theta}-1)(-2j\sin 2\theta)}\right].$$

After some simplification

$$b_{11} = \frac{2}{4 - LC}\left[e^{ka} + e^{-ka/2}\cos(2k-1)\theta\right].$$

Similarly, we can write b_{12} and b_{13} for the nonconfocal case as

$$b_{12} = \frac{2L}{4 - LC}\left[e^{ka} + e^{-ka/2}\left(\sqrt{\frac{4 - LC}{LC}}\sin 2k\theta - \cos 2k\theta\right)\right]$$

$$b_{13} = \frac{2L^2}{LC(4 - LC)}\left[e^{ka} - e^{-ka/2}\cos 2k\theta\right].$$

For the confocal case, $LC = 2$, we are close to a degenerate case where two characteristic roots are close to being equal. If we retain terms to no higher power than 2 in ${\sigma_L}^2$ and ${\sigma_C}^2$ in $\langle M_n \times M_n\rangle^k$ we can write

$$\langle M_n \times M_n\rangle^k = \frac{(M \times M + I)(M \times M + I)}{4}e^{ka} - \frac{(\langle M_n \times M_n\rangle - (1+a)I)(\langle M_n \times M_n\rangle + I)}{2a}(-1)^k e^{-ka} \quad (9)$$

$$+ \frac{(\langle M_n \times M_n\rangle - (1+a)I)(\langle M_n \times M_n\rangle + (1-a)I)}{2a}(-1)^k.$$

We have retained the ${\sigma_L}^2$ and ${\sigma_C}^2$ in the coefficients of the last two terms since ${\sigma_L}^2$ and ${\sigma_C}^2$ appear in the denominators.

We again are interested only in the first row of $\langle M_n \times M_n\rangle^k$. For the confocal case, the elements of the first row are

$$b_{11} = e^{ka}$$

$$b_{12} = L\left[e^{ka} - \frac{2(-1)^k}{a}({\sigma_C}^2 e^{-ka} + {\sigma_L}^2)\right]$$

$$b_{13} = \frac{L^2}{2}\left[e^{ka} - \frac{2(-1)^k}{a}({\sigma_C}^2 e^{-ka} + {\sigma_L}^2)\right].$$

3.3 *Calculation of* $\langle r_n^2 \rangle$

For the nonconfocal case using (6) and the values for the elements in the first row of $\langle M_n \times M_n \rangle^k$, we can calculate $\langle r_n^2 \rangle$.

$$\begin{aligned}\langle r_n^2 \rangle = \frac{2}{4 - LC} \Bigg\{ & r_0^2 [e^{na} + e^{-(na/2)} \cos (2n - 1)\theta] \\ & + r_0 L r_0' \left[e^{na} + e^{-(na/2)} \left(\sqrt{\frac{4 - LC}{LC}} \sin 2n\theta - \cos 2n\theta \right) \right] \\ & + \frac{L^2 r_0'^2}{LC} [e^{na} - e^{-(na/2)} \cos 2n\theta] \\ & + LCy^2 \sum_{k=0}^{n-1} (e^{ka} - e^{-(ka/2)} \cos 2k\theta) \Bigg\}.\end{aligned}$$

In the 3,1 position of $\langle V_n \times V_n \rangle$ we have neglected the σ_c^2 as compared to 1.

The summations can be evaluated as

$$\sum_{k=0}^{n-1} e^{ka} = \frac{e^{na} - 1}{e^a - 1} \approx \frac{e^{na} - 1}{a}$$

$$\begin{aligned}\sum_{k=0}^{n-1} & e^{-(ka/2)} \cos 2k\theta \\ & = \frac{1 - e^{-(a/2)} \cos 2\theta - e^{-(na/2)} \cos 2n\theta + e^{-[(n+1)a/2]} \cos 2(n - 1)\theta}{1 + e^{-a} - 2e^{-a/2} \cos 2\theta} \\ & \approx \frac{1}{2} + \frac{e^{-(na/2)} \sin (2n - 1)\theta}{2 \sin \theta}.\end{aligned}$$

We have used the facts that $a \ll 1$ and $n \gg 1$.

The expected value of the square of the output beam position for the nonconfocal case is therefore,

$$\begin{aligned}\langle r_n^2 \rangle = \frac{2}{4 - LC} \Bigg\{ & r_0^2 [e^{na} + e^{-(na/2)} \cos (2n - 1)\theta] \\ & + r_0 L r_0' \left[e^{na} + e^{-(na/2)} \left(\sqrt{\frac{4 - LC}{LC}} \sin 2n\theta - \cos 2n\theta \right) \right] \\ & + \frac{L^2 r_0'^2}{LC} [e^{na} - e^{-(na/2)} \cos 2n\theta] \\ & + LCy^2 \left[\frac{e^{na} - 1}{a} - \frac{1}{2} - \frac{e^{-(na/2)} \sin (2n - 1)\theta}{2 \sin \theta} \right] \Bigg\}\end{aligned} \tag{10}$$

where

$$a = \frac{2LC}{4 - LC}(\sigma_L^2 + \sigma_C^2).$$

For confocal spacing, $LC = 2$, the square of the expected value of r_n is

$$\begin{aligned}\langle r_n^2 \rangle = {} & r_0^2 e^{na} + r_0 L r_0' \left[e^{na} - \frac{2(-1)^n}{a}(\sigma_C^2 e^{-na} + \sigma_L^2) \right] \\ & + \frac{L^2 r_0'^2}{2} \left[e^{na} - \frac{2(-1)^n}{a}(\sigma_C^2 e^{-na} + \sigma_L^2) \right] \\ & + 2y^2 \left[\frac{e^{na} - 1}{a} - \frac{1}{2} - \frac{\sigma_C^2 e^{-na}}{a} + \frac{(-1)^n \sigma_L^2}{a} \right]\end{aligned} \tag{11}$$

where, for $LC = 2$, $a = 2(\sigma_L^2 + \sigma_C^2)$.

It is of interest to consider how close to $LC = 2$ one must be to have (11) hold rather than (10). If we retain terms to only the first power in a, it can be shown that (11) is valid when

$$| LC - 2 | < \sigma_L^2 + \sigma_C^2.$$

If

$$| LC - 2 | > \sigma_L^2 + \sigma_C^2$$

then (10) holds. Since σ_L^2 and σ_C^2 will be of the order of 10^{-4} we must be very close to $LC = 2$ for (11) to hold.

If the lenses and spacing are perfect so that $a = 0$, the first three terms of (10) and of (11) give the square of the output beam position due to the input beam slope and position. The last term gives the increased displacement due to random lateral lens displacements. Both parts agree with Hirano and Fukatsu[5] when $a = 0$.

The random errors in focal length and spacing cause an exponential increase in the expected value of the square of the output beam displacement. This can be seen more clearly for the case where n is very large. If $na > 2$, then (10) and (11) reduce to the same result. In this case,

$$\langle r_n^2 \rangle \approx \frac{2e^{na}}{4 - LC}\left[r_0^2 + r_0 L r_0' + \frac{L^2 r_0'^2}{LC} \right] + \frac{2LC}{4 - LC} y^2 \left[\frac{e^{na} - 1}{a} \right]. \tag{12}$$

IV. TRANSMISSION LINE WITH RANDOM BENDS

We will now assume the axis of the beam waveguide is bent so that there is some correlation between adjacent lateral lens displacements.

As noted in Section III this correlation will appear in (5) in the last two terms, $\langle (M_n R_{n-1}) \times V_n \rangle$ and $\langle V_n \times (M_n R_{n-1}) \rangle$.

It is shown in the Appendix that these two terms to first order do not contain σ_L^2 or σ_C^2, i.e., they are not affected by the random variations in focal length and spacing. It is also shown that if the axis of the guide is composed of a series of uncorrelated bends whose average bend length is much smaller than the total length of the transmission line, these two terms are not functions of n. This type of bending might typically be the case for a very long transmission line laid to follow the gentle bends of the terrain.

It was shown in Section III that to first order $\langle V_n \times V_n \rangle$ also has these properties, i.e., it is not a function of σ_L^2, σ_C^2, or n. Because of this similarity, the matrix $\langle (M_n R_{n-1}) \times V_n \rangle + \langle V_n \times (M_n R_{n-1}) \rangle$ can be considered as an added part of $\langle V_n \times V_n \rangle$ and can be carried through the analysis in this manner. Hence, the random errors in focal length and spacing affect the beam displacement due to short uncorrelated bends in the same way they affected the beam displacement due to random lens displacements.

