THE BELL SYSTEM Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XLIII JULY 1964 NUMBER 4, PART 2

(*Contents continued on inside back cover*)

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, E. J. McNeely, President, C. E. Wampler, Vice President and Secretary, J. J. Scanlon, Vice President and Treasurer, 195 Broadway, New York, N. Y. 10007. Subscriptions are accepted at $5.00 per year. Single copies $1.25 each. Foreign postage is $1.08 per year or 18 cents per copy. Printed in U.S.A.

THE BELL SYSTEM

TECHNICAL JOURNAL

VOLUME XLIII JULY 1964 NUMBER 4, PART 2

Timing Errors in a Chain of Regenerative Repeaters, III

By B. K. KINARIWALA

(Manuscript received March 2, 1964)

We present here a general and rigorous theory of the jitter accumulation in a chain of regenerative repeaters. The sources of jitter are assumed to be the signal-dependent sources, as distinguished from purely random sources independent of the signal.

Our results show that while the absolute jitter and its dispersion grow without bound with the number of repeaters, the spacing and the alignment jitter remain bounded. In particular, the spacing jitter bounds are quite optimistic for most practical situations, viz., no greater than twice the absolute jitter injected at a single repeater. This result is of importance in that it ensures proper decoding of the binary signal. Its further importance is that it does ensure, in most cases, the validity of the basic model and thus the validity of other results obtained by that model. One such result shows that the alignment jitter is slowly-varying for repeaters farther along the chain. We also include some results which would be of use in computations, together with a simple example.

I. INTRODUCTION

1.1 *Purpose*

Pulse regeneration is an attractive feature of digital communication systems. A regenerator or a regenerative repeater must: (*i*) detect the presence or absence of a pulse at certain time instants which are, ideally, multiples of the basic pulse repetition period, (*ii*) regenerate the pulse shapes, and (*iii*) retime these pulses so they occur at appropriate instants of time in the outgoing signal. In practice, errors in detection due to noise, distortion, etc. in the system and errors in retiming the signal impose limitations on the operation of such systems. Except for noise which may lead to the detection of a pulse where no pulse exists in the original signal and vice versa, the imperfections in the system show up as a jittering of the pulse positions in the outgoing signal. In self-timing repeaters,[1] the jitter from all sources except random noise is also dependent on the signal itself.[2]

Several workers have investigated many different aspects of the timing jitter problem.[1-9] We study here the signal-dependent jitter due to repeated regeneration. This article leans quite heavily on our previous discussion, and in fact it supplements as well as complements the previous results referred to in this article as Parts I and II.[3]

1.2 *Background*

In our previous discussion, we showed how the timing jitter in a pulse train accumulates as the pulse train is repeatedly regenerated. The timing information was assumed to be extracted from the incoming signal itself, and the timing extractor was assumed to be a tuned circuit. Mistuning in the tuned circuit was assumed, as a convenience, to be the major source of jitter. However, the accumulation properties of timing jitter are not dependent on a particular source of jitter. We shall attempt to clarify this point here.

Our previous results showed that the displacements of the pulse positions from their original positions (or the "absolute jitter") increase indefinitely with the number of repeaters. The major component in the absolute jitter was found to be flat delay (i.e., the same displacement at every pulse position). A natural question follows: How does the absolute jitter behave if the flat delay is removed? This is the "dispersion" or the absolute jitter measured against the reference clock delayed by an appropriate amount. It was shown that the dispersion also increases without bound except when the pulse trains are severely constrained (e.g., periodic, finite, etc.). These results are valid even under the constraint that there exists at least one pulse in a predetermined number of the basic periods or "time slots."

It is worth noting that the absolute jitter and the dispersion have as counterparts the average and the variance of the random variable represented by the pulse displacement at any pulse position. We wish to emphasize, however, that our results are independent of any a priori statistics concerning the pulse train ensemble.

1.3 *Results*

In the present article, our most significant result concerns a more important parameter, viz., the "spacing jitter," which is the variation in the spacing between adjacent pulses. We show that the spacing jitter is bounded for an indefinitely long chain of repeaters, and the bound is directly related to the minimum pulse density. Such bounds may be precisely evaluated both for the infinite as well as the finite chain of repeaters. The importance of these results lies in the precise evaluation of the bounds, the means to control these bounds, and in our ability to relate these bounds either with resulting errors in the decoding of the signal,[2] or with distortion in the analog signal,[4] depending on whether the bounds are large or small. As will be seen later, these results also determine the validity or otherwise of all the other previous results on the timing problem, since the present results have a direct bearing on the validity of the model used by most people.

We also present a rather thorough discussion of the computational aspects of the problem in the Appendix. Special situations such as periodic patterns, truncations, pattern transitions, etc. are included in our discussion.

The case of nonidentical repeaters is examined briefly. It appears that the bounds on jitter are not appreciably different if the repeaters are not appreciably different. The "misalignment," or the jitter introduced by a single repeater in an already jittered pulse train, is also examined briefly. We show that the misalignment is slowly-varying for repeaters further along in the chain.

Let us emphasize, in conclusion, that our results are not dependent on any a priori statistics. Our analysis is quite rigorous once the basic model is derived. The basic model is essentially the same as that of other investigators, and the major assumption in the model asserts that a single repeater introduces only slowly-varying jitter in a jitter-free pulse train. Such an assumption is quite reasonable for any practical repeater.

1.4 *Organization*

We start with a mathematical statement of the problem. Our formulation shows that the input and the output jitter sequences are related to each other by a linear operator which maps the space of bounded sequences into itself. The dimensionality of the space is determined by the memory of the system, which is infinite for an infinitely long chain of repeaters. We are thus led to a discussion of the operator in a Banach space of infinite dimensions. The spectral properties of the operator determine the behavior of the absolute jitter and the dispersion. Next, we consider the spacing jitter obtained by a difference operation on the absolute jitter. The brief discussions on the misalignment, unequal repeaters, etc. follow. Finally, we present several results to facilitate computations.

II. STATEMENT OF PROBLEM

The basic model of the repeater is represented by a tuned circuit excited by a train of pulses. The natural frequency of the tuned circuit is assumed to be very close to the pulse repetition frequency. At upward (or downward) zero crossings of the response of the tuned circuit, timing pulses are generated which determine the instants of outgoing pulses; the presence or absence of a particular pulse in the output signal is determined by the presence or absence of a pulse in the input signal.

Let $\{\cdots, -2\tau, -\tau, 0, \tau, \cdots\}$ represent the instants of occurrence of pulses for the ideal pulse train. These would be the centers of the corresponding time slots. The occurrence or nonoccurrence of a pulse at $t = -n\tau$ is determined by the value of the random binary variable $\mathfrak{A}_n$. A pulse is present when $\mathfrak{A}_n = 1$ and no pulse is present when $\mathfrak{A}_n = 0$.

At this point it is convenient to assume that the pulse train consists of impulses located at the actual pulse positions defined above. The actual pulse shapes modify the zero crossings of the response of the tuned circuit, and a term representing such a correction can be added separately.

Finally, let $\xi_k{}^l$ be the displacement of the kth pulse (originally located at $t = -k\tau$) at the output of the lth repeater in a chain of repeaters. At the input of the lth repeater the timing displacement is given by $\xi_k{}^{l-1}$, which is the jitter value at the output of the $(l - 1)$th repeater. The displacement (or jitter) is measured in radians, where 2π corresponds to the pulse interval τ.

In order to determine $\xi_k{}^l$, we merely find the appropriate zero crossing of the response of the tuned circuit excited by a train of pulses. The exciting pulse train is itself jittered and this fact is represented by the values of $\{\xi_k{}^{l-1}\}$. It turns out that the $\xi_k{}^l$ is actually a nonlinear function of the set $\{\xi_{k+n}{}^{l-1}\}$ with $n = 0, 1, 2, \cdots$. Furthermore, it also depends on the original signal represented by the set $\{\mathfrak{A}_n\}$ and, of course, the Q of the resonant circuit. If, however, $\{\xi_{k+n}{}^{l-1}\}$ satisfy certain conditions, it is possible to represent the $\xi_k{}^l$ as a linear function of the variables

$\{\xi_{k+n}{}^{l-1}\}$. The function is still dependent on $\{\alpha_n\}$ and the Q of the circuit. This is the fundamental relation between the input and output jitter and, if there were no jitter introduced by the repeater, it would take the following form:[2]

$$\xi_k{}^l = \frac{\sum_{n=0}^{\infty} \alpha_{n+k}\beta^n \xi_{n+k}{}^{l-1}}{\sum_{n=0}^{\infty} \alpha_{n+k}\beta^n}, \qquad (k = 0,1,2,\cdots; l = 1,2,\cdots), \quad (1)$$

where $\beta = \exp(-\pi/Q) \approx 1 - (\pi/Q)$ for large Q.

The conditions that must be satisfied by $\xi_k{}^{l-1}$ are the following:

$$|\xi_k{}^{l-1} - \xi_{k-1}{}^{l-1}| \ll \pi/Q \qquad \text{for all } k. \quad (2)$$

These conditions are unaltered when (1) is slightly modified to include the jitter introduced by each repeater [cf. (3)]. It is therefore very important to make sure that (2) holds in order for any of the results obtained on the basis of (1) to be valid. The quantity required to be small in (2) is the spacing jitter, whose behavior is important in that if it ever becomes too large there will occur errors and distortion in the decoded signal.[2,4] It is our intention here to investigate thoroughly the behavior of the spacing jitter. Equation (1) represents the way in which jitter propagates along a chain of repeaters. The jitter accumulation properties of a chain are, therefore, a consequence of this basic equation. Of course, at every repeater there is jitter injected in addition to the one propagated from previous repeaters. Since we are not discussing here effects of random noise, the sources of the injected jitter are signal-dependent, and they are identical if we assume identical repeaters. The jitter injected at every repeater by signal-dependent sources is thus identical. For such sources, this injection of jitter is simply additive either at the input end of the repeater[5] or at the output end.[2,3,6,7] For example, dispersion in the channel with a consequent imperfect detection of the pulse positions would be an additive source of jitter at the input of the repeater.[5] Certain nonlinear operations (such as limiting) on the response of the tuned circuit would inevitably alter the zero crossings, and this may be represented as an additive source at the output end of the repeater. If the mistuning of the resonant circuit is small, it can be shown that mistuning represents an additive source at the output end. Finite pulse widths also represent an additive source at the output.[2] In any case, such injection of jitter depends on the signal and repeater parameters. Since these are the same at every repeater, the injected jitter is the same at every repeater. Additivity of these sources is usually a consequence of the fact that the injected jitter is small. For convenience, we will refer all these sources to the output end and represent the injected jitter by $\{\xi_k^1\}$, the output of the first repeater, where the original pulse train is assumed to be jitter-free. It is somewhat inconvenient to refer all sources to the input, since it would involve inverting the functional relationship of (1). Our interest in this article is not to discuss the quantitative evaluation of $\{\xi_k^1\}$. For simple sources such as mistuning and finite pulse widths, it is relatively easy to evaluate $\{\xi_k^1\}$. We shall discuss elsewhere the question of determining $\{\xi_k^1\}$ when all sources are included. However, let us emphasize that we do not restrict the sources of jitter to short memory mechanisms.[5]

In this paper, our interest is to determine the behavior of $\{\xi_k^l\}$ for large l when the basic equation (1) is modified to include the injected jitter:

$$\xi_k^l = \frac{\sum_{n=0}^{\infty} \mathcal{A}_{n+k}\beta^n \xi_{n+k}^{l-1}}{\sum_{n=0}^{\infty} \mathcal{A}_{n+k}\beta^n} + \xi_k^1, \tag{3}$$

subject to condition (2), and where we assume that $\xi_k^0 = 0$. Condition (2) is satisfied by $\{\xi_k^0\}$ since they are all zero and by $\{\xi_k^1\}$ because for a practical repeater the jitter injected by a single repeater must be extremely small. We have made no assumptions on the nature of the signal. Thus, if it can be shown that condition (2) is valid for every l, then the results obtained by using (3) are valid. This is an important point which cannot be overemphasized. We will therefore pay particular attention to the behavior of the spacing jitter.

III. RECAPITULATION

In this section, we recapitulate the basic formulation of the problem developed in Parts I and II of this article. For details, the reader is referred to the original discussion. Define a vector

$$X_l = \{\xi_0^l, \xi_1^l, \xi_2^l, \cdots\}, \qquad (l = 1, 2, \cdots); \tag{4}$$

and $X_0 = 0$. Then, the basic equation (3) may be written as

$$X_l = T_0 X_{l-1} + X_1, \tag{5}$$

where

$$T_0 = \begin{bmatrix} \dfrac{\mathcal{A}_0}{s_0} & \dfrac{\mathcal{A}_1\beta}{s_0} & \dfrac{\mathcal{A}_2\beta^2}{s_0} & \cdots \\ 0 & \dfrac{\mathcal{A}_1}{s_1} & \dfrac{\mathcal{A}_2\beta}{s_1} & \cdots \\ 0 & 0 & \dfrac{\mathcal{A}_2}{s_2} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}, \tag{6}$$

and

$$s_i = \sum_{n=0}^{\infty} \mathcal{A}_{n+i}\beta^n. \tag{7}$$

We have thus expressed our original problem in terms of an operator T_0 which maps the Banach space $\mathbf{1}_\infty$ (the space of bounded sequences) into itself. We are interested in the behavior of X_l as l approaches infinity. The operator T_0 is the most general one. However, there is some interest in the behavior of the jitter when the pulse trains are periodic and the operator T_0 under steady-state periodic condition (cf. Part I) becomes

$$A_0 = \begin{bmatrix} \dfrac{\mathcal{A}_0}{s_0'} & \dfrac{\mathcal{A}_1\beta}{s_0'} & \cdots & \dfrac{\mathcal{A}_{m-1}\beta^{m-1}}{s_0'} \\ \dfrac{\mathcal{A}_0\beta^{m-1}}{s_1'} & \dfrac{\mathcal{A}_1}{s_1'} & \cdots & \dfrac{\mathcal{A}_{m-1}\beta^{m-2}}{s_1'} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{\mathcal{A}_0\beta}{s_{m-1}'} & \cdots & \cdots & \dfrac{\mathcal{A}_{m-1}}{s_{m-1}'} \end{bmatrix}, \tag{8}$$

where $s_n' = (1 - \beta^m)s_n$ for all n. Finally, it is more convenient to write the operator T_0 explicitly to indicate only those positions where the pulses are present. This leads us to the operator

$$T = \begin{bmatrix} \dfrac{1}{S_0} & \dfrac{\beta^{i_1}}{S_0} & \dfrac{\beta^{i_1+i_2}}{S_0} & \cdots \\ 0 & \dfrac{1}{S_1} & \dfrac{\beta^{i_2}}{S_1} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}, \tag{9}$$

where

$$S_0 = s_0 \tag{10}$$

and

$$S_{n-1} = 1 + \beta^{i_n} S_n \,. \tag{11}$$

The vectors X_l are also assumed to be suitably modified. We will use the operator T of (9) to represent our quantities of interest.* The operator T_0 will give identical results. The special restriction to the periodic case will be of interest when we discuss computational aspects. The operator A_0 is also assumed suitably modified to A. Several parameters of interest may now be expressed in terms of the operator T and the injected jitter element X_1.

3.1 *Absolute Jitter*

From (5), we have

$$X_l = \left[\sum_{\nu=0}^{l-1} T^{\nu} \right] X_1 \,, \tag{12}$$

and in the limit

$$Y = \lim_{l \to \infty} \left[\sum_{\nu=0}^{l-1} T^{\nu} \right] X_1 \,. \tag{13}$$

3.2 *Dispersion*

This is obtained by subtracting the average delay from the absolute jitter. A delay element is represented to within a constant by $\{1,1,1, \cdots\}$. This element happens to be an eigenvector of T corresponding to an eigenvalue at $\lambda = 1$. Under the condition that $\lambda = 1$ is a pole (cf. Part II) of T, we can represent the dispersion by

$$X_l^{D} = (I - E_1) X_l = \left[I - E_1 \right] \left[\sum_{\nu=0}^{l-1} T^{\nu} \right] X_1 \,, \tag{14}$$

where E_1 is the projection operator E_1 $(\lambda = 1;\, T)$ which takes on the value "one" in the neighborhood of $\lambda = 1$ and zero elsewhere.

* It should be observed that the condition (2) is also modified. The right-hand side of (2) now becomes π/Q times the number of basic periods τ between two adjacent pulses.

3.3 *Spacing Jitter*

This is obtained by subtracting from the absolute jitter at one pulse position the absolute jitter at the adjacent pulse position (not necessarily the adjacent time slot).

$$X_l^s = \left[I - S\right]\left[\sum_{\nu=0}^{l-1} T^\nu\right] X_1, \tag{15}$$

where S is the shift operator given by

$$S = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & \cdots \\ 0 & 0 & 0 & 1 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}. \tag{16}$$

3.4 *Alignment Jitter*

This quantity is given by the difference of X_{l+1} and X_l :

$$X_{l+1}{}^A = X_{l+1} - X_l = T^l X_1 . \tag{17}$$

Finally, we summarize here without proof the properties of the operator T which were determined in Part II of this article.

(a) The operator T is a bounded operator mapping $\mathbf{1}_p$ into itself for $1 \leqq p \leqq \infty$. In particular, the norm of T in $\mathbf{1}_\infty$ is equal to one (i.e., $| T | = 1$).

(b) The spectrum of T is a subset of the unit disk (i.e., $| \sigma(T) | \leqq 1$), and any pole λ of T with $| \lambda | = 1$ has order one.

(c) All points in the unit circle except $\lambda = 1$ are in $\rho(T)$, the resolvent of T. The point $\lambda = 1$ is an eigenvalue of T with the eigenvector $\{1, 1, \cdots\}$. The dimension of the eigenmanifold is one in this case.

(d) The point $\lambda = 1$ is the limit point of the point spectrum of T if the domain of T is unrestricted. It is a pole of T for very special cases, such as periodic pulse trains, truncated pulse trains, etc.

IV. ABSOLUTE JITTER AND DISPERSION

The results for absolute jitter follow immediately from (13)

$$Y = \lim_{l \to \infty} \left[\sum_{\nu=0}^{l-1} T^\nu\right] X_1 .$$

Observe that $\lambda = 1$ is an eigenvalue of T, and since in general X_1 is any element of $\mathbf{1}_\infty$, the limit in the above equation approaches infinity as l

approaches infinity. In particular, the norm of the operator

$$K_l = \left[\sum_{\nu=0}^{l-1} T^\nu\right] \tag{18}$$

is l. We thus observe that, in general, the absolute jitter grows linearly with the number of repeaters.

If $\lambda = 1$ is a pole of T, the dispersion is given by

$$\begin{aligned} X_{l+1}^D &= \left[I - E_1\right]\left[\sum_{\nu=0}^{l} T^\nu\right] X_1 , \\ &= \left[\sum_{\nu=0}^{l} T^\nu\right]\left[I - E_1\right] X_1 \qquad (\text{since } E_1T = TE_1), \\ &= \left[\sum_{\nu=0}^{l} D^\nu\right]\left[E_D X_1\right] \qquad (\text{cf. Part II}); \end{aligned} \tag{19}$$

where $D = TE_D$ and $E_D = (I - E_1)$. Here E_DX_1 is the dispersion element due to the first repeater. The norm of the operator

$$K_{l+1}^D = \left[\sum_{\nu=0}^{l} D^\nu\right] \tag{20}$$

is bounded and converges to a finite value as l approaches infinity. This follows from the previous discussion (see Part II) where it was shown that

$$| D^m | \leqq M\alpha_0^m , \tag{21}$$

where M is a positive constant and $\alpha_0 < 1$. We therefore find that the dispersion is bounded provided that $\lambda = 1$ is a pole of T. This is true for certain highly constrained situations. In particular, this is true when the domain of T is restricted to a finite dimensional subspace of $\mathbf{1}_\infty$ which is invariant under T. Examples of such cases occur when the pulse trains are periodic, finite, etc.

On the other hand, when $\lambda = 1$ is not a pole of T the projection E_1 does not commute with T. In this case,

$$X_{l+i}^D = \left[I - E_1\right]\left[\sum_{\nu=0}^{l} T^\nu\right] X_1 \tag{22}$$

where

$$E_1T \neq TE_1 . \tag{23}$$

Furthermore, it can be shown that

$$|(I - E_1)T^m| = 1 \qquad \text{for all } m. \tag{24}$$

This is a consequence of the spectral properties of T summarized in the previous section, viz., the point $\lambda = 1$ is a limit point of the point spectrum of T. In fact, we show in Part II of this paper that there exist elements in $\mathbf{1}_\infty$ such that the norm of the operator

$$K_{l+1}^D = \left[I - E_1\right]\left[\sum_{\nu=0}^{l} T^\nu\right] \tag{25}$$

is $(l + 1)$. It follows, therefore, that the dispersion grows without bound in the case of purely unconstrained pulse trains. The result is not altered even if some form of coding is provided to eliminate indefinitely long strings of zeros in the pulse trains.

V. SPACING JITTER

At the output of the lth repeater, the spacing jitter is given by (15),

$$X_l^s = \left[I - S\right]\left[\sum_{\nu=0}^{l-1} T^\nu\right] X_1 = [K_l^s]X_1 . \tag{26}$$

In particular, we are interested in knowing whether the quantity X_l^s remains bounded as l approaches infinity. Secondly, if it remains bounded we wish to determine the least upper bound for each l. Since there are no restrictions on X_1 (other than the requirement of boundedness), we are interested in determining the norm in the limit of the operator K_l^s; or,

$$\lim_{l\to\infty} |K_l^s| = \lim_{l\to\infty} |(I - S)K_l| \tag{27}$$

when we know that

$$\lim_{l\to\infty} |K_l| \to \infty . \tag{28}$$

For physical systems we are also interested in $|K_l^s|$ for all l.

All of our results in this section depend upon an important lemma concerning the operator K_l^s. We assert that K_l^s has a representation simpler than the one given in (26) and prove this assertion by verification. Define an operator

$$B = \text{diag}\cdot\{(S_0 - 1)^{-1}, (S_1 - 1)^{-1}, (S_2 - 1)^{-1}, \cdots\}, \tag{29}$$

where $S_n \neq 1$ are defined in (11).* Then we assert the following *Lemma:*

$$K_l^s = \left[I - S\right]\left[\sum_{\nu=0}^{l-1} T^\nu\right] = B[I - T^l]T^{-1}. \tag{30}$$

Proof: Observe that in (30)

$$B(I - T^l)T^{-1} = B(I - T)T^{-1}\left[\sum_{\nu=0}^{l-1} T^\nu\right].$$

So, we need merely show that

$$[I - S] = B(T^{-1} - I).$$

Or,

$$B^{-1} - B^{-1}S = T^{-1} - I,$$

where

$$\begin{aligned} B^{-1} &= \operatorname{diag}\cdot\{(S_0 - 1), (S_1 - 1), \cdots\} \\ &= \operatorname{diag}\cdot\{S_0, S_1, S_2, \cdots\} - I. \end{aligned}$$

Thus we need to show that

$$\operatorname{diag}\cdot\{S_0, S_1, S_2, \cdots\} - B^{-1}S = T^{-1}.$$

But from (11),

$$(S_{n-1} - 1) = \beta^{i_n} S_n ,$$

so

$$B^{-1} = \operatorname{diag}\cdot\{\beta^{i_1}S_1, \beta^{i_2}S_2, \cdots\}.$$

Therefore, the lemma is proven if

$$T^{-1} = \begin{bmatrix} S_0 & -\beta^{i_1}S_1 & 0 & 0 & 0 & \cdots \\ 0 & S_1 & -\beta^{i_2}S_2 & 0 & 0 & \cdots \\ \cdots & 0 & S_2 & -\beta^{i_3}S_3 & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}. \tag{31}$$

The truth of the above statement is verified directly by considering the products $T^{-1}T = TT^{-1} = I$. The validity of (30) is thus proven.