From (10) or (11) for "a" small we can show how σ_L^2 and σ_C^2 couple to the random displacements by writing

$$\langle r_n^2 \rangle = \langle r_n^2 \rangle_{a=0} \frac{e^{na} - 1}{na}.$$

In this expression, $\langle r_n^2 \rangle_{a=0}$ is the expected value of the square of the beam displacement due to random lens displacements when $a = 0$. Because of the similarity pointed out above, the beam displacement due to short random bends is also multiplied by $(e^{na} - 1)/na$ to account for the focal length and spacing errors. Let us assume a transmission line axis is specified which fits the conditions, i.e., it is composed of a series of uncorrelated bends whose bend length is much shorter than the total line length. From this we can calculate $\langle r_n^2 \rangle$ assuming L and C are perfect. This has been done for some specific cases by Marcuse,[7] Berreman,[6] and Unger.[8] To include random imperfections of L and C if a is small we multiply this result by

$$\frac{e^{na} - 1}{na}.$$

This analysis does not hold if the correlation extends along the entire line (for example a serpentine bend) or if the correlation extends over a large portion of the line.

V. STATISTICAL GROWTH OF BEAM SPOT SIZE

We have been concerned thus far with the behavior of light rays in an imperfect transmission line. Our primary concern, however, is the behavior of Gaussian light beams rather than light rays. It has previously been shown that in the paraxial approximation the center of a Gaussian light beam does behave like a ray.[9] We can regard r_n and r_n' therefore, as the position and slope of the center of a Gaussian beam at lens n and r_0 and r_0' as the initial conditions of the beam center. We lack information on the effects of the transmission line imperfections on the size of the beam.

Using the complex beam parameter law of Kogelnik[13] it would be possible to find the statistical growth of the spot size due to the line imperfections. It will be simpler, however, to use Steier's ray packet equivalent to the Gaussian beam.[10] This approach conveniently uses the already derived statistical behavior of the light rays to find the beam size behavior.

Just to the right of a lens, the ray packet equivalent of the normal Gaussian mode of the transmission line is

$$\left.\begin{aligned} r_0 &= w_0 \cos\varphi + \frac{L}{kw_0}\sin\varphi \\ r_0' &= \frac{-2}{kw_0}\sin\varphi \end{aligned}\right\} \varphi \text{ has all values from 0 to } 2\pi \qquad (13)$$

where

$$w_0 = \text{spot size at the beam waist} = \left[\frac{L(4-LC)}{k^2C}\right]^{\frac{1}{2}}$$

$$k = \frac{2\pi}{\lambda}$$

$$\lambda = \text{wavelength.}$$

If the path through the transmission line of each ray of the packet is found then the behavior of the equivalent Gaussian mode through the transmission line can be found. At any point in the transmission line, the envelope or the distance between the extreme rays of the ray packet is equal to twice the beam spot size and the curves which are perpendicular to the average ray slope are the beam phase fronts.

To find the effect of transmission line imperfections on the beam spot size, let us launch the ray packet given by (13) into the trans-

mission line. If we substitute the value for r_0 and r_0' from (13) into (10) for the nonconfocal case we find

$$\langle r_n^2 \rangle = \frac{2w_0^2}{4 - LC} e^{na} + \frac{2w_0^2}{4 - LC} e^{-(na/2)} [\cos 2\varphi \cos (2n - 1)\theta - \sin 2\varphi \sin (2n - 1)\theta] \tag{14}$$

where φ ranges from 0 to 2π. We have not included the last term of (10) since the lateral lens displacements have no effect on the growth of the spot size.

The expected value of the square of the spot size at the nth lens, $\langle w_n^2 \rangle$, is given by the envelope of these rays. Taking the maximum value of (14) as φ goes from 0 to 2π.

$$\langle w_n^2 \rangle = \frac{2w_0^2}{4 - LC} (e^{na} + e^{-(na/2)}).$$

The normal mode spot size at a lens, w, is given by

$$w^2 = \frac{4w_0^2}{4 - LC}.$$

Hence, for the nonconfocal case

$$\frac{\langle w_n^2 \rangle}{w^2} = \frac{e^{na} + e^{-(na/2)}}{2} \tag{15}$$

where

$$a = \frac{2LC}{4 - LC} (\sigma_L^2 + \sigma_C^2).$$

For the confocal case, $LC = 2$, we substitute from (13) into (11). Taking the maximum value as φ ranges from 0 to 2π.

$$\frac{\langle w_n^2 \rangle}{w^2} = \frac{1}{2} \left\{ e^{na} + \frac{\sigma_C^2 e^{-na} + \sigma_L^2}{\sigma_C^2 + \sigma_L^2} \right\}. \tag{16}$$

If na is small so that

$$e^{na} \approx 1 + na,$$

then for the nonconfocal case (15) becomes

$$\frac{\langle w_n^2 \rangle}{w^2} = 1 + \frac{LCn}{2(4 - LC)} (\sigma_L^2 + \sigma_C^2), \tag{17}$$

and for the confocal case (16) becomes

$$\frac{\langle w_n^2 \rangle}{w^2} = 1 + n\sigma_L^2. \tag{18}$$

Equations (17) and (18) agree with the results of the perturbation analysis of Hirano and Fukatsu.[5]

Hence, for na small, the random errors in C do not affect the spot size in the confocal case. However, as pointed out in Section 3.3, LC must be almost 2 for this to be true. If $|LC - 2| > \sigma_L^2 + \sigma_C^2$ the result for a nonconfocal system should be used. Since $\sigma_L^2, \sigma_C^2 \approx 10^{-4}$, this is a very stringent requirement on LC. It is doubtful if LC can be held close enough to 2 to gain this advantage in reduced spot size growth.

If $na > 2$ then the nonconfocal and the confocal results are very nearly the same and for both cases

$$\frac{\langle w_n^2 \rangle}{w^2} \approx \frac{1}{2} e^{na}.$$

VI. SUMMARY

The results derived here show statistically how imperfections in an optical transmission line affect the output beam from the transmission line. The imperfections cause the beam center to wander from the transmission line axis and cause the beam size to grow. We have considered the errors in focal length and spacing to be random and the lateral lens displacements to be random or with short correlation lengths. For this case, statistically the size of the beam and the distance of the beam center from the axis grow exponentially as the number of lenses when there are many lenses.

For $na > 2$, and random lateral lens displacements, the results can be summarized as follows. The beam center launched at r_0, r_0' has an rms expected value of

$$\sqrt{\langle r_n^2 \rangle} = \sqrt{\frac{2}{4 - LC}} \left[e^{na} \left(r_0^2 + r_0 L r_0' + \frac{L^2 r_0'^2}{LC} \right) + LCy^2 \frac{e^{na} - 1}{a} \right]^{\frac{1}{2}}.$$

The beam spot size has an rms expected value of

$$\sqrt{\langle w_n^2 \rangle} = e^{na/2} \frac{w}{\sqrt{2}}.$$

For random transmission line bends of short correlation length the random variations in L and C increase the expected value of the square

of the output beam displacement as

$$\langle r_n^2 \rangle = \frac{e^{na} - 1}{na} \langle r_n^2 \rangle_{a=0}$$

where $\langle r_n^2 \rangle_{a=0}$ is the value computed assuming no variations in L and C.

If a beam is launched into a straight line on axis with no slope its position is not affected by random errors in f and L and only the size of the beam is affected. This is obviously true since a ray through the center of a lens does not bend no matter what the lens focal length. If, however, the axis is curved or the lenses have random lateral displacements, the beam begins to wander from the axis and is now affected by the errors in f and L. This coupling is clearly shown in these results.

These calculations are pertinent when n is large. This is the case of a transmission line with relatively closely spaced lenses which would be able to control the light beam around gentle bends in the terrain.

As an example, let us consider a confocal beam waveguide with lenses spaced every one meter and built to the following rms expected value tolerances:

(*i*) focal length variations — 1 per cent
(*ii*) spacing variations — much less than 1 per cent
(*iii*) random lateral lens displacements — 2×10^{-2} mm.