* It is quite possible that the S_n approach unity. If any $S_n = 1$, it implies a finite pulse train, and the question is analyzed very simply in a finite dimensional space as was done for the periodic case. Such cases, however, do not give us information for the infinite pulse trains which are required for long chains of repeaters.

We can now state the following
Theorem: The operator K_l^s *is bounded for all* l *if* $\inf_n S_n \geqq \alpha > 1$.

Proof: From (30), we have

$$\begin{aligned} | K_l^s | &= | B(I - T^l)T^{-1} | \\ &\leqq | B | \, | I - T^l | \, | T^{-1} |. \end{aligned}$$

We know that $| T^l | = 1$ for all l and $| T^{-1} |$ is finite from (31). Hence $| K_l^s |$ is bounded if B is bounded, which it is if $\inf_n S_n \geqq \alpha > 1$. The theorem is thus proven.

Corollary:

$$\lim_{l \to \infty} | K_l^s | < \infty .$$

This follows trivially since the boundedness of K_l^s was proven independently of l. Our assumption that $\inf_n S_n \geqq \alpha > 1$ is a simple assertion of the fact that indefinitely long strings of zeros are ruled out on any communication channel.

Next, we wish to determine what precisely is the norm of the element

$$X_l^s = K_l^s X_1$$

as it relates to the norm of X_1. Let us recall that X_1 is the jitter injected (at a single repeater) referred to the output of the repeater. The injected jitter referred to the input of the repeater is

$$X = T^{-1} X_1 , \tag{32}$$

where T^{-1} is defined in (31). We find it more convenient to work with X in what follows. It is obvious, of course, that our entire discussion could have been carried out in terms of X from the very start. We chose not to do so in order to avoid a premature discussion of T^{-1}. The sequence represented by X naturally satisfies condition (2). In fact, the use of X allows a much simpler comparison of the spacing jitter at different repeaters along the chain. We are interested in the behavior of

$$X_l^s = [K_l^s T] X \tag{33}$$

for each l. This is obtained by a precise evaluation of the norm of

$$R_l^s = K_l^s T = B(I - T^l). \tag{34}$$

Let us first define $S_{\mathrm{inf}} = \inf_n S_n$. Then, we state the following

Theorem: The norm of R_l^s is given by

$$|R_l^s| = \frac{2}{S_{\mathrm{inf}}} \sum_{\nu=0}^{l-1} \frac{1}{S_{\mathrm{inf}}^{\nu}}. \tag{35}$$

Proof:

$$R_l^s = B(I - T^l).$$

Let us consider the representation of $(I - T^l)$. The diagonal elements of $(I - T^l)$ are of the form $[1 - (1/S_n^l)]$, whereas the off-diagonal elements are all negative. Also, the sum of the elements in each row must be zero. If we multiply $(I - T^l)$ on the left by B, then the diagonal elements are of the form $[1/(S_n - 1)][1 - (1/S_n^l)]$. Again the off-diagonal elements are all negative and the sum of the elements in each row of $B(I - T^l)$ is zero. Hence, in $\mathbf{1}_\infty$

$$|B(I - T^l)| = 2\left(\frac{1}{S_{\mathrm{inf}} - 1}\right)\left(1 - \frac{1}{S_{\mathrm{inf}}^l}\right)$$

$$= \frac{2}{S_{\mathrm{inf}}} \sum_{\nu=0}^{l-1} \frac{1}{S_{\mathrm{inf}}^{\nu}}.$$

The theorem is thus proven. Observe that this is not just a bound but a precise norm. This value in the norm is taken by the spacing jitter X_l^s for some X whose norm is one. In other words, the value in (35) represents the maximum magnification of X that is possible to yield the value of the spacing jitter. It is interesting to note that this worst case occurs for each l for the same element X, viz., $\{\cdots, 1, -1, -1, \cdots\}$, where the $+1$ corresponds to the position of $1/S_{\mathrm{inf}}$ in the matrix representing the operator T. Finally, to observe the maximum possible growth of the spacing jitter as it compares with the maximum possible spacing jitter at a single repeater, we compare the results for $l = 1$ and $l = \infty$.

$$R_1^s = \frac{2}{S_{\mathrm{inf}}} \tag{36}$$

$$R_\infty^s = \frac{2}{S_{\mathrm{inf}} - 1}. \tag{37}$$

The ratio of the quantities R_∞^s to R_1^s is less than two if S_{inf} is at least greater than two, which is to be expected in most physical situations. For example, if Q is of the order of 100 and there is at least one pulse present in fifteen time slots, then S_{inf} is at least two or greater. Thus, we observe that not only is there a bound on the growth of the spacing jitter but that the growth is monotonic and levels off rather fast. Of course, this is a rather general result which, if desired, can be more specifically stated in terms of the a priori probability distributions of the binary signal. Furthermore, the importance of this result lies in specifying the conditions for the validity of our model. Under most realistic situations it should be clear that conditions (2) are met at every repeater. Of course, there are situations when these conditions are not met and the results obtained by the use of our model [cf. (3)] can no more be relied upon. However, we show that under most situations the results are reliable and very optimistic, as shown by (36) and (37).

The above results are the most crucial ones in this paper. The rest of the paper is devoted to a varied miscellany that has some bearing on different aspects of the timing problem.

VI. ALIGNMENT JITTER

The alignment jitter in the $(l + 1)$th repeater is given by (17),

$$X_{l+1}{}^{A} = T^{l} X_{1}.$$

Obviously, for all l the alignment jitter is no greater in the norm than the absolute jitter X_1. This follows trivially since $| T^l |$ does not exceed one for any l.

For more detailed insight into the behavior of the alignment jitter, we need to discuss specific situations.

(a) If $\lambda = 1$ is a pole of T, then T^l converges to T^∞ and $X_{l+1}{}^A$ settles down to a flat delay element for large l.

(b) If $\lambda = 1$ is not a pole of T, and $\inf\limits_{n} S_n = 1$, T^l does not converge to T^∞. However, for large l the alignment jitter is slowly-varying.

(c) If $\lambda = 1$ is not a pole of T, and $\inf\limits_{n} S_n = \alpha > 1$, the alignment jitter (for large l) varies even more slowly than it does in (b).

All the above results follow from the properties of T. If $\lambda = 1$ is a pole of T, the result is obvious (cf. Part I). If it is not a pole of T, the results follow from the fine structure of the spectrum of T. For example, in situation (b), all points in the point spectrum except $\lambda = 1$ are poles of T, whereas this is not true in situation (c). The corresponding eigenvectors have different structures for the two cases (cf. Part II).

It follows that the situation of (a) is to be preferred over that of (c), which in turn is preferable to that of (b).

VII. NONIDENTICAL REPEATERS

This is a rather difficult matter to discuss with any great generality. What we hope to do here is to briefly indicate perhaps the most convenient formulation and to give some reasons for believing that the orders of magnitude of the jitter parameters are not changed for small differences in the repeaters.

There are essentially three possible ways in which the repeaters may differ: (*i*) injected jitter, (*ii*) Q of the repeater, and (*iii*) mistuning. These differences appear mathematically in terms of different operators, multiplicative coefficients in the power series, and so on. We examine each of these separately.

If we assume that the injected jitter differs slightly at each repeater, then we may write (5) as

$$X_l = TX_{l-1} + X_{\mathrm{av}} + \Delta_l , \tag{38}$$

where X_{av} is the average injected jitter and Δ_l represents the deviation from this average in the lth repeater. The norm of X_l differs at most from the previous case of identical repeaters by

$$| \Delta_l + T\Delta_{l-1} + T^2\Delta_{l-2} + \cdots + T^{l-1}\Delta_1 | \leqq \sum_{i=1}^{l} | \Delta_i |. \tag{39}$$

If the $| \Delta_i |$ are quite small, it is clear that the results will not be appreciably different from the previous ones.

If the Q's are different, then our basic operator is different for each repeater. It would be almost impossible to analyze such a case in general. However, we can make certain observations if we put

$$T_l = T + K_l , \tag{40}$$

where T_l is the operator representing the lth repeater, which is assumed to be an operator T perturbed by an operator K_l. It is reasonable to expect $| K_l | \ll 1$. Then (5) becomes

$$X_l = TX_{l-1} + X_1 + K_lX_{l-1} . \tag{41}$$

If $| K_l | \leqq \epsilon \ll 1$, then the norm of X_l does not differ from the previous results by more than

$$\left(\frac{\epsilon}{1 - \epsilon}\right) | X_1 | \approx \epsilon | X_1 |. \tag{42}$$

Again, we see that the results are not appreciably different.

In the case of mistuning, (5) takes essentially the same form as (38),

$$X_l = TX_{l-1} + X_A + \epsilon_l X_B , \tag{43}$$

where X_A is the injected jitter due to sources other than mistuning, ϵ_l is the mistuning in the lth repeater, and $\epsilon_l X_B$ is the jitter due to mistuning (cf. Part I). Then

$$X_l = \left[\sum_{\nu=0}^{l-1} T^\nu\right] X_A + \left[\sum_{\nu=0}^{l-1} \epsilon_{l-\nu} T^\nu\right] X_B . \tag{44}$$

The contribution due to X_A is unaltered and the contribution due to X_B would depend on the specifications of ϵ_i. However, if we assume that the magnitudes of ϵ_i do not exceed one, then the contribution in the norm due to X_B cannot exceed l times the norm of X_B. Thus the worst case for the absolute jitter does indeed arise from the assumption of equal ϵ_i at their maximum values.

For spacing jitter we can make a slightly different statement for the contribution due to X_B. For one repeater, the worst case (in the norm sense) occurs for the maximum value of $\epsilon_1 = 1$, then the worst case for two repeaters is obtained by setting $\epsilon_2 = 1$. Setting the first two repeaters with $\epsilon_1 = \epsilon_2 = 1$, the worst case for a string of three repeaters is obtained by setting $\epsilon_3 = 1$ and so on. The statement is a simple consequence of the inequality

$$|\,(I - T^{l-1})\,| < |\,(I - T^l)\,|. \tag{45}$$

It is believed that a similar statement can be made for the alignment jitter.

We thus observe that, when the differences in the repeaters are small, it is reasonable to expect that the results are not appreciably different from those obtained by assuming identical repeaters.

VIII. CONCLUSION

We have presented a general and rigorous theory of the jitter accumulation in a chain of regenerative repeaters. The sources of jitter are assumed to be the signal-dependent sources, as distinguished from purely random sources independent of the signal.

Our results show that while the absolute jitter and its dispersion grow without bound with the number of repeaters, the spacing and the alignment jitter remain bounded. In particular, the spacing jitter bounds are quite optimistic for most practical situations, viz., no greater than twice the absolute jitter injected at a single repeater. This result is of importance in that it ensures proper decoding of the binary signal. Its further importance lies in the fact that it does ensure, in most cases, the validity of the basic model and thus the validity of other results obtained by that model. One such result shows that the alignment jitter is slowly-varying for repeaters further along the chain. Finally, a brief discussion shows that the assumption of identical repeaters leads to results which are of the same order of magnitude as would be obtained if the repeaters differed by not too great an amount. Some results which would be of use in computations are to be found in the Appendix, together with an example.

In our discussion so far, we have investigated the accumulation properties of jitter due to repeated regeneration. We have made no attempt to determine the jitter introduced by a single repeater. Analytically, this problem is complicated not only by the nonlinearities involved, but also by a lack of complete knowledge as to the actual mechanisms involved. We propose instead, in a later paper, an experimental approach which allows these measurements to be carried out under steady-state conditions. This experiment also has some bearing on the question of simulation of long chains of repeaters.

IX. ACKNOWLEDGMENTS

I wish to take this opportunity to express my appreciation for the help I received in this entire endeavor from several of my colleagues. In particular, I wish to thank B. Liu and M. R. Aaron for many helpful discussions and criticisms. For their advice and encouragement, my thanks go to J. A. Young and M. Karnaugh.

APPENDIX

Eigenvalues and Eigenvectors of T

Practical systems call for the evaluation of jitter when the number of repeaters is finite. In such cases, we need not concern ourselves with infinite pulse trains. So long as the pulse trains are much longer than the effective memory of the system, the results obtained by considering finite pulse trains will be reliable. The results will also be reliable if the pulse trains are considered periodic with the period being greater than the memory of the system. Actually, as we shall see, the periodic pulse trains are much more difficult to work with than the finite ones. However, we shall discuss the periodic case in detail since much of the experimental work is carried out using periodic pulse trains. Finally, another case of interest is that in which there is a certain quiescent pattern which changes to a different one. This would include a periodic pattern changing to either a nonperiodic or a different periodic pattern. In each case, our interest is in determining the set of eigenvectors. Computations can then be carried out by expressing the injected jitter element in terms of the eigenvectors (cf. Part I). If the set of eigenvectors is not complete, we employ the standard procedure and use the so-called generalized eigenvectors. It should also be observed that the case of nonidentical repeaters is also handled using the same techniques. We present here certain simple algorithms for determining the eigenvalues and the eigenvectors for the several cases of interest.

A.1 *Truncated Pulse Trains*

This is the case where the pulse train is finite. The matrix T in (9) is now finite, upper triangular and the diagonal element in the last row is unity. Such a matrix also arises in the case of pattern transitions where originally the quiescent pattern is periodic with only one pulse present in each period. This is also the more realistic case because the tuned circuit is thus properly excited. We consider the truncated case, therefore, together with the pattern transition case.

A.2 *Pattern Transitions*

Here we have a steady-state periodic pattern which changes to a different pattern. The operator T can be represented as

$$T = \begin{bmatrix} P & C \\ 0 & A \end{bmatrix},$$

where P, representing the new pattern, is an upper triangular square matrix, A represents the quiescent periodic pattern, 0 is the null matrix, and C is the connecting matrix. The matrix A is either an arbitrary positive stochastic matrix for an arbitrary periodic pattern or it is a scalar (viz., unity) for the case of only one pulse present in each period. The latter case also occurs when the tuned circuit is excited by a reference pulse train such as 101010 $\cdots$.

In either case, the eigenvalues of T are given by the eigenvalues of P and those of A. The eigenvalues of the matrix A are discussed in the section dealing with the periodic case. The eigenvalues of P are given by the diagonal elements S_n^{-1}. If A is a scalar, the only other eigenvalue is unity. All the eigenvalues are thus determined.

Next, we observe that the eigenvalues of P are distinct. If not, for some n (say, $n = 0$),

$$S_0 = 1 + \beta^{i_1} + \beta^{i_1+i_2} + \cdots + [\beta^{i_1+i_2+\cdots+i_m}]S_0,$$

which implies a steady-state periodic pulse train, contradicting the transient nature of P. Thus the eigenvalues of T are distinct unless, of course, there is a periodic pattern involved. Let us reserve the periodic case for the next section. Then, the eigenvectors of T form a basis for the space of jitter vectors both for the truncated case and the pattern transition case when the quiescent pattern has only one pulse present in each period.

The eigenvectors are given by the algorithm

$$\xi_{n+1} = \frac{S_n - \frac{1}{\lambda}}{S_n - 1} \xi_n$$

(cf. Part II) for each eigenvalue λ.

The eigenvalues and the corresponding eigenvectors are thus very simply determined when either the reference pattern has only one pulse in each period or the pulse train is finite.

A.3 *Periodic Patterns*

Let us start by assuming that the period is m and there are n pulses in a period. Then

$$m = i_1 + i_2 + \cdots + i_n .$$

Let $\alpha_0 = (1 - \beta^m)$, and $D_i = S_i\alpha_0$. Then,

$$A = \begin{bmatrix} \frac{1}{D_0} & \frac{\beta^{i_1}}{D_0} & \frac{\beta^{i_1+i_2}}{D_0} & \cdots & \frac{\beta^{i_1+\cdots+i_{n-1}}}{D_0} \\ \frac{\beta^{i_2+\cdots+i_n}}{D_1} & \frac{1}{D_1} & \frac{\beta^{i_2}}{D_1} & \cdots & \frac{\beta^{i_2+\cdots+i_{n-1}}}{D_1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{\beta^{i_n}}{D_{n-1}} & \frac{\beta^{i_n+i_1}}{D_{n-1}} & \cdots & \cdots & \frac{1}{D_{n-1}} \end{bmatrix},$$

$$\begin{aligned} \Delta &= \det(\lambda I - A) \\ &= \frac{1}{D_0 D_1 \cdots D_{n-1}} \begin{vmatrix} \lambda D_0 - 1 & -\beta^{i_1} & \cdots & -\beta^{i_1+\cdots+i_{n-1}} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ -\beta^{i_n} & -\beta^{i_n+i_1} & \cdots & \lambda D_{n-1} - 1 \end{vmatrix}. \end{aligned}$$

$$\Delta = \frac{1}{D_0 D_1 \cdots D_{n-1}}$$

$$\cdot \begin{vmatrix} (\lambda D_0 - \alpha_0) & -\beta^{i_1} D_1 \lambda & 0 & 0 \cdots & 0 \\ 0 & (\lambda D_1 - \alpha_0) & -\beta^{i_2} D_2 \lambda & 0 \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -\beta^{i_n} & -\beta^{i_n + i_1} & \cdots & \cdots & (\lambda D_{n-1} - 1) \end{vmatrix}$$

$$= \Bigg[-\frac{1}{D_0} \beta^m \lambda^{n-1} + \left(\lambda - \frac{\alpha_0}{D_0}\right)\left(-\frac{1}{D_1} \beta^m \lambda^{n-2}\right) + \cdots$$

$$+ \left(\lambda - \frac{\alpha_0}{D_0}\right)\left(\lambda - \frac{\alpha_0}{D_1}\right) \cdots \left(\lambda - \frac{\alpha_0}{D_{n-3}}\right)\left(-\frac{1}{D_{n-2}} \beta^m \lambda\right)$$

$$+ \left(\lambda - \frac{\alpha_0}{D_0}\right) \cdots \left(\lambda - \frac{\alpha_0}{D_{n-2}}\right)\left(\lambda - \frac{1}{D_{n-1}}\right)\Bigg]$$

$$= P(\lambda) + Q(\lambda),$$

where

$$P(\lambda) = \sum_{i=0}^{n-1} \left(\lambda - \frac{\alpha_0}{D_i}\right)$$

and

$$Q(\lambda) = \sum_{\nu=0}^{n=1} - \frac{1}{D_\nu} \beta^m \lambda^{n-\nu-1} \prod_{k=1}^{\nu} \left(\lambda - \frac{\alpha_0}{D_{\nu-k}}\right).$$

Finally, after some manipulation it can be shown that

$$\Delta = \left(\frac{\alpha_{0-1}}{\alpha_0}\right) \lambda^n + \frac{1}{\alpha_0} P(\lambda).$$

Thus the eigenvalues of A are given by the zeros of the polynomial

$$R(\lambda) = P(\lambda) - \beta^m \lambda^n,$$

when

$$P(\lambda) = \prod_{i=0}^{n-1} (\lambda - S_i^{-1}).$$

The zeros of $R(\lambda)$ can be obtained from those of $P(\lambda)$ by root-loci considerations or other numerical methods. Since β^m is usually small, this can be handled easily on a digital computer.

Next, let us discuss the eigenvectors corresponding to the eigenvalues given by the zeros of $R(\lambda)$. We show that for each distinct eigenvalue of A, the eigenvector is given by the algorithm

$$\xi_{k+1} = \frac{S_k - \dfrac{1}{\lambda}}{S_k - 1} \xi_k .$$

This would be true if $\xi_{k+n} = \xi_k$ for all k or, if

$$\frac{\left(S_{n-1} - \dfrac{1}{\lambda}\right)\left(S_{n-2} - \dfrac{1}{\lambda}\right) \cdots \left(S_0 - \dfrac{1}{\lambda}\right)}{(S_{n-1} - 1)(S_{n-2} - 2) \cdots (S_0 - 1)} = 1.$$

The above is true if

$$\begin{aligned} S_{n-1}S_{n-2} \cdots S_0 \left(\frac{1}{\lambda^n}\right) P(\lambda) &= \prod_{k=1}^{n} (S_{k-1} - 1) \\ &= \prod_{k=1}^{n} (\beta^{i_k} S_k) \\ &= \beta^m S_1 S_2 \cdots S_n \\ &= \beta^m S_0 S_1 S_2 \cdots S_{n-1} \qquad (\text{since } S_n = S_0), \end{aligned}$$

or if

$$P(\lambda) = \beta^m \lambda^n,$$

which is indeed true for every eigenvalue λ. For a repeated eigenvalue λ, if it exists, one must find a generalized eigenvector in the usual way.

We have thus given simple algorithms for determining the eigenvalues and the corresponding eigenvectors of A.

It should be mentioned that the algorithm for eigenvectors is the same as the one given above when A is a submatrix of T as in Section A.2.

A.4 *Example*

Let us consider a simple example to illustrate some of the points. Consider a system with repeaters having $Q = 100$ and signals having at least one pulse present in 10 time slots. Consider the case of a quiescent pattern (101010 ··· 10) suddenly changing to a new periodic pattern with one pulse in 10 time slots. We are interested in determining the behavior of the first pulse after the transition.

The operator T has the form

$$T = \begin{bmatrix} \frac{1}{1+\beta^{10}(1+\beta^2+\beta^4+\cdots)} & \frac{\beta^{10}}{1+\beta^{10}(1+\beta^2+\beta^4+\cdots)} & \cdots & \cdots \\ 0 & \frac{1}{(1+\beta^2+\beta^4+\cdots)} & \frac{\beta^2}{(\ \ )} & \cdots \\ 0 & 0 & \frac{1}{(1+\beta^2+\beta^4+\cdots)} & \cdots \\ \cdots & \cdots & \cdots & \end{bmatrix}.$$

Let the jitter vector $X_1 = \{\xi_0, \xi_1, \xi_1, \xi_1, \cdots\}$, where ξ_1 is the reference phase jitter for the quiescent pattern and $(\xi_0 - \xi_1)$ is the change due to the transition.

In our simple example it is clear that there are only two eigenvalues of T, viz., $\lambda_0 = (1 - \beta^2)/(1 - \beta^2 + \beta^{10})$ and $\lambda_1 = 1$. The eigenvectors corresponding to these eigenvalues are $e_0 = \{(\xi_0 - \xi_1), 0, 0, \cdots\}$ and $e_1 = \{\xi_1, \xi_1, \xi_1, \cdots\}$.

The absolute jitter obviously increases without bound with the number of repeaters. The dispersion remains bounded, since $\lambda_1 = 1$ is a pole of T in our simple case. In the limit, the dispersion is given by

$$X^D = \left(\frac{1}{1 - \lambda_0}\right) e_0 = \left(\frac{1 - \beta^2 + \beta^{10}}{\beta^{10}}\right) e_0 .$$

The spacing jitter in the limit is given by

$$\left(\frac{1 - \beta^2 + \beta^{10}}{\beta^{10}}\right) (\xi_0 - \xi_1)$$

in the zeroth position and zero elsewhere. Finally, the alignment jitter approaches e_1 in the limit.

The validity of the results is assured if

$$\left(\frac{1 - \beta^2 + \beta^{10}}{\beta^{10}}\right) (\xi_0 - \xi_1) \ll 10 \frac{\pi}{Q},$$

or if

$$(\xi_0 - \xi_1) \ll 10 \frac{\pi}{Q} \left(\frac{\beta^{10}}{1 - \beta^2 + \beta^{10}}\right) \approx \frac{\pi}{10}.$$

Thus, if the *jump* in the phase jitter, for a single repeater, due to the transition in pattern is much smaller than 18°, our results are valid. This requirement can be expected to be satisfied by most practical repeaters.[5]

REFERENCES

1. Sunde, E. D., Self-Timing Regenerative Repeaters, B.S.T.J., **36,** July, 1957, pp. 891–938.
2. Rowe, H. E., Timing in a Long Chain of Regenerative Binary Repeaters, B.S.T.J., **37,** Nov., 1958, pp. 1543–1598.
3. Kinariwala, B. K., Timing Errors in a Chain of Regenerative Repeaters, I and II, B.S.T.J., **41,** Nov., 1962, pp. 1769–1797.
4. Bennett, W. R., Statistics of Regenerative Digital Transmission, B.S.T.J., **37,** Nov., 1958, pp. 1501–1542.
5. Byrne, C. J., Karafin, B. J., and Robinson, D. B., Jr., Systematic Jitter in a Chain of Digital Regenerators, B.S.T.J., **42,** Nov., 1963, pp. 2679–2714.
6. Rice, S. O., unpublished work.
7. Aaron, M. R., and Gray, J. R., Probability Distribution for the Phase Jitter in Self-Timed Reconstructive Repeaters for PCM, B.S.T.J., **41,** Mar., 1962, pp. 503–558.
8. Delange, O. E., The Timing of High-Speed Regenerative Repeaters, B.S.T.J., **37,** Nov., 1958, pp. 1455–1486.
9. Delange, O. E., and Pustelnyk, M., Experiments on the Timing of Regenerative Repeaters, B.S.T.J., **37,** Nov., 1958, pp. 1487–1500.

The Maser Rate Equations and Spiking

By D. A. KLEINMAN

(Manuscript received March 6, 1964)

The rate equations of Statz and De Mars giving the time development of the inversion and photon number in a maser or laser are discussed analytically with the aid of a mechanical analogy in which a particle moves in a potential well under the influence of a viscous damping force. The coordinate of this particle is analogous to the logarithm of the light output of the laser, and the amplitude, period, and damping of the motion can be directly related to the parameters of the rate equations. Simple analytic approximations are developed for all of the quantities of experimental interest in the spiking pattern of a laser. Four relationships are given, which do not contain any of the rate equation parameters, whereby a spike pattern can be tested to determine if it is consistent with the usual rate equations. Systematic procedures are described for extracting all of the information contained in spike patterns.

I. INTRODUCTION

The most fruitful approach for the discussion of maser and laser[1] behavior has been through rate equations describing the time rates of change of the atomic populations and the photon numbers of the electromagnetic field. Bloembergen[2] introduced rate equations for the populations in a paramagnetic maser and based his discussion on the steady-state solutions without explicitly considering the photon field. On the other hand, Shimoda, Takahasi, and Townes[3] have considered the photon rate equations and on this basis have given a theory of maser amplification without explicitly considering the atomic populations. Statz and De Mars[4] have shown that the transient behavior of masers depends upon coupled rate equations for both the populations and the photons. A number of authors have rederived these equations and discussed their applications to various maser systems. Considerable attention has been given to the question of whether these equations have periodic (undamped) solutions. It has been shown by Makhov[5] and by Sinnett[6] that the small-signal solutions are always damped, and it has been presumed and often confirmed by numerical computations that the same is true in the large-signal domain. Statz, et al.[7] have suggested that the damping of experimentally observed "spikes" in the output of lasers is probably sensitive to coherence and noise conditions, and not necessarily related to the damping predicted by the usual rate equations. They also suggest that the complicated spiking patterns frequently observed[8] are due to oscillation in many modes of the laser cavity. Despite these doubts which have been cast on the adequacy of the Statz-De Mars rate equations for describing laser behavior, they still remain the logical starting point for any discussion of the power output of lasers, whether it be transient or steady-state, and of the dependence of power on the quality of the cavity, the intensity of the pumping light, or the linewidth, relaxation time, and concentration of the active atoms.

Several ingenious suggestions have been made for modifying the rate equations so as to obtain periodic solutions. Statz and De Mars[4] and Makhov[5] propose that periodic solutions are obtained if terms are added to the rate equations representing cross relaxation in the inhomogeneously broadened maser transition. On the other hand, Shimoda[9] suggests that periodic spiking can result if the losses due to absorption in the cavity can be partially saturated by the buildup of laser oscillations. Although these and other modifications may ultimately prove to be justified and necessary in laser theory, we shall confine our attention in this paper to the original Statz-De Mars equations which relate the photons in a single cavity mode to a single quantity, the *inversion*, describing the atomic populations. We shall make it our task to understand as fully as possible the damped oscillatory solutions of these equations and how they may be applied to the study of the spiking phenomenon seen[8] in solid-state lasers.

Rate equations in the simple form have been successfully applied by McClung and Hellwarth[10] and by Vuylsteke[11] to the giant-pulse laser, which produces a single very short and very intense burst of radiation. Wagner and Lengyel[12] have shown that an exact analytic solution can be obtained to a simplified rate equation which neglects spontaneous emission and pumping during the pulse. By also neglecting the loss of photons in the cavity Dunsmuir[13] has obtained a still simpler analytic solution which is applicable to the rising portion of a spike or a giant pulse. We shall not consider further these exact solutions or the giant-pulse laser in this paper, but confine ourselves to those solutions representing repetitive pulsations, or spikes, in the ordinary laser. It is in this field that the greatest need now exists for an analytical discussion of the solutions of the rate equations.

A number of authors have endeavored to put the rate equations on a firmer theoretical basis. Following Anderson[14] and Clogston,[15] who first described masers in terms of the density matrix, treatments using the density matrix to derive rate equations have been given by Fain, et al.,[16] Kaplan and Zier,[17] and Pao.[18] By considering directly, without explicit use of the density matrix, the correlation functions of the electromagnetic field which are measured in simple maser experiments, McCumber[19] has shown that the maser medium acts like a dielectric of negative conductivity; he concludes that the field and the dielectric satisfy rate equations of the usual form, providing that the populations change by a small fractional amount during a coherence time (reciprocal linewidth) of the atomic system. Another formal theory, based on a successive approximation approach to the quantum mechanical equations of motion, has been applied by Haken and Sauermann[20] to the frequency shifts and interactions of cavity modes in the laser. An extensive survey with bibliography of the early formal work has been given by Lamb.[21] In the present paper we shall not go into the theoretical basis for the rate equations, but confine ourselves entirely to the problem of solving the equations in the large-signal domain.

Despite the fact that the rate equations are generally accepted, and that the general nature of the solutions has been familiar from digital computer calculations for some time,[5,13,17,22] it is still quite inconvenient in any particular case to compare observed spiking patterns in solid-state lasers with predictions of the rate equations. It has been necessary either to use the small-signal solutions and hope that they are not too inaccurate in the large-signal domain, or to resort to machine calculations and try to fit three or more parameters to the data by trial and error. To be sure, much of the data on spiking is not amenable to analysis, consisting apparently of random spikes with widely varying amplitude, duration, and interval. Nevertheless, several laser systems are now known to give very regular pulsations of the type that might be consistent with the rate equations. Regular spiking patterns have been observed in $CaF_2:U^{+3}$ by Sorokin and Stevenson[23] and by O'Connor and Bostick,[24] and in $CaWO_4:N_d^{+3}$ by Johnson and Nassau.[25] Recently the effect has also been seen in a highly perfect ruby by Nelson and Remeika,[26] and in a confocal ruby by Johnson, et al.[27] More extensive studies of highly regular spiking have been reported by Gürs[28] and by Hercher.[29] Thus it is clear that good data on spiking patterns can be obtained, and it therefore becomes cogent to inquire into practical and convenient means for analyzing this data and obtaining information from it.

The information contained in spike patterns is of two distinct kinds, which we may call qualitative and quantitative. Qualitatively, we can determine quickly by means of relationships given here whether a spike pattern is *consistent* with the rate equations. If we determine that certain patterns are not consistent, we are spared the waste of time that would result from attempting to fit these patterns numerically by trial and error. It is not obvious at this writing that any of the regular patterns that have been reported are consistent, because the application of the consistency relationships requires a measurement of the ratio of the peaks to the valleys in the light output, and only the peaks are seen in pictures published so far. Thus we suggest that by extending spike studies to include the valleys new and interesting information can be obtained. It is to be expected that patterns which are regular but yet not consistent will turn out to be physically the most interesting of all, since they will point the way to new understanding of laser behavior. The quantitative information can only be obtained from patterns which are consistent, at least in some average sense, and consists in obtaining values for the physical constants which appear in the rate equations:

N = the *number* of active laser atoms in the optical cavity

t_r = the *relaxation* time of the upper laser level, usually due to spontaneous emission

t_p = the *photon* lifetime in the laser mode of the cavity

t_m = the *mode* time, the time for spontaneous emission into the laser mode

t_g = the *ground* state time, the time spent by an atom in the ground state before being excited by the pump

s = the *source* strength, the rate of production of laser photons by spontaneous emission, the pumping light, or any other noise source in the cavity, or any signal applied to the cavity. Although s may vary with time, it is convenient here to consider it with the constants.

In principle all of these quantities except s could be directly measured or calculated from independent measurements on the laser material, the cavity, and the pump. The spiking data would then serve as confirmatory evidence. In many cases, however, spike patterns may prove to be the most convenient method of measurement. The last quantity, s, is in some respects the most interesting. The first assumption would be that s is due entirely to spontaneous emission into the laser mode; if so, s could be calculated and comparison with the measured value would reveal the verity of the assumption. Experiments could be carried out with an external weak signal from a monochromator to test the response of the laser as observed in its spiking patterns. Thus we hope that the analysis given here will help to stimulate new experiments by making rate equation analysis more convenient for the experimentalist.

II. FORMULATION OF THE RATE EQUATIONS

The laser is a system consisting essentially of an *optical cavity* with very high Q in a few modes, low Q in all other modes, the *laser medium* containing the active atoms, and a *pump*, usually an intense light source, to excite the atoms into a broad band of excited states. It is a property of the laser medium that the atoms decay from these states in an extremely short time by nonradiative processes to one or more very sharp excited states, called the upper laser levels. The upper laser levels can decay radiatively to a sharp lower level, called the lower laser level, which may be the ground state. If the lower laser level is not the ground state, we shall assume that very rapid nonradiative decay processes return the atom to the ground state. Thus we may always neglect the population of the pumping band and of the lower laser level if the latter is not the ground state. The laser transition takes place from the lowest of the upper laser levels, but it may sometimes be necessary to take into account the populations of nearby levels in thermal equilibrium with this level. The statistical weights of the laser levels must also be taken into account.[30]

The rate equations may always be written in the form (p = photon number, n = inversion)

$$\frac{dp}{dt} = -\frac{p}{t_p} + \frac{pn}{t_m} + s \tag{1}$$

$$\frac{dn}{dt} = \frac{n_0 - n}{t_0} - a\frac{pn}{t_m} \tag{2}$$

as long as we consider only a single mode of the cavity, the laser mode, and neglect all atomic populations except the upper laser level and the ground state. The time t_0 might be called the pumping relaxation time

$$\frac{1}{t_0} = \frac{1}{t_g} + \frac{1}{t_r}, \tag{3}$$

since it represents the characteristic time in the response of the population inversion n to the pump. The inversion n may always be written

$$n = N_u - (1/w)N_l, \tag{4}$$

where N_u, N_l are the populations of the upper and lower laser levels respectively and w is the statistical weight of the lower relative to that of the upper laser level. In view of our assumptions, w will come in only when the lower laser level is the ground state. Let us suppose that the upper laser level is in thermal equilibrium[26,30] with certain other states not directly involved in the laser transition such that there is a temperature-dependent probability P that an excited atom is in the upper laser level. Then we have for a *three-level system,* in which the lower laser level is the ground state,

$$\begin{aligned} a &= P + \frac{1}{w} \\ n_0 &= PN\, t_0 \left(\frac{1}{t_g} - \frac{1}{Pwt_r}\right); \end{aligned} \tag{5}$$

and for a *four-level system,* in which the lower laser level is not the ground state, we have

$$\begin{aligned} a &= P \\ n_0 &= PN\, t_0/t_g\,. \end{aligned} \tag{6}$$

If the relaxation of the upper laser level is predominantly by spontaneous emission to the lower laser level, there is a simple relation between t_r, t_m, and the linewidth $\Delta\nu$ (cps) of the laser transition

$$t_m = (8\pi\nu^2 n_{\text{ref}}^3/c^3)V\, t_r \Delta\nu, \tag{7}$$

where V is the volume of the cavity and n_{ref} is the refractive index. This relation should not be taken too literally, since it takes no account of the anisotropy of the laser medium or the polarization properties of the transition, and V would have to be replaced by a suitable effective "optical volume" if the laser material does not fill the cavity. Nevertheless, it points out the important fact that t_m varies with temperature in the same way as $\Delta\nu$ and therefore is subject to control in spiking studies. Another constant subject to convenient control is t_g, since $1/t_g$ is proportional to the pump intensity. Even when flash lamps are used for pumping it is still approximately valid to assume t_g is constant, since the time constant for the flash will usually be much longer than the interval between spikes. To assure that this is true, it would be advantageous to have the spike pattern commence when the flash is at its maximum intensity. It is not valid to assume without investigation that $1/t_g$ is proportional to the total energy dissipated in the flash. Spiking data should always include a record of the flash as a function of time and an indication of when the spike pattern occurred. Also subject to experimental control is t_p, the photon lifetime in the laser mode of the cavity. Presumably the losses in a good laser cavity can be estimated rather reliably, so that t_p can usually be directly calculated. The total number N of active laser atoms can ordinarily be determined in a given sample, but N does not appear to be a convenient parameter to vary in laser experiments. The signal or source strength s is best determined from an analysis of the spiking data as described in this paper. Although $s = s(t)$ will in general be a function of time, it will be shown that the spiking pattern depends only on the value of s at the instant when the net losses in the laser mode vanish due to the buildup of inversion. We shall denote this time by t_1 and the critical inversion by n_1, where

$$n_1 = t_m/t_p \tag{8}$$

is the celebrated Schawlow-Townes[1] criterion for the buildup of laser oscillation.

We now introduce dimensionless variables and parameters; in terms of the variables

$$\begin{aligned} \tau &= t/t_p \\ \eta &= n(t_p/t_m) = n/n_1 \\ \rho &= p(a\, t_0/t_m), \end{aligned} \tag{9}$$

and the parameters

$$\begin{aligned} \omega &= t_p/t_0 \\ \xi &= n_0(t_p/t_m) = n_0/n_1 \\ \sigma &= s(a\, t_p t_0/t_m), \end{aligned} \tag{10}$$

the rate equations (1), (2) become

$$\dot{\rho} = \frac{d\rho}{d\tau} = \sigma - \rho + \rho\eta \tag{11}$$

$$\dot{\eta} = \frac{d\eta}{d\tau} = \omega(\xi - \eta - \rho\eta). \tag{12}$$

Here ρ represents the *photons,* η the *inversion,* and τ the *time,* while ω represents the *pumping rate,* ξ the *limiting inversion* toward which the pump is tending to drive the system, and σ the *source.* We see that there are really only three parameters in the rate equations; it follows that three relationships among the six relevant physical parameters with which we started can be obtained from spiking studies. Although we shall assume in our analysis that ω and ξ are constant, our results will provide a useful adiabatic approximation for the case of slowly varying ω, ξ. Typical values of the parameters for a ruby laser will be given in the discussion of a numerical example. For the present we need only mention that σ is relevant only in the initial growth of ρ prior to the onset of laser gain.

If we regard σ as a constant, the steady-state solution of (11) and (12) is

$$\begin{aligned} \rho_\infty &= \tfrac{1}{2}\{(\xi + \sigma - 1) + [(\xi + \sigma - 1)^2 + 4\sigma]^{\frac{1}{2}}\} \\ \eta_\infty &= \tfrac{1}{2}\{(\xi + \sigma + 1) - [(\xi + \sigma - 1)^2 + 4\sigma]^{\frac{1}{2}}\}. \end{aligned} \tag{13}$$

The sign of the radical is determined by the requirement that $\rho \geqq 0$. In the limit $\sigma \to 0$ we obtain two cases, depending on whether $\xi < 1$ or $\xi > 1$. For $\xi < 1$

$$\begin{aligned} \rho_\infty &\to \sigma/(1 - \xi) \\ &\sigma \to 0 \\ \eta_\infty &\to \xi(1 - \rho_\infty) \end{aligned} \qquad \xi < 1; \tag{14}$$

this is the case in which the limiting inversion n_0 is less than n_1 given by (8), and laser oscillation does not occur. For $\xi > 1$

$$\begin{aligned} \rho_\infty &\to \xi - 1 \\ &\sigma \to 0 \\ \eta_\infty &\to 1 \end{aligned} \qquad \xi > 1; \tag{15}$$

this describes the steady state of laser oscillation, which is usually approached through a series of sharp pulses in $\rho(\tau)$ called *relaxation oscillations*, or *spikes*.

III. THE NATURE OF THE SPIKING SOLUTIONS TO THE RATE EQUATIONS

In this section we shall consider the nature of the solutions to (11) and (12) when $\xi > 1$ and the initial conditions on ρ and η correspond to very few photons and a small inversion $n \ll n_1$. This may be contrasted with the situation in the giant-pulse laser in which immediately after switching $n \gg n_1$. We shall also assume the *spiking condition*

$$\omega\xi \ll 1, \tag{16}$$

which will be satisfied whenever spiking can be observed. It will be apparent later that when (16) is not satisfied there will be no spikes, but ρ will smoothly approach ρ_∞. This is the case in gas lasers, where the pumping rate $1/t_g$ has to be very high to overcome the high relaxation rate $1/t_r \sim 10^8 \text{ sec}^{-1}$.

In view of (16) and (12), $\eta(\tau)$ will increase slowly with time and $\dot{\rho}$ in (11) can be neglected; thus we have

$$\rho(\tau) \approx \bar{\rho}(\tau) = \sigma(\tau)/(1 - \eta(\tau)) \tag{17}$$

until τ approaches τ_1 and η approaches unity. The behavior of $\rho(\tau)$ is shown in Fig. 1 for the initial condition $\rho(0) = 0$. Initially ρ rises with slope $\sigma(0)$ and asymptotically approaches the adiabatic solution $\bar{\rho}(\tau)$ which it follows for a relatively long time until $\tau \rightarrow \tau_1$. It follows that the initial condition on $\rho(\tau)$ is unimportant. As τ passes through τ_1, where

$$\eta(\tau_1) = 1, \tag{18}$$

$\rho(\tau)$ remains finite and is no longer given by (17); we then leave the *adiabatic phase* and enter the *spiking phase* of the time development of $\rho(\tau)$. According to (11) the slope of $\rho(\tau)$ at τ_1 is $\sigma(\tau_1) \equiv \sigma_1$. Thus we construct an approximate solution by smoothly joining (17) to a line of slope σ_1 as shown in Fig. 1. It follows that

$$\rho(\tau_1) = \rho_1 = 2\sigma_1/(\omega\beta)^{\frac{1}{2}}, \tag{19}$$

where

$$\beta = \xi - 1. \tag{20}$$

This is our first important result. It shows that the source occurs in spiking theory only as a parameter, the single value σ_1. Thus we may drop the subscript on σ and let $\sigma_1 = \sigma$, a constant.

Once we enter the spiking phase, $\eta(\tau)$ remains close to unity, fluctuat-

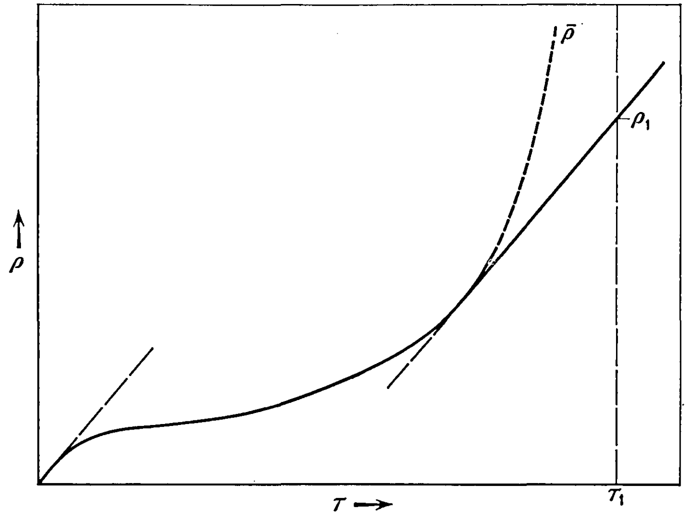


Fig. 1 — The adiabatic phase of the growth of $\rho(\tau)$. The curve is the adiabatic solution (17), and the line segments of slope σ are constructions to approximate $\rho(\tau)$ near $\tau = 0$ and $\tau = \tau_1$ where $\eta = \eta_1 = 1$.

ing up and down and eventually settling down to its steady-state value $\eta_\infty = 1$. Thus we make the approximations

$$\eta - 1 \approx \ln \eta, \tag{21}$$

and

$$\frac{\xi}{\eta} - 1 \approx \beta + \xi(1 - \eta). \tag{22}$$

The rate equations (11), (12) can now be written

$$\frac{d}{d\tau} \ln \rho = \ln \eta \tag{23}$$

$$\frac{d}{d\tau} \ln \eta = \omega(\beta - \xi \ln \eta - \rho). \tag{24}$$

Since the inversion is not directly observed we eliminate ln η from (23), (24); the result is most conveniently written

$$\ddot{\Psi} = \omega\beta(1 - e^{\Psi}) - \omega\xi\dot{\Psi} \tag{25}$$

in terms of the *logarithmic light output*

$$\Psi = \ln (\rho/\beta). \tag{26}$$

The discussion of (25) is greatly facilitated by a mechanical analogy which is shown in Fig. 2. We regard Ψ as the coordinate of a particle of unit mass moving in a one dimensional potential field.

$$\begin{aligned} V(\Psi) &= -\omega\beta \int_0^{\Psi} (1 - e^{\Psi})\, d\Psi \\ &= \omega\beta(e^{\Psi} - \Psi - 1). \end{aligned} \tag{27}$$

There is also a dissipative resistive force $\omega\xi\dot{\Psi}$ as if the particle were moving through a viscous medium. For the moment let us disregard the viscous force, in which case the total energy E of the particle is conserved

$$E = V(\Psi) + \tfrac{1}{2}\dot{\Psi}^2. \tag{28}$$

The particle executes a periodic motion between extreme points $\Psi_m < 0$ and $\Psi_M > 0$ such that

$$V(\Psi_m) = V(\Psi_M) = E. \tag{29}$$

In the spiking phase it is permissible to neglect exp (Ψ_m). From (27)

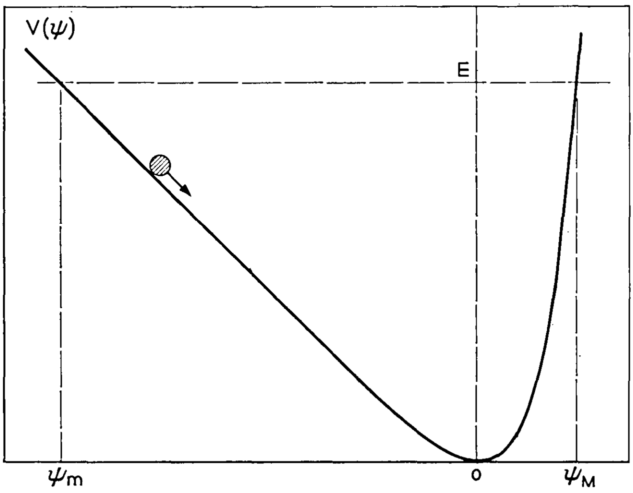


Fig. 2 — The spiking phase of the time development of $\rho(\tau)$ considered in terms of a mechanical analogy in which a particle with coordinate $\Psi = \ln (\rho/\beta)$ moves in a potential $V(\Psi)$ given by (27). The extremes of the motion Ψ_m and Ψ_M are determined by the total energy E of the particle.

and (29) we then obtain a simple relation between Ψ_m and Ψ_M

$$\Psi_M - \Psi_m = \exp (\Psi_M). \tag{30}$$

This is our second important result.