These tolerances give

$$a = 2 \times 10^{-4}$$

$$y^2 = 4 \times 10^{-10} \text{ m}^2.$$

If we assume an rms output beam deviation of 2 millimeters is acceptable, we can go approximately 3.5 kilometers ($n = 3.5 \times 10^3$) with the line described above. If the line is allowed to have gentle circular bends of an average radius of curvature of 5 kilometers and an average bend length[7] of 100 meters then the distance which can be traveled before there is an rms beam deviation of 2 millimeters is reduced to 2.5 kilometers ($n = 2.5 \times 10^3$). This means a beam redirector is required every 2.5 kilometers.

In the above example, the dominant Gaussian mode spot size at each lens for the perfect line is 0.45 mm rad. Because of the line imperfections, this grows to an rms expected value of 0.48 mm rad at $n = 2.5 \times 10^3$. This spot size growth is insignificant compared to the beam deflection. In general, unless LC is very small the spot size growth is not as important as the beam deflection growth for closely spaced lenses.

In calculating these numbers we have assumed the lenses are perfect

and have neglected any aberrations. Additional work is required to determine the effects of aberrations on these results.

These numbers were calculated at $LC = 2$. If we make LC smaller the effect of the random lens displacements becomes less but the effect of correlated bends becomes larger.[7] At very small LC, the effect of spot size growth becomes important. If we increase LC the effect of correlated bends is reduced[7] but the effect of random lens displacements is increased. Clearly there is an optimum LC depending on line construction tolerances and line laying tolerances. It appears this optimum is near $LC = 2$ in a typical case.

Fig. 2 shows rms expected beam deviation as a function of n for a confocal line. This clearly shows how the distance between redirectors must be reduced as the errors in f and L become larger.

VII. ACKNOWLEDGMENTS

This problem and the use of Kronecker products to solve it was suggested by H. E. Rowe. His suggestions and comments were very helpful. Helpful suggestions by D. Marcuse are also acknowledged.

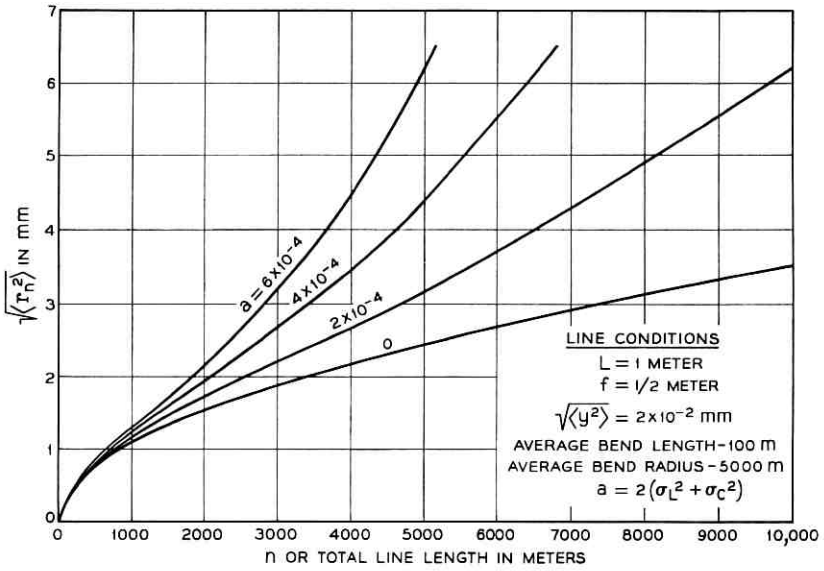


Fig. 2 — Expected value of the beam deflection as a function of transmission line length.

APPENDIX

Analysis of Matrices for Bent Waveguide Axis

We are interested in the last two terms of (5) since they contain the correlation between lateral lens displacements. These terms are

$$\langle (M_n R_{n-1}) \times V_n \rangle + \langle V_n \times (M_n R_{n-1}) \rangle.$$

For simplicity let us consider only $\langle (M_n R_{n-1}) \times V_n \rangle$, since the two terms are very similar. By repeated substitution of

$$R_n = M_n R_{n-1} + V_n$$

into

$$(M_n R_{n-1}) \times V_n ,$$

we find

$$(M_n R_{n-1}) \times V_n = (M_n P_1 R_0) \times V_n + \sum_{k=2}^{n} (M_n P_k V_{k-1}) \times V_n \quad (19)$$

where

$$P_k = M_{n-1} M_{n-2} M_{n-3} \cdots M_{k+1} M_k .$$

If we take the expected value of (19), the first term on the right side is zero since it is the product of independent terms and $\langle V_n \rangle = 0$.

We must look at the individual elements of $(M_n P_k V_{k-1}) \times V_n$. We can write the matrix P_k as

$$P_k = \begin{bmatrix} p_{k(1,1)} & p_{k(1,2)} \\ p_{k(2,1)} & p_{k(2,2)} \end{bmatrix}.$$

The p_k contain L_{n-1}, L_{n-2}, $\cdots$, L_k, and C_{n-1}, C_{n-2}, $\cdots$, C_k but only to the first power.

$$M_n P_k = \begin{bmatrix} p_{k(1,1)} + L_n p_{k(2,1)} & p_{k(1,2)} + L_n p_{k(2,2)} \\ -C_n p_{k(1,1)} + p_{k(2,1)} (1 - L_n C_n) & -C_n p_{k(1,2)} + p_{k(2,2)} (1 - L_n C_n) \end{bmatrix}$$

and

$$(M_n P_k V_{k-1}) \times V_n = \begin{bmatrix} 0 \\ p_{k(1,2)} C_n + p_{k(2,2)} L_n C_n \\ -p_{k(1,2)} C_n^2 + p_{k(2,2)} (C_n + L_n C_n^2) \end{bmatrix} C_{k-1} s_{k-1} s_n .$$

From this result we can write the last two terms in (5) as

$$\langle (M_nR_{n-1}) \times V_n \rangle + \langle V_n \times (M_nR_{n-1}) \rangle$$

$$= \sum_{k=2}^{n} \begin{bmatrix} 0 \\ \langle p_{k(1,2)} \rangle C + LC\langle p_{k(2,2)} \rangle \\ \langle p_{k(1,2)} \rangle C + LC\langle p_{k(2,2)} \rangle \\ 2C^2(1 + \sigma_C{}^2)(L\langle p_{k(2,2)} \rangle - \langle p_{k(1,2)} \rangle) \\ + 2\langle p_{k(2,2)} \rangle C \end{bmatrix} C\langle s_{k-1}s_n \rangle. \quad (20)$$

Since the p_k contain the L's and C's only to the first power in each, $\langle p_k \rangle$ will contain only L and C and will not contain $\sigma_L{}^2$ and $\sigma_C{}^2$. If we neglect $\sigma_C{}^2$ as compared to 1 (the same approximation is used in $\langle V_n \times V_n \rangle$), then $\langle (M_nR_{n-1}) \times V_n \rangle + \langle V_n \times (M_nR_{n-1}) \rangle$ does not contain $\sigma_L{}^2$ or $\sigma_C{}^2$.

We will now find under what conditions $\langle (M_nR_{n-1}) \times V_n \rangle + \langle V_n \times (M_nR_{n-1}) \rangle$ is independent of n.