Let us define

$$\Psi_1 = \ln (\rho_1/\beta); \tag{31}$$

then it follows from (28) that

$$\Psi_m = \Psi_1 - (\sigma_1/\rho_1)^2/2\omega\beta, \tag{32}$$

since $\dot{\Psi}_1 = \sigma_1/\rho_1$. It will be clear from the numerical example in the next section that the second term can be neglected, and we can write

$$\Psi_m \approx \Psi_1 . \tag{33}$$

This says that the small kinetic energy of the particle at the start of the spiking phase can be neglected.

It is convenient to denote the successive times when $\eta(\tau) = 1$ by $\tau_1, \tau_2, \tau_3, \cdots$. From (23) we see that these times correspond to extrema in the motion of Ψ; according to this convention the successive minima and maxima of Ψ are

$$\begin{aligned} &\text{minima: } (\Psi_1), \Psi_3, \Psi_5, \Psi_7, \cdots \\ &\text{maxima: } \Psi_2, \Psi_4, \Psi_6, \cdots . \end{aligned} \tag{34}$$

We have placed Ψ_1 in parentheses, since it is a minimum only in the mechanical analogy and not in the observed light output. In the absence of damping, of course, we would have $\Psi_1 = \Psi_3 = \Psi_5 = \cdots$, and $\Psi_2 = \Psi_4 = \Psi_6 = \cdots$. If (16) is satisfied, the damping will be small enough so that the motion is approximately periodic, or quasi-periodic. Thus to a good approximation we can compute the *first maximum* Ψ_2 from Ψ_1 by means of (30). Before considering the damping we shall consider other quantities of experimental interest which are characteristic of the periodic motion.

The *maximum velocity* $\dot{\Psi}_{\max} = \dot{\Psi}_0$ is capable of being measured experimentally from spiking patterns in which the spikes are well resolved in time. It follows from (28), (29) and (33) that

$$\begin{aligned} \dot{\Psi}_{\max} = \dot{\Psi}_0 &= [2\omega\beta(e^{\Psi_1} - \Psi_1 - 1)]^{\frac{1}{2}} \\ &\approx [2\omega\beta(-\Psi_1 - 1)]^{\frac{1}{2}}. \end{aligned} \tag{35}$$

In numerical applications this can be used to ascertain the validity of (21), since according to (23)

$$\dot{\Psi}_0 = \ln \eta_{\max} \approx \eta_{\max} - 1. \tag{36}$$

In general we can write near Ψ_1

$$\begin{aligned} \dot{\Psi} &= \{2[E - V(\Psi)]\}^{\frac{1}{2}} \\ &\approx [2\omega\beta(\Psi - \Psi_1)]^{\frac{1}{2}}; \end{aligned} \tag{37}$$

thus the time dependence near a minimum is given by

$$\Psi(\tau) \approx \Psi_1 + \tfrac{1}{2}\omega\beta(\tau - \tau_1)^2. \tag{38}$$

Let us denote by m the full width in time of the minimum measured between points e times the minimum in light output; this is the same as the full width of $\Psi(\tau)$ measured between points $\Psi_m + 1$. Thus the *duration of the minima* according to (38) is

$$m = (8/\omega\beta)^{\frac{1}{2}}. \tag{39}$$

This applies to all minima regardless of damping, which provides a very convenient means for determining almost by inspection whether or not a spike pattern is consistent with the assumption of constant $\omega\beta$. Even if the minima are not observed to have the same durations, it may be meaningful, in the sense that our theory provides an adiabatic approximation, to apply (39) to each minimum separately and deduce the variation of $\omega\beta$.

Near the maximum Ψ_2 we write instead of (37)

$$\begin{aligned}\dot{\Psi} &\approx [2\omega\beta(e^{\Psi_2} - 1)(\Psi_2 - \Psi)]^{\frac{1}{2}} \\ &\approx [2\omega\beta(\Psi_2 - \Psi_1 - 1)(\Psi_2 - \Psi)]^{\frac{1}{2}},\end{aligned} \tag{40}$$

where use has been made of (30); thus

$$\Psi(\tau) \approx \Psi_2 - \tfrac{1}{2}\omega\beta(\Psi_2 - \Psi_1 - 1)(\tau - \tau_2)^2. \tag{41}$$

The *duration M of a maximum* will be defined as the time interval measured between points at $1/e$ times the maximum in light output; this is the same as the interval between points at $\Psi_M - 1$. We see from the factor $(\Psi_2 - \Psi_1 - 1)$ in (41) that M will depend upon damping; it may be written in a general way

$$M = m(\Psi_M - \Psi_m - 1)^{-\frac{1}{2}}. \tag{42}$$

There is in this relation a certain ambiguity which is inherent in our method of regarding the motion as quasi-periodic. In applying the relation we may wonder whether the minimum is the one preceding or following the maximum. Within the accuracy of the quasi-periodic approximation it makes no difference: either may be used, or the average of the two.

The most readily observed quantity in spiking experiments is the *interval* between spikes, which can be identified with the period of the quasi-periodic motion

$$I = \oint d\Psi/\dot{\Psi}, \tag{43}$$

where the integral is over one cycle. To evaluate I we use the approximations (37) and (40), which are accurate near the turning points Ψ_1 and Ψ_2 respectively where $1/\dot{\Psi}$ is large. Upon comparing (35) and (37) we see that (37) is reasonably accurate even at $\Psi = 0$ providing $|\Psi_1| \gg 1$. However, (40) is only accurate near Ψ_2, say in the range

$$\Psi_2 - 1 < \Psi \leqq \Psi_2 .$$

Thus we shall use (37) in the range $\Psi_1 \leqq \Psi \leqq \Psi_c$ and (40) in the range $\Psi_c < \Psi \leqq \Psi_2$, where Ψ_c is a crossover point which will be determined presently. The integral (43) can now be carried out to obtain

$$I(\Psi_c) \approx m\left[(\Psi_c - \Psi_1)^{\frac{1}{2}} + \frac{(\Psi_2 - \Psi_c)^{\frac{1}{2}}}{(\Psi_2 - \Psi_1 - 1)^{\frac{1}{2}}}\right]. \tag{44}$$

Since both our approximations tend to underestimate $1/\dot{\Psi}$, we must choose Ψ_c so as to maximize I, which gives the condition

$$\Psi_c = \Psi_2 - 1. \tag{45}$$

Thus the *interval I between spikes* is given by

$$\begin{aligned} I &= m[(\Psi_M - \Psi_m - 1)^{\frac{1}{2}} + (\Psi_M - \Psi_m - 1)^{-\frac{1}{2}}] \\ &= M(\Psi_M - \Psi_m). \end{aligned} \tag{46}$$

It is logical in this case to choose the minimum between the two maxima between which I is measured. There is still an ambiguity, however, in the choice of maxima. Since our approximation tends to underestimate I, it is good to use the larger of the two maxima.

We now return to the equation of motion (25) and consider the damping force $-\omega\xi\dot{\Psi}$. If (16) holds, the damping will be small, and can be computed from the work done per cycle against the damping force.

$$W = \omega\xi \oint \dot{\Psi} d\Psi. \tag{47}$$

With damping present, energy is no longer conserved, but decreases by W every cycle of the motion until the particle eventually settles down at its equilibrium position $\Psi = 0$, corresponding to the steady-state light output given by (15). Near Ψ_1 we can neglect e^{Ψ}, so that (27) gives

$$\omega\beta(\Psi_3 - \Psi_1) = W. \tag{48}$$

We evaluate (47) just as we did (43), using (37) and (40) with a crossover point ψ_c, determined this time by the condition that W should be a minimum; the result is

$$W = (4\sqrt{2}/3)(\omega\xi)(\omega\beta)^{\frac{1}{2}}[(\Psi_M - \Psi_m - 1)^{\frac{3}{2}} + (\Psi_M - \Psi_m - 1)^{\frac{1}{2}}]. \tag{49}$$

Let us denote damping by the increment $\Delta\Psi_m$ between successive minima, or $\Delta\Psi_M$ between successive maxima. From (48) and (49) the general formula for damping of minima is

$$\Delta\Psi_m = (4\sqrt{2}/3)\omega\xi(\omega\beta)^{-\frac{1}{2}}[(\Psi_M - \Psi_m - 1)^{\frac{3}{2}} + (\Psi_M - \Psi_m - 1)^{\frac{1}{2}}]. \tag{50}$$

The choice of Ψ_m is ambiguous, but Ψ_M refers to the maximum between the two minima of $\Delta\Psi_m$. It is obvious from the shape of the potential $V(\Psi)$ shown in Fig. 2 that the maxima will be less damped than the minima. For small damping we have

$$\omega\beta(e^{\Psi_2} - 1)(\Psi_2 - \Psi_4) = W, \tag{51}$$

where W is now computed by integrating (47) from Ψ_2 around the cycle and back to Ψ_2 . The result is

$$\Delta\Psi_M = (4\sqrt{2}/3)\omega\xi(\omega\beta)^{-\frac{1}{2}}[(\Psi_M - \Psi_m - 1)^{\frac{1}{2}} + (\Psi_M - \Psi_m - 1)^{-\frac{1}{2}}] \quad (52)$$

with the familiar ambiguity in choice of Ψ_M . From (52), (46) and (39) we obtain the very convenient formula

$$\Delta\Psi_M = \tfrac{2}{3}(\omega\xi)I \quad (53)$$

relating the damping directly to the interval. We note that all ambiguity has disappeared from (53).

We now have a complete arsenal of formulas with which to attack experimental spiking patterns. Our formulas give all of the minima and maxima of Ψ as well as the durations and intervals and the maximum of $\dot{\Psi}$ in terms of the dimensionless rate equation parameters ω, ξ, σ and $\beta = \xi - 1$. These formulas are valid in the spiking phase where

$$\Psi_M - \Psi_m - 1 \gg 1. \quad (54)$$

As the spikes damp out the solution finally enters the *small-signal phase*

$$\Psi_M - \Psi_m \rightarrow 0. \quad (55)$$

The small-signal solution is well known,[5,6,13,22,29,31] so there is no need to discuss it here, but we give it for ready reference:

$$\begin{aligned}
\eta(\tau) &= 1 + Ae^{-\frac{1}{2}\omega\xi\tau}[\Omega \cos (\Omega\tau + \varphi) - \tfrac{1}{2}\omega\xi \sin (\Omega\tau + \varphi)] \\
\rho(\tau) &= \beta[1 + Ae^{-\frac{1}{2}\omega\xi\tau} \sin (\Omega\tau + \varphi)] \qquad (56) \\
\Omega^2 &= \omega\beta - \tfrac{1}{4}(\omega\xi)^2
\end{aligned}$$

where A is an arbitrary real small amplitude and φ is an arbitrary real phase. Exactly the same small-signal solution is obtained from (25), thereby justifying the approximation (22). It is now easy to see that when (16) breaks down the frequency Ω of the small-signal solution becomes imaginary and there are no oscillations. This may be the case in the gas laser, where we may have $\xi \sim 1$, $\beta \ll 1$, and $\omega > 4\beta$.

IV. A NUMERICAL EXAMPLE

Before attempting to apply our formulas to the analysis of spiking patterns we wish to discuss their accuracy with the aid of a machine calculation. For numerical integration the rate equations (11), (12) are best written in the form

$$\begin{aligned}
\dot{x} &= \omega(\xi - x - xe^{y}) \\
\dot{y} &= \sigma e^{-y} + x - 1,
\end{aligned} \qquad (57)$$

where

$$\begin{aligned} x &= \eta \\ y &= \ln \rho. \end{aligned} \tag{58}$$

As already pointed out following (17), the initial condition on x,y is not critical; it is convenient to start the solution on the adiabatic solution (17), so we take

$$y(0) = \ln \sigma, \qquad x(0) = 0. \tag{59}$$

The constants will have the values

$$\begin{aligned} \omega &= 7.12 \times 10^{-6} \\ \xi &= 5 \\ \sigma &= 2 \times 10^{-9} \end{aligned} \tag{60}$$

which are typical for a ruby laser. The numerical integration of (57) has been performed on the IBM 7090 computer using Hamming's[32] predictor-corrector method. The program provides for automatic halving and doubling of the integration interval under the control of the fractional error of the increment and a preselected tolerance (5×10^{-5}). The stability and accuracy of the program for this problem were checked using initial conditions which lead to the small-signal solution (56).

The results are shown in Fig. 3. The solid curve and scale on the right give $y(\tau)$, while the dotted curve and scale on the left give $x(\tau)$ for τ in the range 3×10^4 to 4×10^4. The origin for τ is completely arbitrary. From (26), (58) and (60) we have

$$y = \Psi + \ln \beta = \Psi + 1.387. \tag{61}$$

We are concerned primarily with y, since the inversion is not ordinarily observed experimentally. The main features of the computed results are summarized in Table I, which lists the values of τ_n, y_n, and m_n or M_n for $n = 1, \cdots, 9$ for the first nine extrema in the notation of (34). These values were obtained from the computed points by fitting a parabola to the three points nearest the extremum. The spacing of the computed points was sufficiently small ($\Delta\tau = 0.005 \times 10^4$) that in all cases all three points lay well within the validity of the parabolic approximations (38) or (41).

We shall now attempt to calculate the information of Table I by the formulas of the preceding section. In every case we shall indicate the

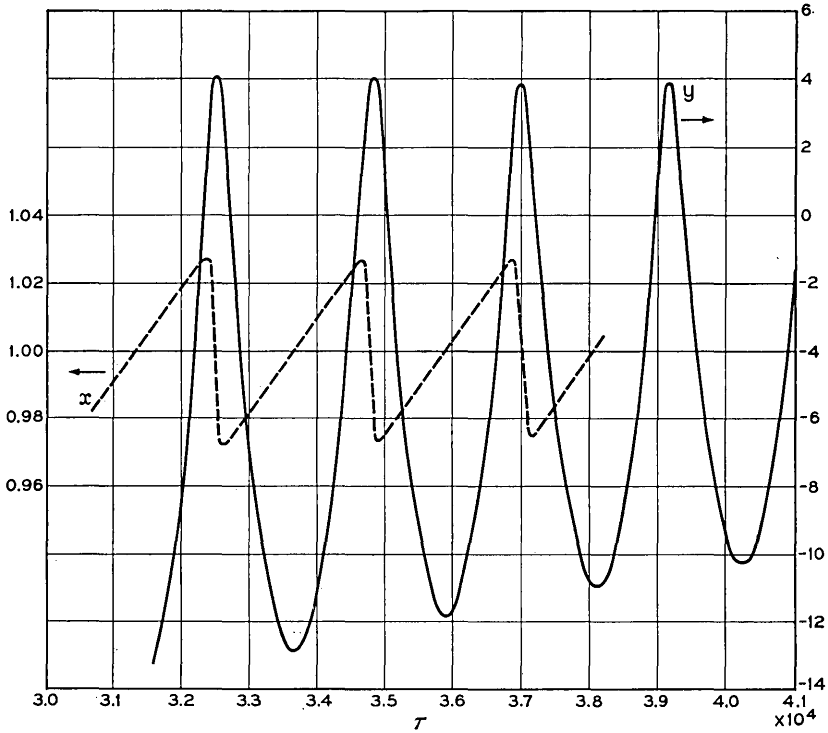


Fig. 3 — A machine calculation of $y(\tau)$, solid curve with scale on the right, and $x(\tau)$, dotted curve with scale on the left, satisfying the rate equations (57) for $\omega = 7.12 \times 10^{-6}$, $\xi = 5$, $\sigma = 2 \times 10^{-9}$. The important results are summarized in Table I. Here $x(\tau)$ represents the inversion and $y(\tau)$ the logarithmic light output.

Table I—Computed Results of Numerical Example

Summary of results of machine solution to the rate equations for $\omega = 7.12 \times 10^{-6}$, $\xi = 5$, $\sigma = 2 \times 10^{-9}$. Also obtained were $x_{\max} - 1 = 0.02830$, $\dot{y}_{\max} = 0.0282$.

n	τ_n	y_n	m_n	M_n
1	3.1344×10^4	−14.570	—	
2	3.2521	4.240		0.01416×10^4
3	3.3642	−12.815	0.0532	
4	3.4806	4.179		0.01415
5	3.5906	−11.795	0.0529	
6	3.7013	4.053		0.0151
7	3.8100	−10.986	0.0531	
8	3.9186	4.064		0.0152
9	4.0239	−10.303	0.0531	

correct value as computed by machine on the right in parentheses. According to (19)

$$\begin{aligned} \rho_1 &= 7.50 \times 10^5 \\ y_1 &= \ln \rho_1 = -14.09 \qquad (-14.57). \end{aligned} \tag{62}$$

Thus we *overestimate* y_1 by 0.48 or 3.3 per cent. From (61) the theoretical value for Ψ_1 is

$$\Psi_1 = -15.48. \tag{63}$$

Putting $\Psi_m = \Psi_1$ and $\Psi_M = \Psi_2$ in (30) gives

$$\Psi_2 = 2.91, \tag{64}$$

or

$$y_2 = 4.30 \qquad (4.240). \tag{65}$$

Thus we *overestimate* y_2 by 0.06 or 1.4 per cent. For the excursion of y we obtain

$$y_2 - y_1 = \Psi_2 - \Psi_1 = 18.39 \qquad (18.81), \tag{66}$$

which is an *underestimate* by 0.42 or 2.2 per cent. Henceforth we need not consider Ψ, but only y. From (50) using $\Psi_M - \Psi_m = y_2 - y_1$ we obtain

$$y_3 - y_1 = 0.965 \qquad (1.355) \tag{67}$$

$$y_3 = -13.12 \qquad (-12.815). \tag{68}$$

We are *underestimating* y_3 by 0.30 or 2.2 per cent. From (52) using $\Psi_M - \Psi_m = y_2 - y_3$ we obtain

$$y_2 - y_4 = 0.0541 \qquad (0.061) \tag{69}$$

$$y_4 = 4.25 \qquad (4.179). \tag{70}$$

By repeating steps (67) and (69) we could obtain values for any number of succeeding maxima and minima

$$y_5 - y_3 = 0.881 \qquad (1.020) \tag{71}$$

$$y_5 = -12.24 \qquad (-11.80) \tag{72}$$

$$y_4 - y_6 = 0.0523 \qquad (0.126) \tag{73}$$

$$y_6 = 4.19 \qquad (4.053). \tag{74}$$

We observe from Table I that $y_4 - y_6$ is unusually large (0.126), while $y_6 - y_8$ is negative (-0.011); the average of these is $\frac{1}{2}(y_4 - y_8) = (0.058)$ in better agreement with the machine-computed $y_2 - y_4$ and with theory. Thus we suspect that the discrepancy in (73) is not significant. There is apparently a slight inaccuracy in the computed solution on the third and fourth spike. One might expect that noise of various kinds would have a similar effect on real lasers, making the damping from spike to spike unreliable. Averaging over several spikes, however, would still give a meaningful value for Δy_M, as it does in our computed spike pattern.

The durations of all the minima should be given by (39)

$$m = 0.0526 \times 10^4 \qquad (0.0531 \times 10^4). \tag{75}$$

We see in Table I that m is constant within 0.4 per cent and the average over the four minima is 0.0531×10^4. The durations of the maxima will be calculated from (42), using for $\Psi_M - \Psi_m$ the values of the excursions just calculated, with the minimum preceding the maximum

$$\begin{aligned} M_2 &= 0.0127 \times 10^4 && (0.01416 \times 10^4) \\ M_4 &= 0.0130 && (0.01415) \\ M_6 &= 0.0134 && (0.0152). \end{aligned} \tag{76}$$

Here we have *underestimated* the spike durations by 11–12 per cent. The intervals are given by (46)

$$\begin{aligned} I_{2\text{-}4} &= 0.226 \times 10^4 && (0.2285 \times 10^4) \\ I_{4\text{-}6} &= 0.220 \times 10^4 && (0.2207 \times 10^4). \end{aligned} \tag{77}$$

The agreement here may be considered perfect. According to (53) we should have

$$\Delta y_M/I = 2.37 \times 10^{-5} \tag{78}$$

for all consecutive spikes; the computed results are as follows:

$$\begin{aligned} &2 \rightarrow 4 && (2.66 \times 10^{-5}) \\ &4 \rightarrow 6 && (5.70 \times 10^{-5}) \\ &6 \rightarrow 8 && (0.51 \times 10^{-5}) \\ &\tfrac{1}{2}[(4 \rightarrow 6) + (6 \rightarrow 8)] && (2.63 \times 10^{-5}). \end{aligned} \tag{79}$$

The last line is considered more significant than the discrepancies in the third and fourth lines. Thus our theory underestimates $\Delta y_M/I$ by

11 per cent. Finally, from (35)

$$\begin{matrix} \dot{y}_{\max} = x_{\max} - 1 = 0.0288 \\ (0.0282) \qquad (0.0283), \end{matrix} \tag{80}$$

which shows that the approximation (21) is excellent.

We may conclude from this confrontation between our formulas and a machine computation that the accuracy is probably adequate for the analysis of experimental spike patterns.

V. APPLICATION OF THE FORMULAS

Experimental data on spiking are ordinarily obtained in the form of an oscilloscope trace proportional to the light output. The trace is proportional to $p(t)$, the number of photons in the laser mode of the cavity, but it is usually considered impractical to calibrate the equipment so as to obtain the absolute value of $p(t)$. And even if $p(t)$ could be measured absolutely it would not fix the absolute value of $\rho(\tau)$ defined in (9). We shall assume that only relative values of $\rho(\tau)$ can be measured. The most significant measurements of ρ are the *peak-to-valley ratios*

$$R = \rho_M/\rho_m , \tag{81}$$

the ratio of the maximum light output of a spike to the minimum output in a neighboring minimum. In analyzing a spike pattern with many spikes, some convention must be adopted on which minimum to choose. From (26) and (61)

$$\Psi_M - \Psi_m = y_M - y_m = \ln R. \tag{82}$$

Thus the excursions of the model particle in the potential $V(\Psi)$ are uniquely fixed by the data. It follows from (30) that both Ψ_M and Ψ_m are fixed by R

$$\begin{aligned} \Psi_M &= \ln (\ln R) \\ \Psi_m &= \ln (\ln R) - \ln R. \end{aligned} \tag{83}$$

Even Ψ_1, the hypothetical minimum not actually observed, can be determined by extrapolating the other minima Ψ_m. Thus all of the extrema may be regarded as immediately fixed by the data. From the extrema alone we obtain 2 relations satisfied by the three parameters ω,ξ,σ, namely (19) and (50) or (52).

It is now clear that a single spike pattern does not contain enough information to determine ω,ξ,σ. The reason for this is that we cannot make use of the measured durations or intervals, since we do not know the ratio t_p between real time t and dimensionless time τ. We might think that we could obtain information from time ratios such as (M/m) or (m/I). It turns out that we can, but it is a different kind of information, the kind we have called "qualitative" in Section I. Consider the three quantities

$$A = (M/I) \ln R \tag{84}$$

$$B = (m/I)[(\ln R - 1)^{\frac{1}{2}} + (\ln R - 1)^{-\frac{1}{2}}] \tag{85}$$

$$C = (\Delta\Psi_M/\Delta\Psi_m)(\ln R - 1), \tag{86}$$

all of which can be determined immediately from the data since they depend only on ratios. It is to be expected that C might have to be averaged over several spikes to get a meaningful value. Except for this difficulty, values of A, B, and C may be calculated for every spike in the data after adopting some convention to handle the usual ambiguity in the definitions (84), (85), (86). One such convention is illustrated by the following example:

$$\begin{aligned} \ln R_2 &= \Psi_2 - \Psi_3 \\ A_2 &= (M_2/I_{2\text{-}4}) \ln R_2 \\ B_2 &= (m_3/I_{2\text{-}4})[(\ln R_2 - 1)^{\frac{1}{2}} + (\ln R_2 - 1)^{-\frac{1}{2}}] \\ C_2 &= [(\Psi_2 - \Psi_4)/(\Psi_3 - \Psi_1)](\ln R_2 - 1). \end{aligned} \tag{87}$$

From (42), (46), (50) and (52) we find the simple relations

$$A = B = C = 1 \tag{88}$$

which do not involve the parameters ω,ξ,σ. Therefore (88) should hold even if ω,ξ are slowly-varying functions of time within our adiabatic approximation. We call the relations (88) *consistency relations*, since they can be used to determine whether data are consistent with the rate equations. If (88) is satisfied reasonably well, at least in an average sense over the spikes, it is a foregone conclusion that a reasonable fit to the data can be obtained from the rate equations. The converse is also true: if (88) is not satisfied the data cannot be fitted from the rate equations.

The importance of observing the minima (valleys) as well as the maxima (peaks) in a spike pattern is abundantly clear. Without the valleys we cannot determine R, m, or $\Delta\Psi_m$, all of which appear in the consistency relations. Therefore a great deal of information is lost unless the valleys can be seen above noise. This presents some difficulty, because ordinarily $R \gg 1$ is much too large to be measured from an oscilloscope trace that responds linearly to the light output and contains both the peaks and the valleys. In spike patterns the valleys usually just correspond to the noise level in the experiment. It is not our purpose to go into experimental details except to point out that with care the valleys should be observable. The basic experimental requirement is that the acceptance cone of the detector should correspond to the radiation cone of the laser so as to exclude extraneous light from the pump and the spontaneous emission of the laser medium. If it is not possible to measure the valleys, there is still one consistency relation that can be applied to the peaks alone. From (46), (82) and (83) we have

$$\begin{aligned} \ln R &= I/M \\ -\Psi_m &= (I/M) - \ln (I/M). \end{aligned} \tag{89}$$

The use of (46) is equivalent to assuming $A = 1$. Thus the minima can be calculated from I/M if we assume $A = 1$. We could obtain the durations m of the minima by putting $B = 1$. The damping of the minima would not be given very reliably by (89), but can be obtained from (86) by putting $C = 1$. Thus we can deduce R, m, and $\Delta\Psi_m$ from the peaks alone, but we lose all of our consistency relations (88). However, from (89), (42), (39) and (35) it follows that

$$D = 1, \tag{90}$$

where

$$D = (M\dot{\Psi}_{\max}/4)[(I/M) - 1]^{\frac{1}{2}}/[(I/M) - \ln (I/M) - 1]^{\frac{1}{2}} \tag{91}$$

can be determined from the peaks if the time resolution is good enough to give a good value for $(M\dot{\Psi}_{\max})$. If (90) is satisfied, it is probably good evidence that the laser obeys the rate equations, and a rate equation analysis is meaningful. However, if all that is desired is to apply the rate equations blindly to obtain quantitative information, it is not necessary to measure $M\dot{\Psi}_{\max}$. All of the quantitative information in a consistent spike pattern can be deduced from I, M, and $\Delta\Psi_M$.

We now consider practical ways of obtaining quantitative information from spike patterns. The *one-pattern method* is to measure t_p by an independent experiment. The measurement of the cavity losses has been discussed by several authors.[26,33] Suffice it to say here that t_p can be measured from the dependence of the threshold flash energy for producing laser action on the temperature and on losses deliberately introduced into the laser mode. Once t_p is measured all of the times m, M and I become known in dimensionless time. From (39)

$$\omega\beta = \omega(\xi - 1) = 8/m^2, \tag{92}$$

and from (53)

$$\omega\xi = \tfrac{3}{2}\Delta\Psi_M/I. \tag{93}$$

These equations can be solved for ω,ξ

$$\begin{aligned} \xi &= \gamma m/(\gamma m - 1) \\ \omega &= 8(\gamma m - 1)/m^2, \end{aligned} \tag{94}$$

where

$$\gamma = \tfrac{3}{16}(m/I)\Delta\Psi_M \tag{95}$$

is independent of t_p. Now σ can be obtained from (19) and (89)

$$\sigma = \tfrac{1}{2}\beta(\omega\beta)^{\frac{1}{2}}(I/M)e^{-(I/M)} \tag{96}$$

with $M = M_2$ and $I = I_{2\text{-}4}$. This procedure makes use of the duration m of the valleys but not the peak-to-valley ratio R. If only the peaks are observed we write instead of (92)

$$\omega\beta = \omega(\xi - 1) = (8/MI)/[1 - (M/I)]. \tag{97}$$

Solving (93) and (97) for ω,ξ gives

$$\begin{aligned} \xi &= \delta M/(\delta M - 1) \\ \omega &= 8(\delta M - 1)/(I - M)M, \end{aligned} \tag{98}$$

where

$$\delta = \tfrac{3}{16}[1 - (M/I)]\Delta\Psi_M \tag{99}$$

is independent of t_p. This kind of analysis can be applied to every spike in a spike pattern. It may be found that ξ and especially ω vary slowly from spike to spike, which is permissible within our adiabatic approximation. If the data satisfy the consistency relations (88), the same values of ω,ξ will be obtained from (94) and (98). The greatest weakness of this method is that (94) fails completely if $\gamma m \leqq 1$; (98) fails if $\delta M \leqq 1$. Thus a great deal depends on the accuracy of the formulas as well as that of the measurements of t_p and γ or δ. It follows that the method will fail whenever $\xi \gg 1$.

As an example we shall apply this method to the machine calculation considered in Section IV. We consider the solid curve of Fig. 3 to be the data, which implies that t_p is known and both the peaks and valleys have been studied. The relevant numbers have already been given in parentheses in (69), (75), (76) and (77); we repeat them here without parentheses

$$\begin{aligned} \Delta\Psi_M &= y_2 - y_4 = 0.061 \\ m &= 0.0526 \times 10^4 \\ M &= M_2 = 0.0142 \times 10^4 \\ I &= I_{2\text{-}4} = 0.2285 \times 10^4. \end{aligned} \tag{100}$$

From (94), (95) and (96) we obtain

$$\begin{aligned} \xi &= 3.4 && (5.0) \\ \omega &= 12 \times 10^{-6} && (7.12 \times 10^{-6}) \\ \sigma &= 10 \times 10^{-9} && (2 \times 10^{-9}). \end{aligned} \tag{101}$$

The lack of accuracy is primarily due to the fact that $\xi = 5$ is a little too large to give good results with this method. Using only the peaks, we obtain from (96), (98) and (99)

$$\begin{aligned} \xi &= 2.9 && (5.0) \\ \omega &= 14 \times 10^{-6} && (7.12 \times 10^{-6}) \\ \sigma &= 8 \times 10^{-9} && (2 \times 10^{-9}). \end{aligned} \tag{102}$$

We now describe a *two-pattern method* which does not require the measurement of t_p. In fact, it yields a value for t_p and may in some cases give better results for ω,ξ than the one just described. It is based upon observing the valleys in two or more spike patterns for which the *relative* values of ω and ξ are known. We must assume that the linewidth $\Delta\nu$ is known as a function of temperature. Let us suppose that we measure the valley durations m and m' in two spike patterns in which the ratios (ξ'/ξ) and (ω'/ω) are known. We know (ξ'/ξ) from the temperatures at which the patterns were obtained. We can obtain ω'/ω from the relative pump intensities at the times when spiking occurred. It immediately follows from (92) that

$$\xi = \frac{(\omega m^2/\omega' m'^2) - 1}{(\omega m^2/\omega' m'^2) - (\xi'/\xi)}\,. \tag{103}$$

This result is meaningful providing that the data give $(\omega m^2/\omega' m'^2)$ lying between (ξ'/ξ) and unity. Once ξ is determined ω can be calculated from

(92) and (93)

$$\omega = \frac{9}{32}\frac{(\xi - 1)}{\xi^2}\left(\frac{m}{I}\right)^2 (\Delta\Psi_M)^2, \tag{104}$$

and σ is again obtained from (96).

As an example of this method we shall apply it to two machine-calculated spike patterns, one of which is that of Fig. 3 with the parameters (60), and the other has the parameters

$$\begin{aligned} \omega' &= \omega \\ \xi' &= 4 \\ \sigma' &= \sigma. \end{aligned} \tag{105}$$

From the machine-calculated pattern we find

$$m' = m_3' = 0.0614 \times 10^4. \tag{106}$$

The value of $m = m_3$ is given in (75). We assume that the ratio

$$\xi'/\xi = 0.8 \tag{107}$$

is known from the temperatures at which the data were taken, and the ratio

$$(\omega m^2/\omega' m'^2) = 0.748 \tag{108}$$

is calculable from (ω/ω') and the two valleys. We now obtain from (103) the value

$$\xi = 4.85 \qquad (5.0). \tag{109}$$

From (104) we now obtain

$$\omega = 9.0 \times 10^{-6} \qquad (7.12 \times 10^{-6}) \tag{110}$$

using the values in parentheses from (75), (77) and (69). From (39), (109) and (110) we obtain the *absolute value* of m

$$m = 0.048 \times 10^4 \qquad (0.0531 \times 10^4). \tag{111}$$

It follows that t_p is given by

$$t_p = m_{\text{exp}}/m, \tag{112}$$

where m_{exp} is the measured duration in laboratory time. We conclude that the two-pattern method is to be preferred to the one-pattern method as a general approach to spike analysis. It should be mentioned that there seems to be no two-pattern method involving only the peaks which is sufficiently accurate to give meaningful results.

The most convenient experiments to perform involve changing the pump intensity $1/t_g$ while all other parameters remain fixed. If $t_g \ll t_r$, as is usually the case in flash-pumped lasers, ξ will be independent of t_g and ω will vary as $1/t_g$. Varying ω does not give quantitative information such as we obtained from varying ξ in the two-pattern method. Nevertheless, it is of interest to consider what effects are to be expected. From (30), (26) and (19) we have roughly

$$\begin{aligned} \rho_M &\sim -\beta\psi_m \\ &\sim \beta \ln \left(\beta\sqrt{\omega\beta}/2\sigma\right). \end{aligned} \tag{114}$$

Thus the maxima in ρ depend only logarithmically on pumping power. It follows from (9) that the observed light output proportional to p should vary as

$$p \propto 1/t_g . \tag{115}$$

The time intervals m, M, and I should vary as

$$m \propto M \propto I \propto t_g^{\frac{1}{2}}. \tag{116}$$

VI. SUMMARY

We have now outlined a comprehensive program for the study of lasers by means of their spiking patterns. The rate equations have been formulated in terms of the light output and the atomic inversion and five physical parameters in (1) and (2). We have written the equations in a general form valid for three- and four-level systems and taking into account statistical weights and thermalization of the upper laser level. These equations were then reduced to dimensionless form in terms of three parameters in (11) and (12). All of the properties of the spiking solutions of experimental interest were then deduced analytically by means of the mechanical analogy shown in Fig. 2, in which a particle moves in a potential well in a viscous medium.

The formulas obtained were illustrated and tested for accuracy against a machine-computed solution to the rate equations. The value of the formulas was confirmed by this comparison, and we went on to discuss their application to experimentally observed spiking patterns. Four relations were given whereby spike patterns can be tested for consistency with the rate equations. These consistency relations do not contain any parameters, only ratios of times and of light outputs.

Two methods were described for obtaining quantitative information. In the one-pattern method everything is deduced from a single spike pattern, but it is necessary to measure the photon lifetime t_p in the cavity by independent experiments. It is possible to apply this method to data in which only the peaks are observed. It is emphasized, however, that the observation of the valleys in light output should be perfectly feasible and very much worthwhile. In the two-pattern method all the parameters and t_p are deduced from two patterns obtained at different temperatures. In this method, the more accurate of the two, it is essential that the valleys be observed.

Our objective has been to provide the researcher with a set of tools for applying the rate equations to experimental spiking data. The spiking phenomenon in solid state lasers can be utilized in research now that it is beginning to come under good experimental control. We have tried here to point out what kind of spiking data is needed, and what additional information is needed, to get the maximum information from spiking. Our analysis applies to the sharp spike region of time in which the rate equations are highly nonlinear. It complements the well known small-signal analysis in which the rate equations become linear. The entire discussion is based upon the rate equations of Statz and De Mars, which we regard as reasonable, relevant and widely accepted. We have not gone into the derivation of these equations or the modifications that have been proposed or might be proposed. We prefer to leave that to the future, when we may reasonably expect that systematic spiking studies will have clearly revealed the adequacy or inadequacy of these equations, and indicated the direction which these modifications must take.

VII. ACKNOWLEDGMENTS

This work actually began several years ago in a joint effort with R. J. Collins, now of the University of Minnesota. Circumstances intervened to prevent the completion of our collaboration, but I am pleased to acknowledge the great benefits that accrued to me through it. For interest, encouragement, and informative discussions I am also indebted to D. F. Nelson, W. K. Kaiser and D. E. McCumber of Bell Telephone Laboratories.

REFERENCES

1. We shall use the term "laser" to mean an optical maser as described by Schawlow, A. L., and Townes, C. H., Phys. Rev., **29,** 1958, p. 1940.
2. Bloembergen, N., Phys. Rev., **104,** 1956, p. 324.
3. Shimoda, K., Takahasi, H., and Townes, C., J. Phys. Soc. Japan, **12,** 1957, p. 686.

4. Statz, H., and De Mars, G., *Quantum Electronics*, Columbia University Press New York, 1960, ed. C. H. Townes, p. 530.
5. Makhov, G., J. Appl. Phys., **33,** 1962, p. 202.
6. Sinnett, D. M., J. Appl. Phys., **33,** 1962, p. 1578.
7. Statz, H., Luck, C., Shafer, C., and Ciftan, M., *Advances in Quantum Electronics*, Columbia University Press, New York, 1961, ed. J. R. Singer, p. 342.
8. Collins, R. J., Nelson, D. F., Schawlow, A. L., Bond, W., Garrett, C. G. B., and Kaiser, W., Phys. Rev. Letters, **5,** 1960, p. 303.
9. Shimoda, K., in *Symposium on Optical Masers*, Interscience Publishers, New York, 1963, ed. Jerome Fox.
10. McClung, F. J., and Hellwarth, R. W., J. Appl. Phys., **33,** 1962, p. 828; and Hellwarth, R. W., *Advances in Quantum Electronics*, Columbia University Press, New York, 1961, ed. J. R. Singer, p. 334.
11. Vuylsteke, A., J. Appl. Phys., **34,** 1963, p. 1615.
12. Wagner, W. G., and Lengyel, B. A., J. Appl. Phys., **34,** 1963, p. 2040.
13. Dunsmuir, R., J. Electronics and Control, **10,** 1961, p. 453.
14. Anderson, P. W., J. Appl. Phys., **28,** 1957, p. 1049.
15. Clogston, A. M., J. Phys. Chem. Solids, **4,** 1958, p. 271.
16. Fain, V. M., and Khanin, Y. I., Soviet Physics JETP (translation), **14,** 1962, p. 1069.
17. Kaplan, J., and Zier, R., J. Appl. Phys., **33,** 1962, p. 2372.
18. Yoh-han Pao, J. Opt. Soc. Am., **52,** 1962, p. 871.
19. McCumber, D. E., Phys. Rev., **130,** 1963, p. 675.
20. Haken, H., and Sauermann, H., Z. Physik, **173,** 1963, p. 261; **176,** 1963, p. 47.
21. Lamb, W. E., Jr., in *Lectures in Theoretical Physics II*, University of Colorado Summer School, 1959, ed. W. E. Brittin and B. W. Downs, Interscience Publishers, Inc., New York, 1960, p. 435.
22. Burch, J. M., in *Proceedings of the Conference on Optical Instruments*, 1961, ed. K. J. Habell, John Wiley, New York, 1963, p. 463.
23. Sorokin, P., and Stevenson, M., Phys. Rev. Letters, **5,** 1960, p. 557; also *Advances in Quantum Electronics*, Columbia University Press, New York, 1961, ed. J. R. Singer, p. 65.
24. O'Connor, J., and Bostick, H., Proc. I.R.E., **50,** 1962, p. 219.
25. Johnson, L. F., and Nassau, K., Proc. I.R.E., **49,** 1961, p. 1704.
26. Nelson, D. F., and Remeika, J., J. Appl. Phys., **35,** 1964, p. 522.
27. Johnson, McMahon, Oharek, and Sheppard, Proc. I.R.E., **49,** 1961, p. 1942.
28. Gürs, K., Z. Naturforsch, **17a,** 1962, p. 990; **18a,** 1963, p. 510; **18a,** 1963, p. 418; Gürs, K., and Müller, R., Physics Letters, **5,** 1963, p. 179.
29. Hercher, M. M., unpublished thesis, 1963, University of Rochester.
30. Collins, R. J., and Nelson, D. F., Ref. 22, p. 441.
31. Kaiser, W., Garrett, C. G. B., and Wood, D. L., Phys. Rev., **123,** 1961, p. 766.
32. Hamming, R. W., J. Assoc. Computing Machinery, **6,** 1959, p. 37.
33. Masters, J. I., Nature, **199,** 1963, p. 442; D'Haenens, I. J., and Asawa, C. K., J. Appl. Phys., **33,** 1962, p. 3201.

Design of Wideband Sampled-Data Filters

By R. M. GOLDEN and J. F. KAISER

(Manuscript received March 6, 1964)

A design procedure is presented for readily obtaining sampled-data filter representations of continuous filters. The procedure utilizes the bilinear z transformation and preserves the essential amplitude characteristics of the continuous filter over the frequency range between zero and one-half the sampling frequency. It is shown that the procedure can yield meaningful sampled-data filter designs for many of those filters where the standard z transform cannot be used directly.

I. INTRODUCTION

Sampled-data filter representations for continuous filters can be obtained using several different design procedures.[1] A particular design method utilizing the bilinear transformation is developed herein. The method is especially useful in designing wideband* sampled-data filters which exhibit relatively flat frequency-magnitude characteristics in successive pass and stop bands. Filters of this type are widely used in network simulation and data processing problems.[2] The design method possesses two chief advantages over the standard z transform.[3] The first is that the transformation used is purely algebraic in form. This means it can be applied easily to a continuous filter having a rational transfer characteristic expressed in either polynomial or factored form. The second advantage is the elimination of aliasing[4] errors inherent in the standard z transform. Thus, the sampled-data filter obtained by this design method exhibits the same frequency response characteristics as the continuous filter except for a nonlinear warping of the frequency scale. Compensation for this warping can be made by a suitable frequency scale modification. Some of the more common filter networks to which the design method can be applied effectively are the Butterworth, Bessel, Chebyshev, and elliptic filter structures.

The essential properties of the bilinear transformation are presented

* A sampled-data filter design will be termed "wideband" if the frequency range of useful approximation approaches half the sampling frequency.

in the next section. For comparison purposes, properties of the standard z transform are also given. This is followed by a detailed description of a filter design procedure using frequency transformations. Examples illustrating the design procedure are then presented. A short discussion concerning computer simulation of the obtained sampled-data filters is also included.

II. THE STANDARD AND BILINEAR z TRANSFORMATIONS

In this section it is assumed that a satisfactory rational expression is known for the transfer function of a continuous filter for which a sampled-data approximation is sought. What is then necessary is a transformation which will convert this transfer characteristic into a sampled-data transfer function rational in z^{-1}, the unit delay operator.

Two transformations applicable to this problem are the standard z transform and the bilinear z transformation.

The standard z transform applied to $H(s)$, the transfer function of the filter, is[3]

$$H^*(s) = \sum_{m=-\infty}^{\infty} H(s + jm\omega_s) \tag{1}$$

or equivalently in terms of the impulse response, $h(t)$, of the filter

$$\mathcal{H}^*(z) = T \sum_{l=0}^{\infty} h(lT) z^{-l} \tag{2}$$

where

$$\begin{aligned}
s &= \sigma + j\omega \\
H(s) &= \text{Laplace transform of } h(t) \\
\omega_s &= 2\pi/T = \text{radian sampling frequency} \\
H^*(s) &= \text{Laplace transform of the sampled filter impulse response} \\
z^{-1} &= \exp\,(-sT) = \text{the unit delay operator} \\
\mathcal{H}^*(z) &= H^*(s)\mid_{s=(\ln z)/T} = z \text{ transform of } h(t).
\end{aligned}$$

The behavior of $H(s)$ for s greater than some critical frequency $j\omega_c$ is assumed to be of the form

$$H(s)\mid_{\text{all } s > j\omega_c} = K/s^n, \qquad n > 0 \tag{3}$$

where K is a determined constant.

Equation (1) or (2) is the transfer function of a sampled-data filter which *approximates* the continuous filter. In the time domain, the impulse response of the sampled-data filter is the sampled impulse response of the continuous filter. This can be shown by taking the inverse transform of (2). Equation (1) shows that in the baseband

$$(-\omega_s/2 \leqq \omega \leqq \omega_s/2)$$

the frequency response characteristics of the sampled-data filter, $H^*(s)$, differ from those of the continuous filter, $H(s)$. The difference is the amount added or "aliased"[4] in through terms of the form

$$H(s + jm\omega_s), m \neq 0.$$

If $H(s)$ is bandlimited to the baseband, i.e., $| H(s) | = 0$ for $\omega > \omega_s/2$, then there is no aliasing error and the sampled-data filter frequency response is identical to that of the continuous filter. Unfortunately, when $H(s)$ is a rational function of s, it is not bandlimited and therefore $H(s) \neq H^*(s)$ in the baseband.