The n dependence of (20) is in the terms

$$\sum_{k=2}^{n} \langle p_{k(1,2)} \rangle \langle s_{k-1}\ s_n \rangle$$

and

$$\sum_{k=2}^{n} \langle p_{k(2,2)} \rangle \langle s_{k-1}\ s_n \rangle.$$

Since

$$\langle P_k \rangle = \langle M \rangle^{n-k}$$

$$\langle p_{k(1,2)} \rangle = \frac{L \sin\ (n - k)\ \theta}{\sin \theta}$$

$$\langle p_{k(2,2)} \rangle = -\frac{L^2C \sin\ (n - k)\ \theta}{2 \sin \theta} + L \cos\ (n - k)\ \theta.$$

And we can write

$$\frac{\sin \theta}{L} \sum_{k=2}^{n} \langle p_{k(1,2)} \rangle \langle s_{k-1}\ s_n \rangle = \langle s_1s_n \rangle \sin\ (n - 2)\ \theta$$

$$+ \langle s_2s_n \rangle \sin\ (n - 3)\ \theta \cdots + \langle s_{n-2}s_n \rangle \sin \theta + \langle s_{n-1}s_n \rangle 0.$$

We will assume, as did Marcuse,[6] that

$$\langle s_ks_n \rangle = f(n - k).$$

The correlation depends on the distance between the lenses. We also assume that

$$f(n - k) = 0 \quad \text{for} \quad n - k > N.$$

That is, the correlation length is finite and extends only N lenses away. This means the waveguide axis is a series of random bends, the "average length" of each bend is NL. Therefore, if $n > N$ we can write

$$\frac{\sin \theta}{L} \sum_{k=2}^{n} \langle p_{k(1,2)} \rangle \langle s_{k-1}\, s_n \rangle = f(2) \sin \theta + f(3) \sin 2\theta$$

$$\cdots f\,(N - 1) \sin\,(N - 2)\theta + f(N) \sin\,(N - 1)\theta,$$

which is not a function of n. We can write a similar equation for

$$\sum_{k=2}^{n} \langle p_{k(2,2)} \rangle \langle s_{k-1}\, s_n \rangle.$$

However, we must consider all n down to 1. For these small n, $\langle (M_n R_{n-1}) \times V_n \rangle + \langle V_n \times (M_n R_{n-1}) \rangle$ will be a function of n. This means that all bends contribute the same to the output beam displacement except the initial bend which is within NL of the beginning of the transmission line. However, if we assume that $n \gg N$ (the average bend length is much smaller than the length of the transmission line) the contribution of this initial bend will be very small and can be neglected.

In summary, the conditions imposed on the transmission line axis are that it is composed of a series of random bends whose average length is NL. The average bend length is much smaller than the total length of the line. We have neglected the effect of any bend within NL of the beginning of the line. These are essentially the same conditions used by Marcuse,[7] Berreman,[6] and Unger[8] in their analyses of random bends of the transmission line.

Under these conditions $\langle (M_n R_{n-1}) \times V_n \rangle$ and $\langle V_n \times (M_n R_{n-1}) \rangle$ are independent of n.

REFERENCES

1. Boyd, G. D. and Gordon, J. P., Confocal Multimode Resonator for Millimeter Through Optical Wavelength Masers, B.S.T.J., *40*, March, 1961, pp. 489–508.
2. Goubau, G. and Schwering, F., On the Guided Propagation of Electromagnetic Wave Beams, Trans. IRE, *AP-9*, May, 1961, p. 248.
3. Berreman, D. W., A Lens or Light Guide Using Convectively Distorted Thermal Gradients in Gases, B.S.T.J., *43*, July, 1964, pp. 1469–1475.
4. Rowe, H. E., Private communication.
5. Hirano, J. and Fukatsu, Y., Stability of a Light Beam in a Beam Waveguide, Proc. IEEE, *52*, Nov. 1964, pp. 1284–1292.
6. Berreman, D. W., Growth of Oscillations of a Ray About the Irregularly Wavy Axis of a Lens Light Guide, B.S.T.J., *44*, Nov., 1965.

7. Marcuse, D., Statistical Treatment of Light-Ray Propagation in Beam-Waveguides, B.S.T.J., *44*, Nov., 1965.
8. Unger, H. G., Light Beam Propagation in Curved Schlieren Guides, Archiv der elektrischen Ubertragung, *19*, April, 1965, pp. 189–198.
9. Tien, P. K., Gordon, J. P., and Whinnery, J. R., Focusing of a Light Beam of Gaussian Field Distribution in Continuous and Periodic Lens-Like Media, Proc. IEEE, *53*, Feb., 1965, pp. 129–136. This result has also been derived independently by H. E. Rowe.
10. Steier, W. H., Ray Packet Equivalent to a Gaussian Light Beam, To be published.
11. Bellman, R., Introduction of Matrix Analysis, McGraw-Hill Book Company, Inc., New York, 1960, Chapter 12.
12. Korn, G. A. and Korn, T. M., Mathematics Handbook for Scientists and Engineers, McGraw-Hill Book Company, Inc., New York, 1961, Chapter 13.
13. Kogelnik, H., Imaging of Optical Modes — Resonators with Internal Lenses, B.S.T.J., *44*, March, 1965, pp. 455–494.

Signal-Noise Ratio Maximization Using the Pontryagin Maximum Principle

By J. M. HOLTZMAN

(Manuscript received November 26, 1965)

The applicability of the Pontryagin maximum principle to signal-noise ratio maximization is explored. Attention is focused on the reformulation of the problem so that the maximum principle may be used. The basic aspect of the reformulation is to cast the problem into the form of differential equations instead of integral equations.

Two problems are solved. The first, a variation of the matched filter problem, could have been solved by other methods. However, the maximum principle provided a very neat and systematic approach. The second problem, signal design with both an energy and an amplitude constraint imposed on the signal, is solved numerically. It appears to be intractable by other methods. One of the advantages of the maximum principle formulation is that, by working with differential equations rather than with integral equations, numerical techniques may be more easily used.

I. INTRODUCTION

The Pontryagin maximum principle may be considered to be a generalization of methods of calculus of variations that permits solution of optimization problems with inequality contraints. During the last few years, it has been extensively used to attack control theory problems. The use of the principle to solve signal optimization problems is introduced in Ref. 1. The maximum principle is briefly discussed in connection with wave-form optimization in Ref. 2.

The purpose of this present paper is to develop techniques for the application of the maximum principle, in particular to problems of signal-noise ratio maximization. We shall show how the maximum principle may be used to solve some problems with inequality constraints (e.g., the amplitude of a signal may be constrained to be less than or equal to some maximum value) which were heretofore considered intractable. We shall also show how a problem, solvable by other methods, may be

very conveniently attacked with the formalism of the maximum principle.

It is interesting to note that the maximum principle is, in a sense, more applicable to communication theory than to control theory for which it was originally developed (this is also pointed out in Ref. 1). The maximum principle yields a function of time to maximize a functional subject to constraints and for prescribed initial conditions. The answer to most communication theory problems is a function of time. On the other hand, in control problems, the function of time for specific initial conditions is called an "open loop" solution. What is actually needed is the "closed loop" or "feedback" solution which is a function of the present state. This is only indirectly determined using the maximum principle.

After introducing the maximum principle, we shall first solve a constrained matched filter problem and then a signal design problem. Attention will be focused on the reformulation of the problems so that the maximum principle is applicable.

II. THE MAXIMUM PRINCIPLE

We shall briefly discuss the maximum principle. Our discussion is an abstraction of some material in Ref. 3. Another excellent introduction to the maximum principle, which is presently available, is the Introduction and Chapter I of Ref. 4. Consider a system whose state is described by a vector $x = (x_1, x_2, \cdots x_n)$ which satisfies the differential equation

$$\dot{x} = f(x,u,t) \qquad t \in [t_0, t_1] \tag{1}$$

where $u = (u_1, \cdots, u_r)$ is an r-dimensional control vector and $f(x,u,t) = (f_1(x,u,t), f_2(x,u,t), \cdots, f_n(x,u,t))$ is a given n-dimensional vector valued function of x, u, and t. We require that

$$u(t) \in \Omega \qquad t \in [t_0, t_1] \tag{2}$$

where the set Ω is the set of admissible control vectors. Let F be the class of all piecewise continuous functions* from $[t_0, t_1]$ into Ω. If $\mathcal{u}$ is a control function in the class F we denote the trajectory corresponding to $\mathcal{u}$ by $x(t;\mathcal{u})$ which satisfies the following relations

$$\dot{x}(t;\mathcal{u}) = f(x(t;\mathcal{u}), u(t), t) \quad \text{a.e. } t \in [t_0, t_1] \tag{3}$$

$$x(t_0 ; \mathcal{u}) = 0. \tag{4}$$

* Corinthian script (e.g., $\mathcal{u}$,$\mathcal{v}$) is used to denote control functions. Small English letters (e.g., $u(t)$, $v(t)$) denote values of functions at specific times. The function $\mathcal{u}$ is the function whose value at time t is $u(t)$.

(There is no loss of generality in assuming zero initial conditions; a transformation of coordinates may be used for nonzero initial conditions.)

The optimization problem is as follows. Let $\{s_1, s_2, \cdots, s_m\}$ be a given set of real numbers where $0 \leqq m \leqq n-1$. We prescribe the final values (at $t = t_1$) of the first m coordinates of the state vector x to be

$$\{s_1, s_2, \cdots, s_m\}$$

and we require the final value of x_n to be maximum. The optimization problem is formally stated as follows: we are given the set

$$S = \{x: x_i = s_i \quad \text{for} \quad i = 1, \cdots, m\} \tag{5}$$

and we want to find a control function φ in the class F such that

(*i*) $x(t_1 ; \varphi) \in S$

(*ii*) for all $u \in F$ such that (6)

$$x(t_1 ; u) \in S \tag{7}$$

the following relation holds:

$$x_n(t_1 ; u) \leqq x_n(t_1 ; \varphi). \tag{8}$$

The control function φ is called the optimal control function and $x(t;\varphi)$ is the optimal trajectory.

The Pontryagin maximum principle is a necessary condition that an optimal control function must satisfy. To state the principle, we first define the Hamiltonian, $H(x,u,t,p)$,

$$H(x,u,t,p) = \langle f(x,u,t) \mid p \rangle \tag{9}$$

where $p = (p_1, \cdots, p_n)$ is an n-dimensional vector and $\langle a \mid b \rangle$ denotes the scalar product of a and b.

2.1 *Maximum Principle*

If φ is an optimal control function then there exists a nonidentically zero continuous vector valued function $p(t)$ such that

(*i*) $H(x(t;\varphi), v(t), t, p(t)) \geqq H(x(t;\varphi), u, t, p(t))$

for a.e $t \in [t_0, t_1]$ and all $u \in \Omega$, (10)

$$(ii)\ \dot{p}(t) = -\left[\frac{\partial f(x,v(t),t)}{\partial x}\right]^T_{x=x(t;\mathcal{U})} \cdot p(t) \tag{11}$$

for a.e. $t \in [t_0, t_1]$
(superscript T denotes transpose),

$$(iii)\ p_i(t_1) = 0 \qquad i = m+1, \cdots, n-1, \tag{12}$$

$$(iv)\ p_n(t_1) \geqq 0. \tag{13}$$

Relation (i) states that the Hamiltonian, evaluated along the optimal trajectory, takes on its maximum value with $v(t)$. Note that the maximization is over u, with $x(t;\mathcal{U})$ and $p(t)$ held fixed.

Relation (ii) may alternatively be expressed as follows:

$$\dot{p}_i(t) = -\sum_{j=1}^{n}\left[\frac{\partial^j f(x,v(t))}{\partial x_i}\right]_{x=x(t;\mathcal{U})} \cdot p_j(t), \qquad i = 1, 2, \cdots, n. \tag{14}$$

Relation (iii) states that the final value of an element of the vector $p(t)$ is zero if it corresponds to an element of the vector $x(t)$ which is left free at $t = t_1$.

Relation (iv) states that the nth element of $p(t)$, which corresponds to the element of $x(t)$ which is being maximized at $t = t_1$, is nonnegative at $t = t_1$.

Verification that the maximum principle is satisfied is seen to be equivalent to verification that a set of differential equations with mixed boundary values is satisfied. Conditions on $x(t;\mathcal{U})$ must be satisfied at both t_0 and t_1 and $p(t)$ must satisfy conditions at t_1. This boundary value problem is not always solvable analytically but much progress has been made in the numerical solution of such problems.

Communication theory problems are not usually stated in the form just described with differential equations. So the first order of business in applying the maximum principle to a communication theory problem is to convert it into the appropriate form with differential equations.

III. A MATCHED FILTER PROBLEM

To illustrate the formalism involved, we first solve a variation of the matched filter problem. The use of the maximum principle, in this case, is actually equivalent to using the classical calculus of variations. The basic matched filter problem is as follows (Ref. 5, p. 244). We have signal, $s_1(t)$, and noise, $n(t)$, entering a linear filter and we wish to design the linear filter so that the output signal-noise ratio is maximized at a specific time, t_1. This problem can be trivially solved using the maximum principle and by other methods. To make the problem a

little more interesting, suppose that we also specify that the output to a second signal input, $s_2(t)$, is to be equal to some real number, α. For example, if we chose $\alpha = 0$, we could be interested in detecting the presence of $s_1(t)$ while discriminating against $s_2(t)$.

We assume white noise with correlation function

$$R_n(t - u) = N\delta(t - u). \tag{15}$$

Then the mean square noise at t_1, $\sigma^2(t_1)$, is (if we start the problem at $t = 0$ and if we employ the usual formal operations with white noise)

$$\sigma^2(t_1) = N \int_0^{t_1} h^2(\tau) d\tau \tag{16}$$

where $h(t)$ is the impulse response of the linear filter. The outputs due to $s_1(t)$ and $s_2(t)$ at $t = t_1$ are, respectively,

$$y_1(t_1) = \int_0^{t_1} h(\tau)\, s_1(t_1 - \tau)\, d\tau \tag{17}$$

$$y_2(t_1) = \int_0^{t_1} h(\tau)\, s_2(t_1 - \tau)\, d\tau. \tag{18}$$

The problem is then to choose $h(t)$ to maximize

$$\frac{[y_1(t_1)]^2}{\sigma^2(t_1)} \tag{19}$$

while satisfying the relationship

$$y_2(t_1) = \alpha. \tag{20}$$

If we let

$$\dot{x}_1(t) = u(t)s_1(t_1 - t), \tag{21}$$

$$\dot{x}_2(t) = u(t)s_2(t_1 - t), \tag{22}$$

$$\dot{x}_3(t) = -Nu^2(t), \tag{23}$$

$$x_1(0) = x_2(0) = x_3(0) = 0, \tag{24}$$

and if we identify $u(t)$ with $h(t)$, then

$$x_1(t_1) = y_1(t_1), \tag{25}*$$

$$x_2(t_1) = y_2(t_1), \tag{26}$$

$$x_3(t_1) = -\sigma^2(t_1). \tag{27}$$

* Note that for $t \neq t_1$, $x_1(t)$ does not necessarily equal $y_1(t)$.

An equivalent problem is to choose $u(t)$ so as to maximize $x_3(t_1)$ subject to

$$x_1(t_1) = 1 \tag{28}$$

$$x_2(t_1) = \alpha. \tag{29}$$

That is, we can minimize $\sigma^2(t_1)$ with $y_1(t_1)$ constrained since the signal-noise ratio is not changed if $h(t)$ is multiplied by a constant factor.

Now that we have recast the problem into differential equation form, we can solve it using the maximum principle. Using (14) we see that, since f is independent of x.

$$p_1(t) = \text{constant} \equiv p_1 \tag{30}$$

$$p_2(t) = \text{constant} \equiv p_2 \tag{31}$$

$$p_3(t) = \text{constant} \equiv 1 \tag{32}$$

(we let $p_3(t) = 1$ for convenience).*
The Hamiltonian, H, is

$$H = p_1 u(t) s_1(t_1 - t) + p_2 u(t) s_2(t_1 - t) - N u^2(t). \tag{33}$$

Since there are no constraints on $u(t)$, we maximize H by differentiating and get

$$u(t) = \frac{1}{2N} [p_1 s_1(t_1 - t) + p_2 s_2(t_1 - t)]. \tag{34}$$

To satisfy (28) and (29), we must have

$$\frac{1}{2N} \int_0^{t_1} [p_1 s_1(t_1 - \tau) + p_2 s_2(t_1 - \tau)]\, s_1(t_1 - \tau)\, d\tau = 1 \tag{35}$$

$$\frac{1}{2N} \int_0^{t_1} [p_1 s_1(t_1 - \tau) + p_2 s_2(t_1 - \tau)]\, s_2(t_1 - \tau)\, d\tau = \alpha. \tag{36}$$

We can then solve for p_1 and p_2 and get

$$p_1 = \frac{2N(S_2 - \alpha S_{12})}{S_1 S_2 - S_{12}^2} \tag{37}$$

$$p_2 = \frac{2N(\alpha S_1 - S_{12})}{S_1 S_2 - S_{12}^2}, \tag{38}$$

* This involves an assumption of normality (in the sense of the classical calculus of variations).

where

$$S_1 = \int_0^{t_1} s_1^2(t_1 - \tau)\, d\tau \tag{39}$$

$$S_2 = \int_0^{t_1} s_2^2(t_1 - \tau)\, d\tau \tag{40}$$

$$S_{12} = \int_0^{t_1} s_1(t_1 - \tau)\, s_2(t_1 - \tau)\, d\tau. \tag{41}$$

As a simple example, let

$$s_1(t) = 1 \quad t \in [0,t_1] \tag{42}$$

$$\begin{aligned} s_2(t) &= 1 \quad t \in [0,t_1/2] \\ &0 \quad t \in (t_1/2,t_1]. \end{aligned} \tag{43}$$

We would then get

$$u(t) = h(t) = \frac{2(1-\alpha)}{t_1}\, s_1(t_1 - t) + \frac{2}{t_1}\,(2\alpha - 1)\, s_2(t_1 - t). \tag{44}$$

IV. A SIGNAL DESIGN PROBLEM

The following problem is taken from the thesis by M. I. Schwartz (Ref. 6). The system is depicted in Fig. 1. We have a signal passing through a linear time-invariant filter, represented by impulse response function $h(t)$, after which the signal is corrupted by noise. The resultant signal plus noise is processed by a correlation-type receiver. The object is to maximize the signal-noise ratio at the output of the receiver at $t = t_1$ by choosing forms for both $s(t)$ and receiver function $q(t)$. M. I.