The magnitude of the errors resulting from aliasing is directly related to the high-frequency asymptotic behavior of $H(s)$ as defined in (3). If n is large and $\omega_c \ll \omega_s/2$, then the aliasing errors will be small and the standard z transform generally will yield a satisfactory sampled-data filter design. However, in wideband designs, ω_c is usually an appreciable fraction of $\omega_s/2$. Furthermore, many continuous filter designs result in transfer functions in which n is no greater than 1. These two conditions, namely $\omega_c \approx \omega_s/2$ and $n = 1$, can create large aliasing errors in the frequency response characteristics, thus yielding an unusable result.

Fortunately, even when $\omega_c \approx \omega_s/2$ and $n = 1$, a design method employing the bilinear z transformation* may provide satisfactory wideband designs. This z form is defined from the mapping transformation,

$$s = (2/T) \tanh (s_1 T/2) \tag{4}$$

where

$$s_1 = \sigma_1 + j\omega_1 .$$

The right-hand side of (4) is periodic in ω_1 with period $2\pi/T$. Considering only the principal values of ω_1, $-\pi/T < \omega_1 < \pi/T$, it is seen that the transformation given by (4) maps the *entire* complex s plane into the strip in the s_1 plane bounded by the lines $\omega_1 = -\pi/T$ and $\omega_1 = +\pi/T$. For this reason the bilinear transformation can be looked upon as a bandlimiting transformation. Therefore, when this transformation is applied to a transfer function $H(s)$, the entire s-plane frequency characteristics of $H(s)$ are uniquely carried over into the frequency characteristics of $H(s_1)$.

* This transformation will be referred to as the bilinear z form or z form.

With the substitution

$$z^{-1} = e^{-s_1 T},$$

(4) can be written immediately as

$$s = \frac{2}{T}\frac{(1 - z^{-1})}{(1 + z^{-1})}. \tag{5}$$

Thus,

$$\mathcal{H}(z) \equiv H(s)\Big|_{s = \frac{2}{T}\frac{(1-z^{-1})}{(1+z^{-1})}} \tag{6}$$

where $\mathcal{H}(z)$ denotes a sampled-data transfer function obtained by the use of the bilinear z form.*

The transfer function $\mathcal{H}(z)$ obtained by means of the bilinear z form and the function $\mathcal{H}^*(z)$ obtained by means of the standard z transform are both rational in z^{-1} and of the same denominator order as the continuous filter. These two functions, $\mathcal{H}(z)$ and $\mathcal{H}^*(z)$, become essentially equal to each other as the sampling frequency becomes large compared to the moduli of each of the poles of the continuous filter function, $H(s)$. When the sampling frequency is not large, representation of some filters by the standard z transform can be quite unsatisfactory because of aliasing errors. However, the bilinear z form, with its absence of aliasing errors, may give a satisfactory representation for these filters. A particular set of filters to which the bilinear z form always can be applied successfully are those which exhibit relatively flat frequency-magnitude characteristics in successive pass and stop bands. This follows directly from (6).

Thus, sampled-data filters designed by using the bilinear z form preserve the essential amplitude characteristics of the continuous filter. In the baseband $(-\omega_s/2 \leqq \omega \leqq \omega_s/2)$, the frequency characteristics of the sampled-data filter are identical to those of the continuous filter *except* for a nonlinear warping of the frequency scale.

This warping is found from (4) upon substituting $j\omega$ for s and is

$$\omega = (2/T)\tan(\omega_1 T/2). \tag{7}$$

For small values of ω_1, the relation is essentially linear, producing

* It should be noted that the bilinear transform is used here in a distinctly different way than it is commonly used in sampled-data control system design. In the control system literature it is used to transform the sampled-data function $\mathcal{H}^*(z)$ from the discrete domain back to the continuous domain for conventional stability and frequency response analysis. See for example J. T. Tou, *Digital and Sampled-Data Control Systems*, McGraw-Hill, New York, 1959, pp. 244–247 and pp. 466–470.

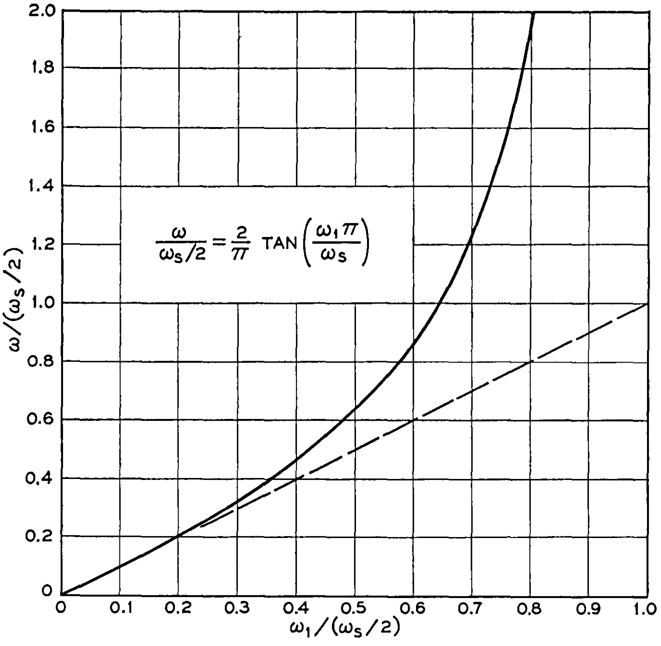


Fig. 1 — The frequency scale warping of the bilinear z transformation.

negligible warping at the lower end of the frequency scale. Fig. 1 shows the nature of this warping. Compensation for the effect of warping can be made by prewarping the band-edge frequencies of the continuous filter in such a way that application of the z-form transformation will shift the band-edge frequencies back to the desired values. The incorporation of this prewarping compensation into a sampled-data filter design procedure is discussed in the next section.

III. A SAMPLED-DATA FILTER DESIGN METHOD

A sampled-data filter design may be obtained by applying the z-form transformation of (6) to the rational transfer characteristic for a continuous filter. However, in order to compensate for the frequency warping imposed by the z form, the frequency characteristics of the continuous filter first must be altered or prewarped. Hence the transfer characteristic for the continuous filter must be redesigned such that the band-edge (cutoff) and maximum loss frequencies are computed according to,

$$\omega_c = (2/T) \tan (\omega_d T/2) \tag{8}$$

where:

$$\omega_c = \text{computed cutoff or loss frequency}$$
$$\omega_d = \text{desired cutoff or loss frequency.}$$

The redesign of the continuous transfer characteristic cannot be accomplished simply by applying (8) to each pole and zero of the original transfer characteristic. On the contrary a completely new transfer characteristic must be obtained for the continuous filter. It is then possible to obtain the desired sampled-data filter by applying the z-form of (6) directly to the redesigned transfer characteristic of the continuous filter. The sampled-data filter so obtained will then have the desired magnitude-frequency characteristics.

Compensation for frequency warping becomes especially simple to apply if the original continuous filter design was obtained by applying a frequency-band transformation to a suitable low-pass design such as Butterworth, Chebyshev, etc. Thus the extensive literature available on tabulated low-pass filter designs can be used to great advantage to simplify the filter design problem. The well-known frequency transformations which convert a normalized low-pass filter to a low-pass, a bandpass, a bandstop, or a high-pass design are

$$s_n = s/\omega_u \text{ low-pass to low-pass} \tag{9}$$

$$s_n = \frac{(s^2 + \omega_u\omega_l)}{s(\omega_u - \omega_l)} \text{ low-pass to bandpass} \tag{10}$$

$$s_n = \frac{s(\omega_u - \omega_l)}{(s^2 + \omega_u\omega_l)} \text{ low-pass to bandstop} \tag{11}$$

$$s_n = \omega_u/s \text{ low-pass to high-pass} \tag{12}$$

where:

s_n = the complex variable of the normalized low-pass filter transfer function

s = the complex variable of the desired filter transfer function

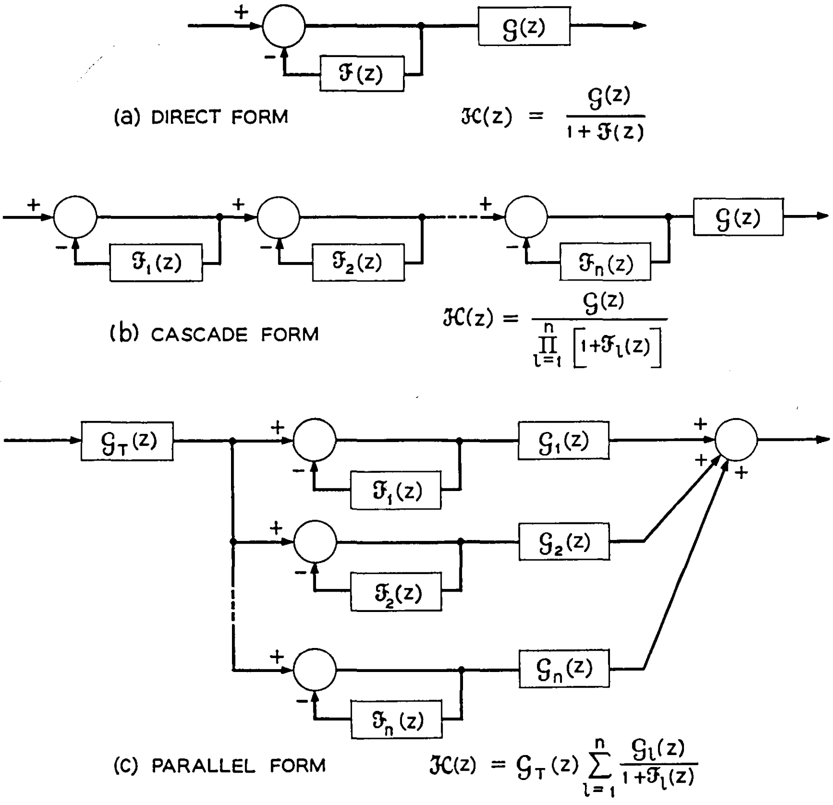
ω_u = the upper radian cutoff frequency

ω_l = the lower radian cutoff frequency.

When continuous filters are designed with the aid of these transformations, prewarping is accomplished by properly choosing the cutoff frequencies used in the frequency transformations. The choice of these cutoff frequencies is determined from the desired cutoff frequencies by means of (8). Using these values, the new prewarped transfer function is determined by applying (9), (10), (11) or (12) to the original low-pass function. Transformation is made to a sampled-data filter by applying (5) to the prewarped continuous function. This sampled-data filter will now have the correct cutoff frequencies. The transfer function thus obtained can be used directly in a digital computer simulation.

IV. SIMULATION OF SAMPLED-DATA FILTERS

Application of either the standard z transform or the bilinear z form to a rational transfer function yields a transfer function rational in z^{-1} for the sampled-data filter. The programming or simulation of this sampled-data filter on a digital computer can be accomplished by either the direct, the cascade or the parallel form. These forms, as commonly defined, are shown in Fig. 2. In this figure $\mathcal{G}(z)$ and $\mathcal{F}(z)$ represent finite polynomials in z^{-1} for feed-forward and feedback transmissions re-



+
−
$\mathcal{G}(z)$
$\mathcal{F}(z)$
(a) DIRECT FORM
$\mathcal{H}(z) = \frac{\mathcal{G}(z)}{1+\mathcal{F}(z)}$
$\mathcal{F}_1(z)$
$\mathcal{F}_2(z)$
$\mathcal{F}_n(z)$
(b) CASCADE FORM
$\mathcal{H}(z) = \frac{\mathcal{G}(z)}{\prod_{l=1}^{n}\left[1+\mathcal{F}_l(z)\right]}$
$\mathcal{G}_T(z)$
$\mathcal{G}_1(z)$
$\mathcal{G}_2(z)$
$\mathcal{G}_n(z)$
(C) PARALLEL FORM
$\mathcal{H}(z) = \mathcal{G}_T(z)\sum_{l=1}^{n}\frac{\mathcal{G}_l(z)}{1+\mathcal{F}_l(z)}$

Fig. 2 — Some possible simulation forms for sampled-data filters.

spectively; where subscripted, the order of the polynomial is at most second.

The choice of which of the three forms to use for simulation of the sampled-data filter depends on the complexity of the filter function $H(s)$, on the form of $H(s)$, and on the particular z transformation used. Generally, simulation by the direct form requires considerably greater accuracy in the determination of the filter parameters than either of the other two forms. This is especially true when the order of $H(s)$ is large and when $H(s)$ has poles with real parts that are a very small fraction of the sampling frequency. For this reason either the cascade or parallel form may be preferred.

The choice between using the cascade or the parallel form depends largely on which z-transform method is used to obtain the sampled filter and how that particular method is applied. Realization in the cascade form requires calculation of the numerator polynomial, $\mathcal{G}(z)$, or its factors. This computation consists of a simple algebraic substitution when the bilinear z form is applied to a filter function $H(s)$ expressed in the form,

$$H(s) = \frac{G(s)}{\prod_{k=1}^{m} (s - \alpha_k)}. \tag{13}$$

Determination of $\mathcal{G}(z)$ in polynomial or product form respectively allows either of the following cascade realizations to be made:

$$\mathcal{H}(z) = \mathcal{G}(z) \prod_{k=1}^{n} \left(\frac{1}{1 + b_{1k}z^{-1} + b_{2k}z^{-2}} \right) \tag{14}$$

or

$$\mathcal{H}(z) = \prod_{k=1}^{n} \left(\frac{a_{0k} + a_{1k}z^{-1} + a_{2k}z^{-2}}{1 + b_{1k}z^{-1} + b_{2k}z^{-2}} \right). \tag{15}$$

If the numerator $G(s)$ is in polynomial form, considerable care must be taken in the calculation of the coefficients of the polynomial $\mathcal{G}(z)$, as this computation involves differencing nearly equal numbers.

For realization in the parallel form the partial fraction expansion of $H(s)$ must be known. Since in the standard z transform method obtaining the partial fraction expansion is a necessary step, simulation of filters designed by this method is most directly accomplished in the parallel form. Here the continuous filter transfer characteristic is represented by

$$H(s) = \sum_{k=1}^{N} \frac{P_{1k}s + P_{0k}}{Q_{2k}s^2 + Q_{1k}s + Q_{0k}}. \tag{16}$$

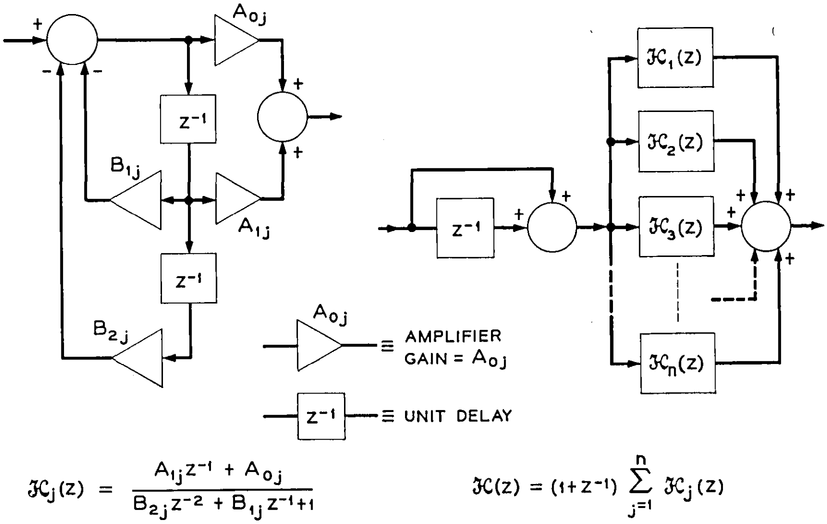
Transforming this expression by use of the standard z transformation yields

$$\mathcal{H}(z) = \sum_{k=1}^{N} \frac{A_{1k}z^{-1} + A_{0k}}{B_{2k}z^{-2} + B_{1k}z^{-1} + 1} \tag{17}$$

whereas transforming by use of the bilinear z form yields the similar expression

$$\mathcal{H}(z) = (1 + z^{-1}) \sum_{j=1}^{N} \frac{A_{1j}z^{-1} + A_{0j}}{B_{2j}z^{-2} + B_{1j}z^{-1} + 1}. \tag{18}$$

Each rational function in either of the above summations can be synthesized by the recursive structure shown in functional block diagram form in Fig. 3(a). This recursive structure uses only two delays, four multiplications, and five additions. The complete realization of (18) is shown in Fig. 3(b).



$$\mathcal{H}_j(z) = \frac{A_{1j}z^{-1} + A_{0j}}{B_{2j}z^{-2} + B_{1j}z^{-1} + 1}$$

(a)

TERM IN PARTIAL FRACTION EXPANSION

$$\mathcal{H}(z) = (1 + z^{-1}) \sum_{j=1}^{n} \mathcal{H}_j(z)$$

(b)

COMPLETE TRANSFER FUNCTION EXPRESSED AS A PARTIAL FRACTION EXPANSION

Fig. 3 — Simulation in parallel form of a sampled-data filter obtained by the bilinear z transformation.

The programming of these sampled-data filters for computer simulation can be greatly simplified if a compiler such as the Block Diagram (BLODI) compiler[5] developed at Bell Telephone Laboratories is used. The compiler permits specification of a sampled-data system in functional block diagram form.

In the following section an example is presented for a filter designed, synthesized, and simulated by the foregoing method.

V. DESIGN, SYNTHESIS, AND SIMULATION OF A WIDEBAND BANDSTOP SAMPLED-DATA FILTER

As an example of the application of the bilinear z-form to the simulation of a practical filter, consider the design of a particular bandstop filter. The filter is to exhibit at least 75 db loss in a rejection band which extends between 2596 cps and 2836 cps. Below 2588 cps and above 2844 cps, the loss is to be between 0 and +0.5 db. The sampling rate of the discrete filter is to be 10 kc. The complexity or order of the filter is to be held to a minimum. Fig. 4 shows a sketch of the amplitude response characteristics desired of the filter.

Minimum filter complexity and sharp transition between pass and stop bands suggest the use of an elliptic filter[6] (equiripple) as the basic low-pass type. However, before a suitable low-pass structure can be determined, the above specified critical frequencies must be prewarped by

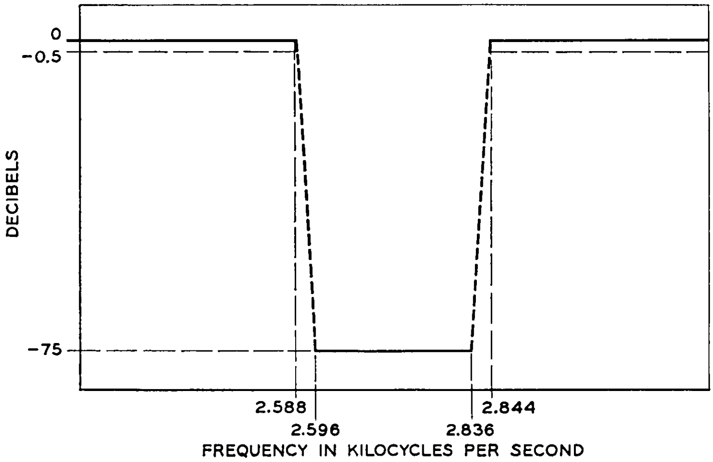


Fig. 4 — Desired amplitude response characteristic of a bandstop filter.

means of (8). The warped values are:

f_{lp} = lower cutoff frequency in passband (2588 cps) = 3364.15 cps
f_{lr} = lower cutoff frequency in rejection band (2596 cps)
= 3381.13 cps
f_{ur} = upper cutoff frequency in rejection band (2836 cps)
= 3937.54 cps
f_{up} = upper cutoff frequency in passband (2844 cps) = 3957.84 cps.

The warped values at the lower band edge require a low-pass filter with a transition ratio of 0.93792, while the values at the upper band edge require a transition ratio[6] of 0.93658. Therefore, to meet the original specifications, the larger of the two transition ratios must be chosen. Hence specifications required for the basic low-pass elliptic filter are:

in-band ripple = 0.5 db
out-of-band minimum attenuation = 75.0 db
transition ratio = 0.93792.

Application of elliptic filter design procedure with these specifications yields a basic low-pass structure of eleventh order. The poles and zeros for the transfer function of this low-pass filter are listed in Table I. The low-pass filter has been normalized to have a cutoff frequency of one radian per second and amplitude gain of unity at zero frequency.

TABLE I—POLES AND ZEROS OF NORMALIZED ELEVENTH-ORDER ELLIPTIC LOW-PASS FILTER

In-band ripple = 0.500 db	
Minimum attenuation = 76.504 db	
Transition ratio = 0.937917	
Gain factor = 0.0011060	
Poles	
	$-0.0069130 \pm j\,1.0010752$
	$-0.0257616 \pm j\,0.9756431$
	$-0.0615122 \pm j\,0.9063786$
	$-0.1269215 \pm j\,0.7504391$
	$-0.2142976 \pm j\,0.4483675$
	-0.2611853
Zeros	
	$\pm j\,1.0695414$
	$\pm j\,1.1009005$
	$\pm j\,1.1946271$
	$\pm j\,1.4652816$
	$\pm j\,2.5031313$

The desired bandstop filter is obtained next by applying the low-pass-to-bandstop transformation, given in (11), to the normalized elliptic low-pass filter. The cutoff frequencies used are the warped values obtained above for f_{lp} (3364.15 cps) and f_{up} (3957.84 cps). The resulting bandstop filter is then transformed by the bilinear z form to yield the required sampled-data filter. Table II lists the coefficients of the resulting sampled-data filter needed for parallel realization of the form shown in Fig. 3. Fig. 5 shows the frequency response characteristics of this sampled-data filter. It is seen that the original filter specifications are met by the sampled-data filter. For comparison purposes, the standard z transform was applied directly to the twenty-second-order continuous filter. The frequency response characteristic of this filter is shown in Fig. 6. It is seen that the standard z transform has yielded an unusable result.

VI. SUMMARY

The need for sampled-data filters in wideband simulations of many processing systems has led to a synthesis method which overcomes the shortcomings of the standard z transform. The method presented consists of directly transforming a suitable continuous transfer function to a sampled-data filter by means of the bilinear z form. For wideband filters the method is particularly suited to those filters that exhibit relatively constant magnitude-frequency characteristics in successive pass and stop bands. Conventional design techniques of continuous filters are used

TABLE II—PARTIAL-FRACTION EXPANSION COEFFICIENTS FOR PARALLEL REALIZATION OF SAMPLED-DATA BANDSTOP FILTER

Term	Numerator Coefficients		Denominator Coefficients	
	A_1	A_0	B_2	B_1
1	0.0001628	0.0008827	0.9987854	0.1106416
2	−0.0009283	−0.0001764	0.9989898	0.4285348
3	−0.0024098	−0.0027894	0.9956089	0.1063723
4	0.0031774	0.0026966	0.9957459	0.4317548
5	0.0102446	0.0026026	0.9879911	0.0940731
6	−0.0037799	−0.0112135	0.9883051	0.4414974
7	−0.0277640	0.0127415	0.9651789	0.0616261
8	−0.0108027	0.0289421	0.9661438	0.4663508
9	0.0272223	−0.1163873	0.8694592	−0.0204564
10	0.1206914	−0.0054765	0.8742300	0.5186036
11	0.2973946	−0.2973227	0.5283651	0.2074591

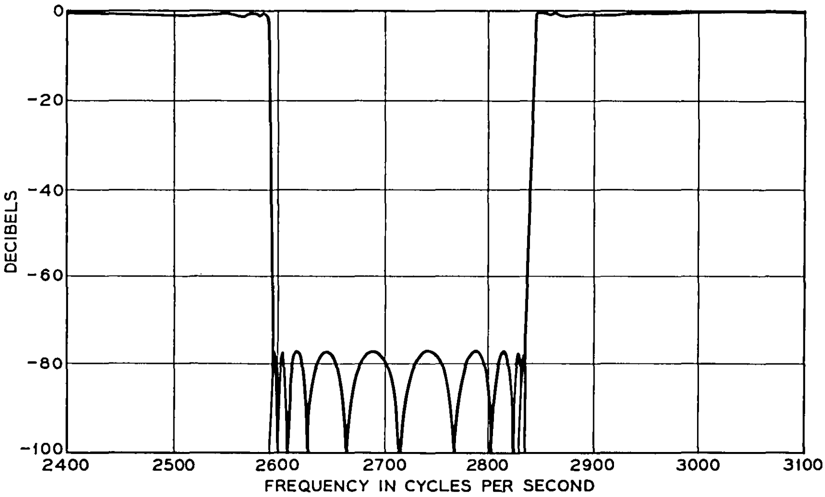


Fig. 5 — Frequency response characteristics of the sampled-data bandstop filter designed by the bilinear z transformation.

directly in the synthesis procedure of the sampled-data filters. (Thus the synthesized filters have frequency characteristics comparable to those of continuous filters.) An example has been presented of a filter function synthesized by this procedure and easily programmed for a simulation. Results obtained from this example demonstrate the usefulness and accuracy of the bilinear z-form method.

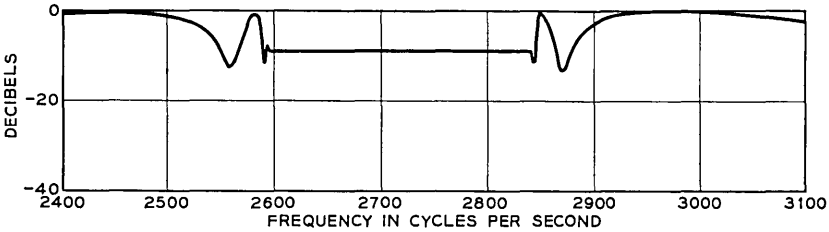


Fig. 6 — Frequency response characteristics of the sampled-data bandstop filter designed by the standard z transformation.

REFERENCES

1. Kaiser, J. F., Design Methods for Sampled-Data Filters, Proc. First Allerton Conference on Circuit and System Theory, Nov., 1963, Monticello, Illinois.

2. Golden, R. M., Digital Computer Simulation of a Sampled-Data Voice-Excited Vocoder, J. Acoust, Soc. Am., **35**, Sept., 1963, pp. 1358–1366.
3. Wilts, C. H., *Principles of Feedback Control,* Addison-Wesley, 1960, pp. 197–207.
4. Hamming, R. W., *Numerical Methods for Scientists and Engineers*, McGraw-Hill, New York, 1962, pp. 277–280.
5. Kelly, J. L., Jr., Lochbaum, C., and Vyssotsky, V. A., A Block Diagram Compiler, B. S. T. J., **40**, May, 1961, pp. 669–676.
6. Storer, J. E., *Passive Network Synthesis*, McGraw-Hill, New York, 1957, pp. 293–296.

The ALPAK System for Nonnumerical Algebra on a Digital Computer — III: Systems of Linear Equations and a Class of Side Relations

By J. P. HYDE

(Manuscript received March 6, 1964)

This is the third and last in a series of papers describing the ALPAK system for nonnumerical algebra on a digital computer. The first paper[1] *is concerned with polynomials in several variables and truncated power series with polynomial coefficients. The second paper*[2] *is concerned with rational functions in several variables and truncated power series with rational-function coefficients. The present paper discusses systems of linear equations with rational-function coefficients and a certain class of side relations.*

The ALPAK system has been programmed within the BE-SYS-4 monitor system on the IBM 7090 computer, but the language and concepts are machine independent. Several practical applications are described in Ref. 1.

This paper is divided into six sections. The first two assume that the reader has no knowledge of computers or computer programming and the last four assume that the reader is familiar with basic computer programming and Refs. 1 and 2. Section I is a general description of ALPAK and this paper; Section II discusses the different forms in which a linear system can occur, including canonical form; Section III describes the ALPAK linear system operations for converting these forms; Section IV discusses side relations; Section V describes list naming operations; and Section VI discusses possible future developments and improvements.

I. INTRODUCTION

This is the third and last in a series of papers describing the ALPAK system, a programming system for performing routine manipulations of algebraic expressions on a digital computer. The system can perform the operations of addition, subtraction, multiplication, division, substitution, and differentiation. The first paper[1] is concerned with polynomials in several variables and truncated power series with polynomial coefficients. The second paper[2] is concerned with rational functions in several variables and truncated power series with rational-function coefficients. The present paper describes the ALPAK facilities for manipulating and solving by Gaussian elimination systems of equations linear in certain variables with coefficients which are rational functions of other variables. The facilities for handling a certain class of side relations are also described.

The ALPAK system has been programmed within the BE-SYS-4 monitor system on the IBM 7090 computer, but the language and concepts are machine independent. Several practical applications are described in Ref. 1.

This paper is divided into six sections, of which the first two do not presuppose any knowledge of computers or computer programming and the last four assume that the reader is familiar with the basic concepts of computer programming and Refs. 1 and 2. Section I is a general description of ALPAK and deals with basic concepts. Section II describes the different forms in which a linear system can occur, including especially the canonical form of a linear system. Section III discusses the ALPAK linear system operations for converting these forms.

Section IV describes the way in which ALPAK has been programmed to simplify a rational function, using a certain class of side relations. The most important relations in the allowed class are of the form $X^2 = C$ (C a rational function independent of X). This includes in particular $i^2 = -1$ and $s^2 = 1 - c^2$ where s and c can stand for sin α and cos α, respectively. The simplification is done by a special rearrangement of the ALPAK format statement and has certain limitations.

Section V discusses list naming operations, a convenient set of auxiliary operations for handling arguments of ALPAK subroutines which are lists (one-dimensional arrays). Finally, Section VI discusses possible future developments and improvements.

1.1 *An Example of the ALPAK Language*

The simplicity of handling linear systems by ALPAK is illustrated in the process of solving the following system of two linear equations, *EQ1* and *EQ2*, in two unknowns, *X1* and *X2*, with polynomials in a as coefficients.

$$\mathrm{EQ1}: 3aX1 + 2aX2 - 1 = 0$$

$$\mathrm{EQ2}: 2aX1 + 5a^2X2 - 3 = 0$$

We first extract a coefficient matrix, SYS, for the equations with -1 and -3 moved to the right side.

$$\text{SYS:} \left\| \begin{matrix} 3a & 2a & 1 \\ 2a & 5a^2 & 3 \end{matrix} \right\| .$$

The matrix is then put into canonical form using Gaussian elimination.[3,4]

$$\text{SYS:} \left\| \begin{matrix} 1 & 0 & \dfrac{-6 + 5a}{-4a + 15a^2} \\ 0 & 1 & \dfrac{7}{-4a + 15a^2} \end{matrix} \right\| .$$

The fact that the original coefficient matrix and the canonical form matrix both have the name *SYS* does not imply that they are equal but rather that the latter replaces the former *physically* in the computer. The expressions for the unknowns are then extracted from the coefficient matrix.

$$\text{X1:} \ \frac{-6 + 5a}{-4a + 15a^2}$$

$$\text{X2:} \ \frac{7}{-4a + 15a^2} .$$

The following program illustrates how these operations are performed by ALPAK.

```
SYS    SYSRES    2,2
```

Reserve space in the computer for the physical representation of the set of system coefficients which will be obtained from two linear equations in two unknowns and name the set *SYS*.

```
       SYSFRM    SYS,(EQ1,EQ2),=2
```

Extract the 2 × 3 coefficient matrix from the equations *EQ1* and *EQ2* and place it in *SYS*. The "=2" says that there are two unknowns and the third column of the matrix is used for the terms of the equations which are independent of the unknowns.

```
       SYSPRT    SYS
```

Print the system coefficients.

```
        SYSCFM    SYS
                                   Put the system into canonical form
                                   (described in next section) using
                                   Gaussian elimination. If the system
                                   were triangular, the row selection
                                   strategy would cause this to be done
                                   in the obvious way. The canonical
                                   system retains the same name as the
                                   original system.
        SYSSLV    (X1,X2),SYS
                                   Fill X1 and X2 with the solutions
                                   for the unknowns in SYS. The op-
                                   eration SYSSLV assumes that SYS
                                   is in canonical form.
        RFNPRT    X1
        RFNPRT    X2
                                   Print X1 and X2.
X1
X2
                                   These are names of single cells in
                                   memory which will be filled in with
                                   "pointers"* to the physical repre-
                                   sentations of the solutions in the
                                   computer.
```

The usefulness of the linear system operations was demonstrated in a problem from queuing theory, proposed by L. Takacs,† in which a truncated power series of 813 terms was involved in forming a system of nine linear equations in nine unknowns. One of these unknowns was the third moment of a probability distribution. Its numerator had 200 terms in five variables with maximum degrees 1, 1, 3, 7, and 9 and its denominator had 39 terms in two variables with maximum degrees 7 and 10.

II. LINEAR SYSTEMS

In this section are discussed the different forms of linear systems as they are dealt with in the ALPAK context. It is important in writing ALPAK programs to remember what these forms are. The next section discusses the ALPAK subroutines for changing one form to another.

* See Ref. 2, p. 795.

† See Ref. 1, pp. 2090–2092.

2.1 *System of Equations*

A linear system of m equations in n unknowns, x_j, is a set of m rational functions, (1), of v variables ($v \geqq n$), each of which is linear in the x_j and is implicitly equal to zero.

$$\sum_{j=1}^{n} \lambda_{ij} x_j - c_i = 0 \qquad (1 \leqq i \leqq m). \tag{1}$$

Thus for each i in (1) the λ_{ij} are the coefficients of the x_j and together with c_i may be thought of as $n + 1$ rational functions with a common denominator.

2.2 *System Coefficients*

Consider (1) written in the form:

$$\sum_{j=1}^{n} \lambda_{ij} x_j = c_i \qquad (1 \leqq i \leqq m). \tag{2}$$

The λ_{ij} and the c_i of (2) shall be referred to as the *system coefficients* of the linear system (1). In ALPAK they form an array of $m(n + 1)$ rational functions stored row-wise and forwards.

2.3 *System Canonical Form*

Let $x_{a_1}, \cdots, x_{a_r}$ be a subset of the unknowns $x_1, \cdots, x_n$ which we shall call the *dependent set*, and let $x_{a_{r+1}}, \cdots, x_{a_n}$ be the remaining unknowns, which we shall call the *independent set*. The dependent set is said to be *valid* if r is the rank of the system and if the associated columns of system coefficients are linearly independent over the field to which they belong. The system

$$x_{a_i} + \sum_{j=r+1}^{n} \lambda_{ij} x_{a_j} = c_i \qquad (1 \leqq i \leqq r) \tag{3}$$

and its array of system coefficients are both said to be in *canonical form* with respect to such a dependent set.* It can be shown that for any linear system and a given valid dependent set, there exists a unique canonical form which is obtainable from the original system and which is satisfied by the same values of the x_{a_i}. One obtains this canonical form by Gaussian elimination; i.e., operating on the system by suitably chosen row operations and column interchanges.† When there is a choice of row interchange, the row with the most zero coefficients is selected to minimize the work involved. If, in the derived canonical form, $r < m$,

* The dependent set in (3) is clearly valid.

† See Refs. 3 and 4.

the last $m - r$ rows should be of the form $0 = 0$. If they are not, the system is said to be *inconsistent.* If $r < n$, the system is said to be *singular.*

2.4 *System Solution*

The solution of a linear system is a set of r rational functions, (4), of v variables ($v \geqq n - r$), each of which is linear in the x_{a_j} ($r + 1 \leqq j \leqq n$) and is implicitly equal to x_{a_i}.

$$x_{a_i} = c_i - \sum_{j=r+1}^{n} \lambda_{ij} x_{a_j} \qquad (1 \leqq i \leqq r). \tag{4}$$

The solution is easily produced once the system is in canonical form, and if the system is nonsingular the solution is of the form $x_{a_i} = c_i$ $(1 \leqq i \leqq n)$.

III. LINEAR SYSTEM OPERATIONS

3.1 *General Remarks*

In this section are discussed the ALPAK subroutines for converting the different linear system forms discussed in Section II. The name of a set of system coefficients must be defined by operations SYSNAM or SYSRES if it is to be used in any other operations. This name is the BSS address of a three-word system heading in which are stored the five system parameters. These parameters are the BSS address of the system coefficients, the number of equations, the number of unknowns, an ALPAK format address, and the number of leading variables in this format of which the equations are independent. They are set at assembly time by operations SYSNAM and SYSRES or at run time by operations SYSSET and SYSMPR.

The $m(n + 1)$ system coefficient pointers are stored row-wise and forwards, and a block of $n + 1$ cells must immediately follow to be used by ALPAK as work space. In the ALPAK format statement of the system equations, the n unknowns must have consecutive variable numbers $k + 1$ through $k + n$ ($k \geqq 0$). If $k > 0$, the system equations must be independent of the first k variables, and thus the system coefficients are independent of the first $k + n$ variables. The system parameter *fmt* is normally this ALPAK format statement and is referenced in any system operations involving the names of the unknowns. If it is not supplied by SYSNAM, SYSRES, SYSFRM, or SYSSET, all such operations must refer to variables by number (VARTYP NUM or VARTYP NUM*).

Those arguments of operations SYSFRM, SYSCFM, and SYSSLV which are lists are specified according to the conventions established in Section V. System parameters and system names are not indexable, but the addresses where they are stored or to be stored are. As in other ALPAK operations, index registers are preserved with the exception of index register four.

3.2 *Notational Conventions*

The following conventions of notation are used in descriptions of instructions. Upper-case letters are used for operation codes (including macro names) and for any parameters which must appear exactly as shown. Dummy parameters are indicated by lower-case letters. A dummy parameter usually stands for the symbolic address of a cell or block of cells in the program where the argument is stored. Those dummy parameters which are the arguments themselves are in boldface. Finally, optional parameters are enclosed in brackets, and parameters which usually have subarguments are enclosed in parentheses. All integer arguments are decimal. By this notation, then, the instructional description

sys SYSRES **m**,**n**,[fmt],[k]

specifies certain properties and restrictions about the arguments of the following call:

COEFF SYSRES 9,9,,INDEP

Thus, only SYSRES must appear exactly as shown and all other parameters are dummies with the third one omitted, as it is optional. The number of equations is nine, but the number of leading variables of which the equations are independent is in the cell whose symbolic address is INDEP.

3.3 *Linear System Operations*

sys	SYSNAM	bss,**m**,**n**,[fmt],[k]	name	(a)
sys	SYSRES	**m**,**n**,[fmt],[k]	reserve	(b)
	SYSPRT	sys	print	(c)
	SYSFRM	sys,(listr),n,[k]	form	(d)
	SYSCFM	sys,[(listv)],[**inc**],[**ids**]	canonical form	(e)
	SYSSLV	(listr),sys,[(listv)]	solve	(f)
	SYSOBT	[(abss)],[(m)],[(n)],[(afmt)], [(k)],sys	obtain parameters	(g)
	SYSSET	sys,[abss],[m],[n],[afmt],[k]	set parameters	(h)
	SYSMPR	sys,[(**op** oper)], [(**op** oper)] [(**op** oper)], [(**op** oper)], [(**op** oper)]	modify parameters	(i)

sys = name of system (symbolic address of heading)
bss = BSS address of the array of system coefficients
abss = address where *bss* is or is to be stored
m = the number of equations in the system
n = the number of unknowns in the system
fmt = the address of the system's ALPAK format statement
afmt = address where *fmt* is or is to be stored
k = the number of leading variables in this format statement of which the system equations are independent
listr = list of rational functions (see Section V)
listv = list of variables (specified in the manner indicated by the last previous VARTYP declaration — see Section V)
(op oper) = a 7090-94 machine operation and an operand separated by a blank
inc = inconsistency return
ids = invalid dependent set return.

3.4 *Descriptions*

(a) sys SYSNAM bss,**m,n**,[fmt],[k]

Declare a block of length $(m + 1)$ $(n + 1)$ starting at *bss* to be a set of linear system coefficients and work space, and name it *sys* by reserving remotely a three-word system heading. This heading is filled in with *bss*, *m*, *n*, *fmt*, and *k*. If *fmt* and/or *k* is omitted, the corresponding fields in the system heading are filled in with zeros.

(b) sys SYSRES **m,n**,[fmt],[k]

Reserve remotely a block of length $(m + 1)$ $(n + 1)$ for a set of system coefficients and work space, and name the set *sys* by reserving remotely a three-word system heading as in SYSNAM. *sys* is to be filled in at run time (e.g., by SYSFRM).

(c) SYSPRT sys

Print the set *sys* of system coefficients.

(d) SYSFRM sys,(listr),n,[k]

Replace *sys* by the set of system coefficients formed from the set *listr* of system equations and remove the common factors between the coefficients of any given equation (*listr* is destroyed). The contents of *n* and *k* and the number of rational functions in *listr* together with their format are copied into the heading of *sys*. If *k* is not supplied, it is assumed to be zero. If SYSFRM is not used to fill in *sys*, the system parameters must be filled in with operations (a), (b), or (h).

(e) SYSCFM sys,[(listv)],[inc],[ids]

Replace the set *sys* of system coefficients by its associated canonical set, using Gaussian elimination. *listv* is a list of unknowns (specified in the manner indicated by the last previous VARTYP declaration) to be included in a valid dependent set. If *listv* is not supplied, the list is assumed to be empty. If *sys* is found to be inconsistent, control will be transferred to *inc* (or to the REMARK subroutine if *inc* is not supplied) and *sys* will have a canonical form with an inconsistency. If the set of unknowns in *listv* cannot be included in a valid dependent set, control will be transferred to *ids* (or to the REMARK subroutine if *ids* is not supplied) and *sys* will have a canonical form with some subset of *listv* in the dependent set. At *inc* or *ids* it is possible to call SYSSLV, SYSPRT, or to go to some other part of the program.

(f) SYSSLV (listr),sys,[(listv)]

Replace *listr* (whose length must not be less than that of *listv*) by the solutions for the list of unknowns *listv* (specified in the manner indicated by the last previous VARTYP declaration). *sys* is assumed to be in canonical form. If *listv* is not supplied, all the unknowns in the dependent set are solved for in the order in which they were at the start of SYSCFM.

(g) SYSOBT [(abss)],[(m)],[(n)],[(afmt)],[(k)],sys

Obtain the system coefficient parameters of the system whose name is *sys*. Each optional argument is a memory location in whose address field the parameter is to be stored. Thus the parameter *bss* is stored in the location *abss* specified by SYSOBT, etc. Each optional argument may actually be several arguments, and if an argument is an integer equal to seven or less, it refers to an index register.

(h) SYSSET sys,[abss],[m],[n],[afmt],[k]

Set the system coefficient parameters of the system whose name is *sys* from the locations specified by the bracketed arguments. Thus the parameter *bss* is set to the contents of the location *abss* specified by SYSSET, etc.

(i) SYSMPR sys,[(**op** oper)],[(**op** oper)]

[(**op** oper)], [(**op** oper)], [(**op** oper)]

Modify the system coefficient parameters of the system whose name is *sys* using the 7090-94 machine operations *op* with operands *oper*. Thus, the parameter *bss* is modified by the first operation and operand, m is modified by the second, n by the third, *fmt* by the fourth, and k by the fifth. Each operation and operand may be different. Typically, the operation is ADD or SUB and the operand is the address of some increment or decrement.

IV. SIDE RELATIONS

4.1 *General Remarks*

The ALPAK programmer may find that expressions involving radicals occur in his problem. A radical can be handled by assigning it a variable name and writing the rational functions using this name. Thus in the polynomial $a + 2a\sqrt{3}$, we let $X = \sqrt{3}$ and the expression becomes $a + 2aX$. The problem is that in the outputs of arithmetic operations involving such rational functions, X can have an exponent greater than one and the fact that $X^2 = 3$, $X^3 = 3X$, $X^4 = 9$, $\cdots$ will be ignored. The implicit equation $X^2 = 3$ is called a *side relation* of degree two on X. A subroutine is provided for simplifying rational functions using side relations of the general form

$$X^{2^j} = C \qquad (j \text{ an integer} \geqq 1) \qquad (C \text{ a rational function independent of } X). \tag{5}$$

This category includes especially $i^2 = -1$ and $s^2 = 1 - c^2$ where s and c can stand for sin α and cos α, respectively.

4.2 *Limitations*

Many limitations exist in the present handling of side relations. In the relation $X^n = C$, n must be a power of two and X a single variable. Dependencies between relations are not observed; i.e., $R^2 = 2$ and $S^2 = 3$ and $T^2 = 6$ will not result in the implicit relation of $T = \pm RS$. Moreover, relations are not handled automatically by the lowest-level subroutines, thus causing exponents to grow unnecessarily until simplification is done at main program level. A more sophisticated version of ALPAK would prevent this by including the relations as part of the format statement. To repair these limitations would require a great deal of extra programming, and it turns out in practice that these limitations do not usually matter. The general problem of dependencies is especially difficult, as it involves algebraic extensions of the field of rational functions of several variables.

4.3 *Implementation*

The simplification of a rational function, *RF*, by a side relation $X^{2^j} = C$ is accomplished with the aid of a specially constructed temporary ALPAK format statement. The temporary format is the same as the original one except that the exponent field of X is split into two parts. The right j bits are assigned the name X and the remaining left-hand bits form a temporary exponent field which is assigned any name and stands effectively for X^{2^j}. The rational function C is then substituted for the temporary variable by the call SIDREL. If several side relations are defined on several variables, then a single temporary format statement can be used to split up these variables. There will then be a list of rational functions to be substituted for the temporary variables by a single call to SIDREL (see Section 3.2).

```
SIDREL      rf,(listv),(listr),tfmt
```

rf = rational function to be simplified
listv = list of temporary variables (specified in the manner indicated by the last previous VARTYP declaration — see Section V)
listr = list of rational functions to be substituted for these variables and specified in the same order (see Section V)
tfmt = address of temporary format.

The format of *rf* after simplification is the format of the items in *listr*, or if none of these items has a format, the format of *rf* is unaltered.

4.4 *Example*

Suppose it is desired to simplify the function *RF* using the side relation $I^2 = -1$. This is done by the following program.

```
FMT     POLCVF    (X,5,Y,5,I,5,Z,21)
                            Permanent format.
TFMT    POLCVF    (X,5,Y,5,ISQ,4,I,1,Z,21)
                            Temporary format with I split up
                            into ISQ with four bits and I with
                            one bit.
        VARTYP    NAM
        POLSTC    MON,=-1
        SIDREL    RF,ISQ,MON,TFMT
            ⋮
RF
MON
```

Testing equality of rational functions R and S which have been simplified by a side relation should always be done by subtracting and testing for zero as follows:

```
RFNSUB    TEMP,R,S
SIDREL    TEMP,ISQ,MON,TFMT
   ⋮                 The side relation applied to R and
                     S is applied to TEMP.
RFNZET    TEMP
                     Test if TEMP is zero.
```

This procedure will recognize that the expressions $(1 + i)/(1 - i)$ and i are equal.