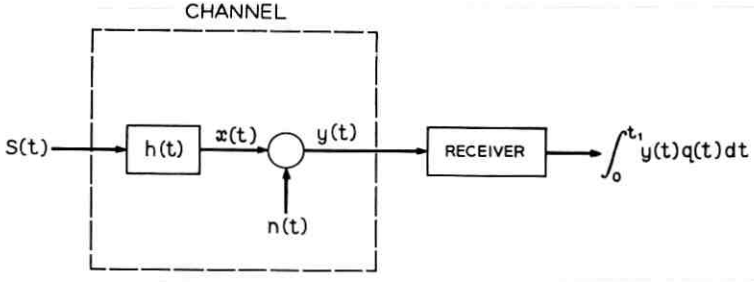


Fig. 1 — Signal design problem.

Schwartz solves this problem with an energy constraint on $s(t)$. We will show how to simultaneously handle an energy constraint and an amplitude constraint on $s(t)$. That is, we require that

$$\int_0^{t_1} s^2(\tau)\, d\tau = \mathcal{E} \tag{45}$$

and

$$| s(t) | \leqq S_{\max} . \tag{46}$$

To make the problem meaningful, we require that

$$(S_{\max})^2 t_1 > \mathcal{E}.$$

It may be easily shown that the problem is equivalent to the problem with the equality of (45) replaced by $\leqq$.

Again, we assume that the noise is white, i.e.,

$$R_n(t - u) = N\delta(t - u), \tag{47}$$

and we further specify that $h(t)$ has a rational Fourier transform with just poles* for simplicity. The assumption of rational Fourier transform facilitates recasting the problem into the differential equation form. Thus,

$$\mathfrak{F}[h(t)] = \frac{\alpha}{D(i\omega)} \tag{48}$$

where α is a real number and $D(i\omega)$ is a polynomial in $i\omega$. Letting

$$D(i\omega) = (i\omega)^l + a_1(i\omega)^{l-1} + a_2(i\omega)^{l-2} + \cdots + a_{l-1}(i\omega) + a_l , \tag{49}$$

(i.e., we have an lth order differential equation in $h(t)$) and

$$x_1 = x \tag{50}$$

$$x_2 = \dot{x} \tag{51}$$

$$\vdots$$

$$x_l = x^{(l-1)} \tag{52}$$

we can represent the effect of $h(t)$ by the following set of first-order differential equations

* If we also assumed zeros in the Fourier transform, we would solve for a derivative of $h(t)$.

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_l \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ & & & \vdots & & \\ 0 & 0 & 0 & & & 1 \\ -a_l & -a_{l-1} & -a_{l-2} & & \cdots & -a_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_l \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \alpha s \end{bmatrix} \tag{53}$$

$$x_1(0) = x_2(0) = \cdots = x_l(0) = 0.$$

We have converted an lth order differential equation into l first-order differential equations.
Let

$$x^* = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_l \end{bmatrix} \tag{54}$$

and let B be the $l \times l$ matrix in (53). Then (53) may be more concisely written as

$$\dot{x}^* = B\, x^* + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \alpha s \end{bmatrix} \tag{55}$$

$$x^*(0) = 0.$$

Now that we have put the effect of $h(t)$ into differential equation form, it remains to cast the signal-noise considerations into differential equations. Recall that the object is to maximize $[s_0(t_1)]^2/\sigma^2(t_1)$ subject to (45) and (46) where

$$s_0(t_1) = \int_0^{t_1} dt\, q(t) \int_0^t du\, s(u)\, h(t-u) \tag{56}$$

$$\sigma^2(t_1) = N \int_0^{t_1} q^2(t)\, dt. \tag{57}$$

If we let

$$\dot{z}_1(t) = s^2(t) \tag{58}$$

$$\dot{z}_2(t) = Nq^2(t) \tag{59}$$

$$\dot{z}_3(t) = x_1(t)q(t) \qquad (z_3(t_1) = s_0(t_1)) \tag{60}$$

$$\dot{z}_4(t) = 2z_3(t)x_1(t)q(t) \qquad (z_4(t_1) = s_0^2(t_1)) \tag{61}^*$$

$$z_1(0) = z_2(0) = z_3(0) = z_4(0) = 0 \tag{62}$$

$$\text{control vector} = u(t) = (s(t), q(t)),$$

then the optimization problem is to choose $s(t)$, subject to relation (46), and $q(t)$ to maximize $z_4(t_1)$ (which equals $s_0^2(t_1)$) subject to $z_1(t_1) = \varepsilon$ and $z_2(t_1) = \sigma^2$. That is, we fix $\sigma^2(t_1)$ at some arbitrary real number and maximize $s_0^2(t_1)$. Just as in the matched filter problem, multiplying $q(t)$ by a constant does not affect the signal-noise ratio.

Now our state vector is the $(l + 4)$-vector, $(x^*, z_1, z_2, z_3, z_4)$. Equation (14) will take the following form ($p(t)$ is an $(l + 4)$-vector):

$$\dot{p}(t) = -\begin{bmatrix} \begin{bmatrix} B^T \\ (l \times l) \\ \end{bmatrix} & \begin{matrix} 0 & 0 & q & 2z_3q \\ 0 & 0 & 0 & 0 \\ & & & \vdots \\ 0 & 0 & 0 & 0 \end{matrix} \\ \begin{matrix} 0 & \cdots & 0 \\ 0 & \cdots & 0 \\ 0 & \cdots & 0 \\ 0 & \cdots & 0 \end{matrix} & \begin{matrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2x_1q \\ 0 & 0 & 0 & 0 \end{matrix} \end{bmatrix} p(t) \tag{63}$$

(B is the $l \times l$ matrix in (53)).
The final conditions on $p(t)$ are

$$p_1(t_1) = 0 \tag{64}$$

$$p_2(t_1) = 0 \tag{65}$$

$$\vdots$$

$$p_l(t_1) = 0, \tag{66}$$

$p_{l+1}(t_1)$, $p_{l+2}(t_1)$ unspecified (they correspond to $z_1(t)$ and $z_2(t)$ which are fixed at $t = t_1$)

* This differential equation is derived as follows:

$$dt\,(z_3^2(t)) = 2z_3(t)\dot{z}_3(t) = 2z_3(t)x_1(t)q(t)$$

$$z_4(t_1) = z_3^2(t_1) = s_0^2(t_1).$$

$$p_{l+3}(t_1) = 0 \tag{67}$$

$$p_{l+4}(t_1) \geqq 0. \tag{68}$$

An example of the use of the maximum principle for this signal design problem is given in the next section.

V. EXAMPLE OF SIGNAL DESIGN

We shall consider the case of

$$\mathfrak{F}[h(t)] = \frac{\alpha}{i\omega + \alpha}. \tag{69}$$

Following the method described in Section IV, the problem is first recast into the following differential equation form

$$\dot{x} = -\alpha x + \alpha s \tag{70}$$

$$\dot{z}_1 = s^2 \qquad (z_1(t_1) = \varepsilon) \tag{71}$$

$$\dot{z}_2 = Nq^2 \qquad (z_2(t_1) = \sigma^2) \tag{72}$$

$$\dot{z}_3 = xq \qquad (z_3(t_1) = s_0(t_1)) \tag{73}$$

$$\dot{z}_4 = 2z_3xq \qquad (z_4(t_1) = s_0^2(t_1)) \tag{74}$$

$$x(0) = z_1(0) = z_2(0) = z_3(0) = z_4(0) = 0. \tag{75}$$

For this problem the necessary condition (maximum principle) for an optimal solution is that there exists a nonzero vector function $p(t) = (p_1(t), p_2(t), \cdots, p_5(t))$ such that

$$H = p_1(-\alpha x + \alpha s) + p_2s^2 + p_3Nq^2 + p_4xq + p_52z_3xq \tag{76}$$

is maximized over the allowable s and q and such that

$$\dot{p}_1 = \alpha p_1 - qp_4 - 2z_3q \tag{77}$$

$$\dot{p}_2 = 0 \tag{78}$$

$$\dot{p}_3 = 0 \tag{79}$$

$$\dot{p}_4 = -2xq \tag{80}$$

$$\dot{p}_5 = 0 \tag{81}$$

$$p_1(t_1) = 0 \tag{82}$$

$$p_4(t_1) = 0 \tag{83}$$

$$p_5(t_1) \geqq 0 \tag{84}$$

(we can let $p_5(t_1) = p_5(t) = 1$ under a normality assumption). The maximization of H leads to

$$q(t) = -\left[\frac{p_4(t)\,x(t) + 2z_3(t)\,x(t)}{2p_3N}\right] \tag{85}*$$

$$s(t) = s^*(t) \quad \text{if} \quad |s^*(t)| \leqq S_{\max} \tag{86}$$

$$= \frac{s^*(t)}{|s^*(t)|} S_{\max} \quad \text{if} \quad |s^*(t)| > S_{\max}, \tag{87}$$

where

$$s^*(t) = \frac{-p_1(t)\alpha}{2p_2}.$$

To verify satisfaction of the maximum principle, it is necessary to solve the differential equations (70) to (74) and (77) to (81) with satisfaction of the above mentioned initial and final conditions and in such a way that maximization of the Hamiltonian is satisfied.

The numerical method of satisfying the maximum principle is based on iteration of the initial values of p-vector to successively improve the final conditions. That is, we know the initial conditions of x, z_1, z_2, z_3, z_4 (see (75)) and we wish to constrain the final values of z_1, z_2, p_1, and p_4 :

$$z_1(t_1) = \varepsilon \tag{88}$$

$$z_2(t_1) = \sigma^2 \tag{89}$$

$$p_1(t_1) = 0 \tag{90}$$

$$p_4(t_1) = 0. \tag{91}$$

Suppose we guess at $p(0) = (p_1(0), p_2(0), p_3(0), p_4(0), 1)$ and integrate the differential equations (70) to (74) and (77) and (81) and evaluate the following error in final conditions

$$E = |z_1(t_1) - \varepsilon| + |z_2(t_1) - \sigma^2| + |p_1(t_1)| + |p_4(t_1)|. \tag{91}\dagger$$

We wish to decrease E. To this end, let

$$(p_1(0))_{\text{new}} = (p_1(0))_{\text{old}} + \delta p_1(0) \tag{92}$$

and re-integrate the differential equations. If E decreases, change

* Since $\dot{p}_4(t) + 2\dot{z}_3(t) = 0$, $q(t)$ is proportional to $x(t)$. Thus, we are equivalently maximizing the signal energy into the receiver (see (73)) and then correlating with $x(t)$. This is consistent with (and, in fact, rederives) well-known properties of matched filters.

† Actually, the second term is not required as σ^2 can be adjusted by changing $q(t)$ by a multiplicative factor (without changing the signal-noise ratio).

$p_2(0)$. If E does not increase, try

$$(p_1(0)_{\text{new}} = (p_1(0))_{\text{old}} - \delta p_1(0).$$

Again see whether E has decreased. If it has, change $p_1(0)$ to $(p_1(0))_{\text{new}}$ and change $p_2(0)$. If E has not decreased, retain $(p_1(0))_{\text{old}}$ and try changing $p_2(0)$.

Thus, the method is to successively change $p_1(0)$, $p_2(0)$, $p_3(0)$, $p_4(0)$ to decrease E. When E becomes sufficiently small, the maximum principle may be said to be satisfied. After we present some results, we will discuss the method further.

5.1 *Results*

Two cases were run. They were for the following parameters:

$$t_1 = 1$$

$$\mathcal{E} = 1$$

$$N = 1$$

$$\alpha = 1.$$

(As mentioned previously, σ^2 is determined by the scaling of $q(t)$.) The difference between the two cases is that in the first, the amplitude constraint was not imposed and in the second, $S_{\max}$ was set at 1.1. The first case was already treated by other methods in Ref. 6. Our results for that case were in agreement with those of Ref. 6. Fig. 2 shows $q(t)$ for both cases and Fig. 3 shows $s(t)$. For the case of no amplitude constraint the signal-noise ratio $(\sqrt{s_o^2(t_1)/\sigma^2})$ was 0.44 and the signal-noise ratio for the amplitude constrained case was 0.43. One could (nonoptimally) impose the amplitude constraint by scaling down the results for the amplitude unconstrained case (and not use all the signal energy available). That is, the peak amplitude of the signal in the first case is 1.27. The entire signal $(s(t)$ for $t \in [0,t_1])$ could be reduced by a factor of 1.1/1.27. The signal-noise ratio would also be reduced by that factor (0.865). Whether or not this signal-noise ratio reduction is significant is not actually germane to this investigation. What is of consequence is the fact that the optimum can be determined and any sub-optimum scheme can be compared with it.

5.2 *Comments on the Numerical Method*

The basic method is similar to that of Ref. 7. In Ref. 7, the gradient (relating changes in the error to changes in $p(t_0)$) is evaluated and used

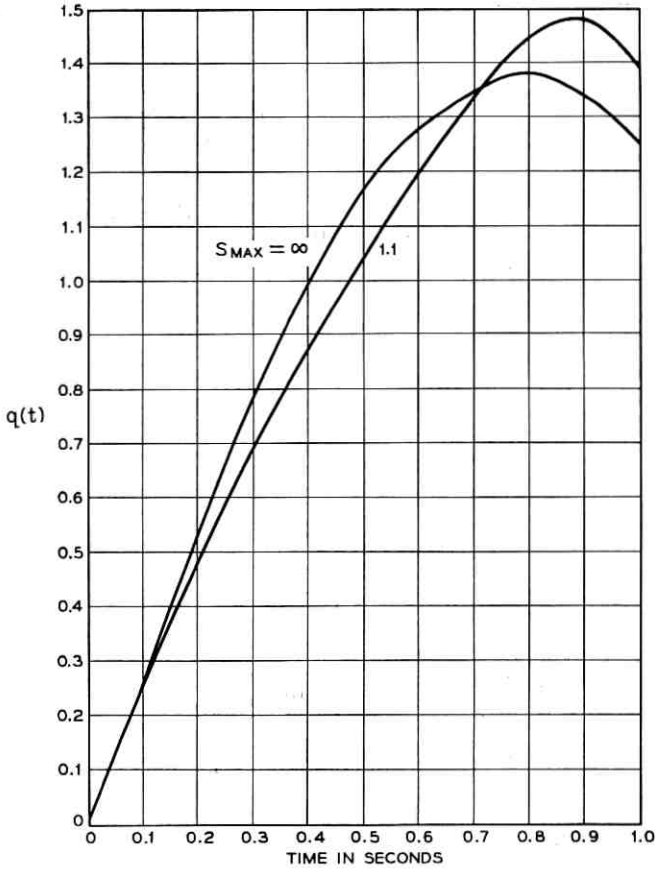


Fig. 2 — Optimum correlation waveforms.

to determine a steepest descent change in $p(t_0)$. This was not done in the present problem for two reasons. First of all, the evaluation of the gradient (as done in Ref. 7) is not valid for the problem with inequality constraints. Secondly, even if the gradient can be conveniently evaluated, it still requires extra integrations and the problem of step size is left unresolved. (This is not intended as criticism of the method of Ref. 7 which may be quite useful in many applications.) We decided to frankly treat the problem as a systematic trial and error. Our method of seeing how the error changes as $p_i(t_0)$ is changed to $p_i(t_0) + \delta p_i(t_0)$ may be loosely interpreted as evaluating $\partial E/\partial p_i(t_0) \cdot \delta p_i(t_0)$.

The numerical method may be considered to be semiautomatic. There is little *a priori* information available as to the initial choices of the

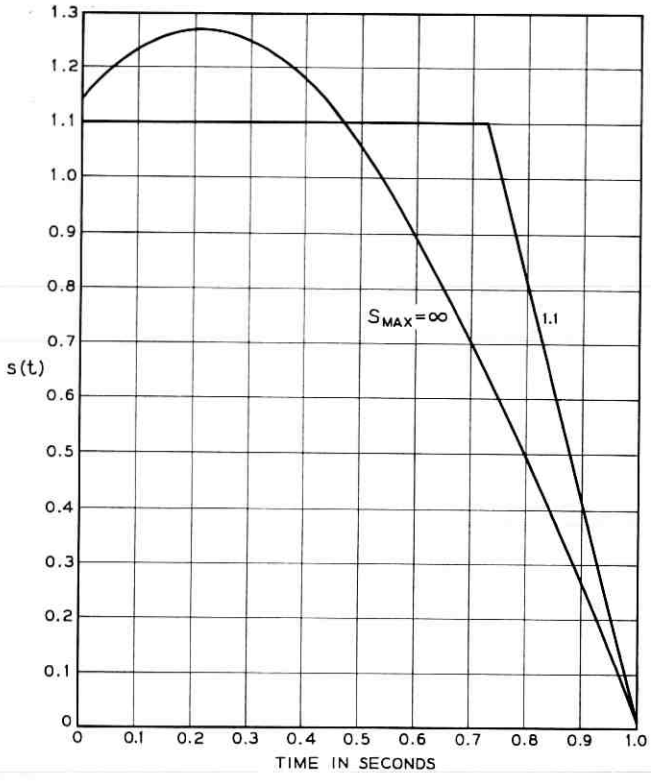


Fig. 3 — Optimum signal waveforms.

$p_i(t_0)$ and the $\delta p_i(t_0)$. A few runs on the computer offer the optimizer some insight as to appropriate choices. About 30 iterations were found to be needed for convergence (approximately 0.01 hours of computer time). No convergence proof is offered for this method. In fact, even though it did not happen in the problem considered in this paper, it is conceivable that $\partial E/\partial p_i(t_0)$ (assuming the derivative exists) can be so large that the smallest $\delta p_i(t_0)$ that can be used by the computer would result in a much too large change in E. There is also the possibility of local minima of E (with $E > 0$). These problems (which may not even occur; we are trying to anticipate the worst) could be presumably resolved by changing the metric defining E and by trying a wide range of $p_i(t_0)$.* It may be noted that convergence proofs do not seem to be avail-

* More efficient methods of adjusting the $p_i(t_0)$ may be possible. See, for example, Wilde, D. J., *Optimum Seeking Methods*, Prentice-Hall, 1962.

able for competitive algorithms (e.g., steepest descent) for these optimization problems.

VI. CONCLUSION

The maximum principle has been used here to attack two signal-noise ratio maximization problems. The first one (matched filter problem) could have been solved without the maximum principle. However, the maximum principle provided a very neat and systematic approach. The second problem (with the amplitude constraint included) appears to be unsolvable except by the maximum principle.* In this paper, it was assumed that the noise is white. The handling of non-white noise merits further attention, in particular, the conversion to differential equations and the presence of impulses (see Ref. 5, Appendix 2). It should also be noted that the maximum principle is not conceptually limited to time-invariant problems.

VII. ACKNOWLEDGMENTS

I am very grateful to M. I. Schwartz for many valuable discussions during the course of this work. I also wish to express my appreciation to T. Burford, I. Jacobs, and W. L. Nelson for helpful comments on the paper, and to Miss N. M. Klujber for her excellent programming.

REFERENCES

1. Schweppe, F. C., Optimization of Signals, Lincoln Laboratories Group Report 1964-4, January 16, 1964.
2. Tufts, D. W. and Shnidman, D. A., Optimum Waveforms Subject to Both Energy and Peak-Value Constraints, Proc. of the IEEE, September, 1964, pp. 1002–1007.
3. Halkin, H., Mathematical Foundations of System Optimization, to appear as a chapter in *Optimization Techniques*, Vol. 2, ed. by G. Leitmann.
4. Pontryagin, L. S., et al, *The Mathematical Theory of Optimal Processes*, (translation) John Wiley and Sons, Inc., New York, 1962.
5. Davenport, W. B., Jr. and Root, W. L., *An Introduction to the Theory of Random Signals and Noise*, McGraw-Hill Book Company, Inc., New York, 1958.
6. Schwartz, M. I., Binary Signal and Receiver Design for Linear Time Invariant Channels, Sc.D. Thesis, New York University, September, 1964.
7. Knapp, C. H. and Frost, P. A., Determination of Optimum Control and Trajectories Using the Maximum Principle in Association with a Gradient Technique, IEEE Trans. on Automatic Control, *AC-10*, No. 2, April, 1965, pp. 189–193.

* Another possible approach to the problem would be to assume that $s(t) = S_{max}\sin[\theta(t)]$ and use classical methods to solve for $\theta(t)$ (since the problem of inequality constraint is thereby avoided). However, the maximum principle attacks the problem directly.

RECENT RELATED REFERENCES*

1. Schweppe, F. C. and Gray, D., Radar Amplitude Modulation Using Pontryagin's Maximum Principle, Lincoln Laboratory preprint, 1965.
2. Athans, M. and Schweppe, F. C., On the Design of Optimal Modulation Schemes Via Control Theoretic Concepts. I: Formulation, Lincoln Laboratory preprint, Aug. 31, 1965.

* I wish to thank Prof. M. Athans of MIT for sending these references.

Contributors to This Issue

MORGAN M. BUCHNER, JR., B.E.S., 1961, Ph.D., 1965, The Johns Hopkins University; Bell Telephone Laboratories, 1965—. Mr. Buchner has been engaged in a study of impulse noise in an effort to better understand its characteristics and its effects upon data communications. Member, IEEE, Tau Beta Pi, Sigma Xi, Eta Kappa Nu.

ROGER M. GOLDEN, B.S., 1954, M.S., 1955, Ph.D., 1959, California Institute of Technology; Fulbright student Technical Institute at Eindhoven, 1959–1960; Bell Telephone Laboratories, 1960—. Since joining Bell Laboratories, Mr. Golden has been working on speech bandwidth compression devices, vocoders, and speech analysis-synthesis systems for telephone communications. He is presently studying such systems by means of newly-developed digital computer simulation techniques. Member, Acoustical Society of America, IEEE, Sigma Xi, Tau Beta Pi, Association for Computing Machinery.

JACK M. HOLTZMAN, B.E.E., 1958, City College of New York; M.S., 1960, University of California (Los Angeles); Hughes Aircraft Company, 1958–1963; Bell Telephone Laboratories, 1963—. At present, Mr. Holtzman is working toward the Ph.D. degree in system science at Polytechnic Institute of Brooklyn. His work has been primarily in various aspects of systems and control theory. Member, IEEE.

DANIEL LEED, B.S., 1941, College of the City of New York; M.E.E., 1957, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1946—. Mr. Leed heads a group concerned with the development of instrumentation for measuring the frequency characteristics of passive and active networks with VHF and microwave frequency ranges. This work has led to the generation of special techniques for the broadband characterization of solid-state devices.

WILLIAM H. STEIER, B.S.E.E., 1955, Evansville College; M.S.E.E., 1957 and Ph.D. (E.E.), 1960, University of Illinois; Bell Telephone Laboratories, 1962—. Mr. Steier first worked on the millimeter wave circular waveguide transmission system. More recently he has worked

on optical transmission lines and gas lenses. Member, American Physical Society, IEEE.

AARON D. WYNER, B.S., 1960, Queens College; B.S.E.E., 1960, M.S., 1961, and Ph.D., 1963, Columbia University; Bell Telephone Laboratories, 1963—. Mr. Wyner has been engaged in research in various aspects of information theory. He is also Adjunct Assistant Professor of Electrical Engineering at Columbia University. Member, IEEE, Tau Beta Pi, Eta Kappa Nu, Sigma Xi.