V. LIST NAMING OPERATIONS

5.1 *General Remarks*

Whenever an ALPAK subroutine argument is a list, the list may be specified either by actually listing the contents; e.g.,

```
SYSSLV    (P,Q,R),SYS,(X,Y,Z)
```

or by name; e.g.,

```
SYSSLV    (LISTP,*),SYS,(LISTV,*)
```

Here the asterisk indicates that the list has been specified by name. Both methods may be used within the same command; e.g.,

```
SYSSLV    (P,Q,R)SYS(LISTV,*)
```

This section describes a set of operations LSTNAM, LSTMAK, and LSTRES for assigning names to lists and blocks of storage, thus enabling one subsequently to call them by these names in the appropriate subroutines. The operations LSTOBT, LSTSET, and LSTMPR serve as auxiliary operations. The facilities are especially useful whenever the items of the list are to be filled in at run time or whenever the items do not form a contiguous block in core. A list has two parameters, which are its BES address and its length. These can be set at assembly time by LSTNAM, LSTMAK, and LSTRES or changed at run time by LSTSET and LSTMPR (see Section 3.2). List parameters and list names are not indexable, but the addresses where they are stored or to be stored are. Each item in the specified contents of a list may be tagged. Index registers are preserved with the exception of index register four.

5.2 *List Naming Operations*

ttl	LSTNAM	bes,**lng**,[VAR]	name	(a)
ttl	LSTMAK	(items),[VAR]	make	(b)
ttl	LSTRES	**lng**,[VAR]	reserve	(c)
	LSTOBT	[(abes)],[(lng)],ttl	obtain	(d)
	LSTSET	ttl,[abes],[lng]	set	(e)
	LSTMPR	ttl,[(**op** oper)],[(**op** oper)]	modify	(f)

ttl = name for list
bes = BES address of list
abes = address where "bes" is or is to be stored
lng = length of list
items = contents of list
(**op** oper) = 7090–94 machine operation and an operand separated by a blank.

5.3 *Descriptions*

(a) ttl LSTNAM bes,**lng**,[VAR]

Declare a set of items in a contiguous block of length *lng* to be a list whose BES address is *bes*, name it *ttl*, and set the list parameters to *bes* and *lng*. If *VAR* is present, the list is assumed to consist of variables (specified in the manner indicated by the last previous VARTYP declaration). If *VAR* is not supplied, the list is assumed to consist of rational functions, polynomials, etc. (i.e., of symbolic addresses of pointers.)

(b) ttl LSTMAK (items),[VAR]

Declare the set whose elements are the subarguments in *items* to be a list, name it *ttl*, and set the list parameters accordingly. *VAR* is as described in LSTNAM. The items need not be in a contiguous block as in LSTNAM.

(c) ttl LSTRES **lng**,[VAR]

Reserve remotely a block of length *lng* for a list, name it *ttl*, and set the list parameters to the BES address of the block and *lng*. *VAR* is as described in LSTNAM. The list is to be filled in at run time (e.g., by SYSSLV).

(d) LSTOBT [(abes)],[(lng)],ttl

Store the BES address of the list whose name is *ttl* in location *abes* and store its length in *lng*. The bracketed arguments may actually consist of several subarguments, and if an argument is an integer equal to seven or less, it refers to an index register.

```
(e)            LSTSET     ttl,[abes],[lng]
```

Set the BES address of the list whose name is *ttl* to the contents of *abes* and set its length to the contents of *lng.*

```
(f)            LSTMPR     ttl,[(op oper)],[(op oper)]
```

Modify the BES address of the list whose name is *ttl* using the 7090-94 machine operation *op* with operand *oper* specified by the first bracketed argument. Modify the length of the list in a similar manner as indicated by the second bracketed argument. Typically, the operation is ADD or SUB and the operand is the address of some increment or decrement.

5.4 *Example*

The following example shows how list naming can be used to good advantage. We are given polynomials ($A_1, \cdots, A_y$; $y \leqq 15$), a set of variable names ($m_1, \cdots, m_y$; $y \leqq 15$), and polynomials ($F_1, \cdots, F_y$; $y \leqq 15$). It is desired to form a set of polynomials ($G_1, \cdots, G_y$; $y \leqq 15$) in the following way, where m_i: A_i means A_i is substituted for m_i

$$\begin{aligned} G_1 &= F_1(m_1{:}A_1) \\ G_2 &= F_2(m_1{:}A_1, m_2{:}A_2) \\ \vdots\ & \quad \vdots \\ G_y &= F_y(m_1{:}A_1, m_2{:}A_2, \cdots, m_y{:}A_y). \end{aligned}$$

Assume that the F_i's, A_i's, M_i's, and G_i's are stored forwards in blocks whose BES addresses are F, A, M and G respectively and that the parameter y is in location Y. The following program will perform the substitution.

```
POLS   LSTNAM    A,15
VARS   LSTNAM    M,15
                           Define the lists thus setting the list
                           parameters to (A,15) and (M,15)
       LSTMPR    POLS(SUB  Y) (SUB  =15)
       LSTMPR    VARS(SUB  Y) (SUB  =15)
                           Initialize the list parameters to (A −
                           y,0) and (M − y,0).
       LXA       Y,1
```

```
LOOP    LSTMPR    POLS(ADD   =1) (ADD   =1)
        LSTMPR    VARS(ADD   =1) (ADD   =1)
        POLSST    (G,1) (F,1) (POLS,*) (VARS,*)
        TIX       LOOP,1,1
                             Increment the list parameters by
                             one at each repetition of the above
                             loop.
               .
               .
               .
F       BES       15
A       BES       15
M       BES       15
G       BES       15
Y
```

VI. OUTLOOK

Our experience has shown us that the present handling of linear systems has its limitations. Large linear systems are always difficult to put into canonical form, and even a relatively simple set of system coefficients can grow quite rapidly throughout the course of SYSCFM and cause some form of overflow. This growth becomes coupled with the growth produced by the greatest common divisor algorithm,* thus making the inadequacies of the latter most apparent. The success or failure of SYSCFM depends less on the dimensions of the system and more on the internal structure and size of the individual coefficients. Moreover, it is very difficult to tell by looking at the input array whether the structure and size at a later stage of the reduction will cause trouble. This difficulty is illustrated in that SYSCFM succeeded in reducing a 9 × 9 array with large, apparently complex, entries,† but failed in a related queuing theory problem to reduce a 10 × 10 array whose entries averaged only two or three terms. It can at least be said that there will be no GCD problems if the original array consists of all rational numbers.

The subroutine SYSCFM is perhaps too comprehensive. A series of orders which would enable one to perform the Gaussian elimination method a step at a time, leaving the choice of row and column permutations completely up to the user, might be useful. SYSPRT could then be called at any time during the reduction. A routine for evaluating determinants, if available, would enable the solution of nonsingular systems by Cramer's method as an alternative.

* See Ref. 2, pp. 791–794.

† See Ref. 1, pp. 2090–2092.

The growth problem could be reduced by allowing multiple precision polynomial coefficients and by allowing a polynomial to be represented as a product of polynomials (not necessarily irreducible). Thus one could compute the GCD as a product of simpler GCD's. To do this would require the ability to have a data structure hierarchy in the data buffer more complicated than that of a rational function.* This ability would also enable a linear system itself to be such a data structure rather than an array in the main program.

A new version of ALPAK (to be called ALPAKB) is now being developed. Its foundation is a programming system[5] called OEDIPUS (Operating Environment with Dynamic storage allocation, Input-output, Public push down list, Unhurried diagnostics, and Symbolic snaps) which provides for the dynamic storage allocation of such data structures, among other things. ALPAKB will also include multiple precision integer arithmetic which will handle polynomial coefficient overflow.

VII. ACKNOWLEDGMENT

I would like to thank W. S. Brown for many valuable suggestions and discussions concerning every aspect of this paper.

REFERENCES

1. Brown, W. S., The ALPAK System for Nonnumerical Algebra on a Digital Computer — I: Polynomials in Several Variables and Truncated Power Series with Polynomial Coefficients, B.S.T.J., **42,** Sept., 1963, p. 2081.
2. Brown, W. S., Hyde, J. P., and Tague, B. A., The ALPAK System for Nonnumerical Algebra on a Digital Computer — II: Rational Functions of Several Variables and Truncated Power Series with Rational-Function Coefficients, B.S.T.J., **43,** March, 1963, p. 785.
3. Hyde, J. P., unpublished work.
4. Stoll, Robert R., *Linear Algebra and Matrix Theory,* McGraw-Hill, New York, 1952. (See especially Chap. 1.)
5. Brown, W. S., and Leagus, D. C., OEDIPUS: Operating Environment with Dynamic storage allocation, Input-output, Public push down list, Unhurried diagnostics, and Symbolic snaps, to be published.

* See Ref. 2, p. 794.

A Technique for Measuring Small Optical Loss Using an Oscillating Spherical Mirror Interferometer

By A. J. RACK and M. R. BIAZZO

(Manuscript received March 19, 1964)

The measurement of very small optical losses (the order of a few per cent) by conventional methods becomes very difficult because of the extreme accuracy required. This article shows that both high mirror reflectances and low transmission losses can be readily measured using an oscillating mirror interferometer as a frequency spectrum analyzer. The theory developed shows that when this type of interferometer is excited by a continuous gaseous laser, the total optical loss is proportional to the frequency resolution or the finesse. The theory also shows that the first-order velocity effect produced by having the mirror move at a velocity of one foot per hour can be large if the total optical loss is about 0.25 per cent. For the velocities and optical losses we have encountered so far in our measurement system, the first-order mirror velocity effect can be neglected. The range of reflectance of mirrors we have measured is from 94 to 99.5 per cent, and the measurements for the optical transmission loss range from 0.2 to 3 per cent. The accuracy to which a 1 per cent loss can be repeated is 1.0 ± 0.1 per cent. It was found that the transmission loss through an optical grade of fused quartz (Homosil) at 6328 Å is about 1 db per meter, and that for Plexiglas II is about 2 db per meter.

I. INTRODUCTION

In developing a long-distance optical communication system employing a large number of components, it is essential to be able to measure accurately the optical transmission loss of each component. These components may include mirrors with reflectances in the order of 99 per cent, Brewster angle output windows, various lenses, and other passive elements with optical transmission losses of 1 per cent or less. For such small losses, the conventional measuring techniques become increasingly difficult because of the extreme accuracy required.

In conventional measuring systems, the loss (or reflectance) is calculated by comparing the electrical output of a photodetector for two different optical conditions, first with the unknown in the system, and then with the unknown removed. As the magnitude of the optical loss decreases, this comparison becomes more and more inaccurate, as it requires the measurement of a small difference between two relatively large photodetector outputs. In these measuring systems, a number of methods have been developed to minimize these errors, which are principally optical.[1] However, it is believed that measuring techniques to be described below will more readily measure very small optical losses.

The proposed method uses a frequency spectrum analyzer at optical frequencies. Such an analyzer can be obtained by using a Fabry-Perot type of interferometer as a transmission element between an optical source and a photodetector.[2,3] When the mirror separation of the interferometer is varied periodically by moving one of the mirrors linearly, a large photocell output will be obtained whenever the optical cavity is in resonance at any frequency that may be present in the optical source. If the photodetector output is observed on an oscilloscope whose sweep is synchronized with the mirror drive, the scope will display the energy distribution of an optical source as a function of frequency. If this type of Fabry-Perot interferometer is illuminated by a continuous laser with its extremely narrow line output (one or two cycles wide), the linewidth of the pattern displayed on the scope is determined by the optical losses in the cavity itself, and by the velocity of the moving mirror. If the excursion of the moving mirror is several optical wavelengths, the scope pattern will repeat several times during a single sweep trace. That is, as the mirror separation increases, the mth harmonic of the cavity becomes resonant with the source; a short time later, the $(m + 1)$th harmonic is resonant with the same optical frequency, then the $(m + 2)$th, and so on. At each resonant point, there will be several output scope pulses, since the laser usually has an output at more than one frequency. The ability of any interferometer to separate or resolve two adjacent optical frequencies is determined by the finesse of the system. In the moving-mirror interferometer, the finesse is equal to the ratio of the fundamental pulse group spacing to the half-power-height width of any pulse. It will be shown below that the total power loss of the Fabry-Perot cavity, expressed as a ratio, is equal to π divided by the finesse.

This interferometer method of measuring small optical losses has been suggested on several occasions,[4] but it is believed that this is the first time such a system has been so fully developed. Since this system of measurements requires an optical source with an output linewidth very narrow compared to that of the optical cavity, loss measurements cannot be made over a continuous range of wavelengths, but only at the discrete wavelengths at which cw lasers have been developed. Since some light must be transmitted through the cavity, the reflectance can be measured only for those mirrors not coated with an opaque reflecting surface, such as the multiple-layer dielectric coated mirror.

The interferometer loss measuring technique is very sensitive to building and floor vibrations, and to air currents. The reason for this is as follows: if the separation between the two mirrors in the cavity is changed by one half wavelength, the resonant frequency of the interferometer is changed by the natural frequency of the cavity, which in our system is about 1 kmc, whereas the bandwidth of the cavity for 1 per cent optical loss is only about 3.0 mc. Thus, even very small random variations in the mirror spacing in both the laser and in the cavity, produced by either vibrations or air currents, will cause the output pulse pattern displayed on the scope to have large random time position variation.

II. THEORY

The relation for which the mirror reflectance and other optical losses can be calculated from measurable quantities is derived in the Appendix, and will be given briefly here. The effects of the velocity of the moving mirror were included in the Appendix.

Let

R_1, R_2 = power reflectance of the two mirrors expressed as a ratio
g = power loss per single pass through any material within the cavity, also expressed as a ratio
T_c = fundamental pulse group spacing
T_p = half-power output pulse width
v = velocity of moving mirror
d_0 = mirror spacing
t = time
λ = optical wavelength in free space
c = velocity of light in free space

$$\beta_0 = \frac{4\pi d_0}{\lambda} \frac{v}{c}$$

$$x = g(R_1R_2)^{\frac{1}{2}}$$

$$\alpha \equiv -\frac{4\pi vt}{\lambda}$$

$$g_0 \equiv [g(1 - R_1)(1 - R_2)]^{\frac{1}{2}}.$$

An exceedingly good approximation for the time response of the interferometer appropriate for computer calculation is given by (5) of the Appendix:

$$I_T \approx \left[\sum_{n=0}^{\infty} g_0x^n \cos\,(\alpha n + \beta_0 n^2)\right]^2 + \left[\sum_{n=0}^{\infty} g_0x^n \sin\,(\alpha n + \beta_0 n^2)\right]^2. \quad (5)$$

Since x is very nearly unity, a large number of terms must be taken in the series. For a maximum error of ϵ in stopping the series after the first $N - 1$ terms, we have

$$N \approx \frac{\log\,(2/\epsilon)}{1 - x}. \quad (6)$$

For example

$$\epsilon = 1/1000 \quad \text{and} \quad x = 0.9975, \qquad N = 3040.$$

The values of I_T shown in Fig. 1 were calculated for the following three

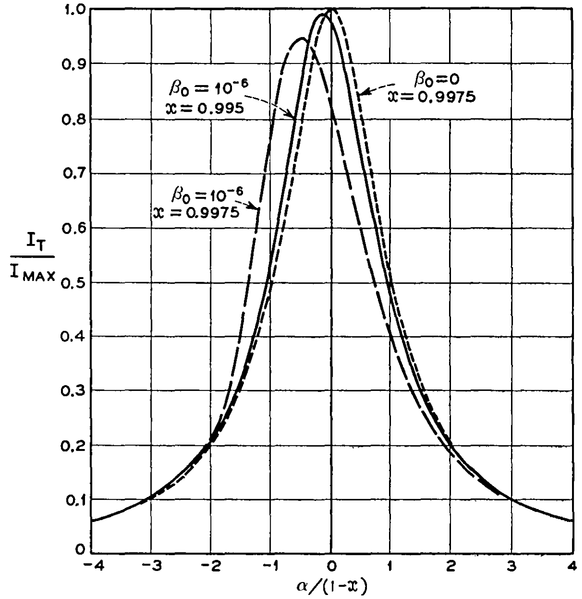


FIG. 1 — Interferometer time response as affected by mirror velocity.

conditions: (a) $\beta_0 = 10^{-6}$, and $x = 0.9975$, (b) $\beta_0 = 0$, $x = 0.9975$, and (c) $\beta_0 = 10^{-6}$ and $x = 0.995$. The curves given in Fig. 1 show that the first-order mirror velocity effect is to decrease the maximum response, increase the half-power-height pulse width, and make the response unsymmetrical about the maximum response.

From a number of computed values for (5), of which only a few are given in Fig. 1, it can be shown that the first-order effect of the moving mirror's velocity is to increase the half-power pulse width according to the relation (see Appendix)

$$\frac{Tm}{Tp} \approx \left[1 + 13.8\,\frac{\beta_0^2}{(1-x)^4}\right]^{\frac{1}{2}}. \tag{7}$$

This relation holds only when T_m is within a few per cent of T_p. For larger values, the computed results are less than those given in (7). Thus, in order to have the width increase by less than 1 per cent,

$$\frac{\beta_0}{(1-x)^2} \leqq \frac{1}{25}.$$

This relation can readily be satisfied unless the total loss becomes extremely small. In our measurements, the mirror spacing is varied about one micron at a 20-cycle frequency. The mirror spacing is about 15 cm and the wavelength is 6328 Å. Then, if $x = 0.995$ (the largest we have measured),

$$v = \tfrac{1}{250}\ \text{cm/sec} \approx 0.4\ \text{ft/hour}$$

and

$$\frac{\beta_0}{(1-x)^2} = \frac{0.4 \times 10^{-6}}{25 \times 10^{-6}} = \frac{1}{62.5}.$$

Hence, for optical losses of 0.5 per cent or greater, we can neglect the first-order mirror velocity effects.

If mirror velocity effects can be neglected, then the system response given by (10) and (11) of the Appendix can be used. The relation between the optical loss and the finesse of the system is

$$g(R_1R_2)^{\frac{1}{2}} = 1 - (\pi T_p/T_c) + \tfrac{1}{2}\,(\pi T_p/T_c)^2 + \cdots \tag{11}$$

where T_c/T_p is defined as the finesse. This equation points out the potential accuracy with which small losses can be measured. The interferometer method actually measures how much the combined loss, $g(R_1R_2)^{\frac{1}{2}}$, differs from unity. Obviously, the smaller this difference, the greater can be the experimental errors to obtain a fixed accuracy in

$g(R_1R_2)^{\frac{1}{2}}$. For a value of $g(R_1R_2)^{\frac{1}{2}}$ near 0.99, a 10 per cent error made in the measurement of T_p/T_c would give an error of only one part in a thousand in the value of $g(R_1R_2)^{\frac{1}{2}}$.

To obtain a good signal-to-noise ratio in the electrical output of the photodetector, it is important to obtain a maximum amount of light transmission through the interferometer. From (10), the fraction of the incident light transmitted through the cavity is given by (see Appendix)

$$I_m = \frac{g(1 - R_1)(1 - R_2)}{[1 - g(R_1R_2)^{\frac{1}{2}}]^2}. \tag{10}$$

Equation (10) shows that for $g = 1$ and $R_1 = R_2$, I_m is unity for any value of reflectance, but if $R_1 \neq R_2$, I_m will be less than unity. For example, if $R_1 = 0.995$, $R_2 = 0.97$ and $g = 1$, then $I_m = 0.49$. The remaining 51 per cent of the light is reflected back towards the source. Therefore the two mirror reflectances should be identical for maximum transmission. For measurements of transmission losses in the cavity, there is some advantage to be obtained by not having the mirror reflectances too great. As an example, for $g = 0.97$ and $R_1 = R_2 = 0.99$, $I_m = 0.062$. For the same loss and $R_1 = R_2 = 0.97$, $I_m = 0.25$. In the second case, however, there is a greater chance of making an error in measuring g since it is a smaller fraction of the total loss.

The shape of the output pulse expressed as a function of time is given by (12) of the Appendix

$$I(t) \approx \frac{I_m}{1 + 4t^2/T_p^2}. \tag{12}$$

The frequency spectrum of this time pulse is

$$F(\omega) = \frac{\pi}{2} T_p I_m \exp\left(-\frac{T_p}{2}\,|\omega|\right).$$

Now, the value of T_p is a function of the velocity of the moving mirror and hence, subject to the limitations on velocity discussed above, may be made as large or small as is desired. In our laboratory, the motion of the moving mirror was so selected that for a 1 per cent total cavity loss, the pulse width is about 30 μsec, and the frequency spectrum is down to 1 per cent of its low-frequency value at about 50 kc. The required bandwidth of the photodetector and its associated electrical circuits, including the viewing oscilloscope, is increased by a decrease in the total optical loss, as this decreases the pulse width, T_p. In order that the system be capable of measuring optical losses as small as 0.25 per cent, the over-all bandwidth of the electrical components should be at least 200 kc.

The above theory was developed for the assumptions that the incident beam of monochromatic light was collimated and that the plane surface mirrors were infinite in size. For finite-size mirrors with plane or spherical surfaces, and for a finite-size input light beam diameter, the electromagnetic energy inside the interferometer can be described by the familiar TEM modes.[5,6,7] For a given input light beam condition and for a given set of mirrors, the energy within the cavity can be characterized by selecting the appropriate amplitudes of the TEM modes. If the input beam spot size is too large or too small, a large portion of the input energy will be found in higher transverse modes of quite large order. If the laser is adjusted to operate in only the TEM_{00q} mode, then it has been shown[8] that the light energy will be principally in the fundamental TEM_{00q} mode in the cavity, if the spot size and the surfaces of constant phases of the input beam are both equal to those for the TEM_{00q} resonant cavity mode. As is usual in matching problems, these conditions are not too critical.

When the measuring interferometer has spherical mirrors at nonconfocal spacing, an incident light ray at a small angle off the system axis will produce repeated reflections, which in general will trace an ellipse on the mirrors.[9] Under special conditions, the points of reflection lie on a circle and are displaced by some angle after every round trip. When this angle is a multiple of 2π, the rays will exactly retrace their paths, and the trace of reflections on a mirror will break up in a number of separate and equally spaced dots. Under these reentrant conditions, the cavity will become resonant not only at a multiple of the fundamental cavity frequency, but also at multiples of a much lower frequency which is not quite a subharmonic of the fundamental cavity frequency. For these conditions, the oscilloscope pattern of the photocell output will show, in addition to the main response, a number of smaller equally spaced pulses. Since the response of one of the "off-axis" modes can coincide (or nearly coincide) with the main response, the measurement of the loss under these conditions can be considerably in error unless these off-axis mode responses are made very small.

In general, the system should be designed to be nondegenerate. That is, the length of the interferometer must be so selected as to avoid any possible overlapping of the cavity resonant response of the different orders of the TEM modes to any of the several optical frequencies present in the laser source. Usually, this is not difficult to accomplish.

III. MEASUREMENT SYSTEM

A block diagram of the components in the interferometer measuring system is given in Fig. 2. The He-Ne laser has external spherical mirrors

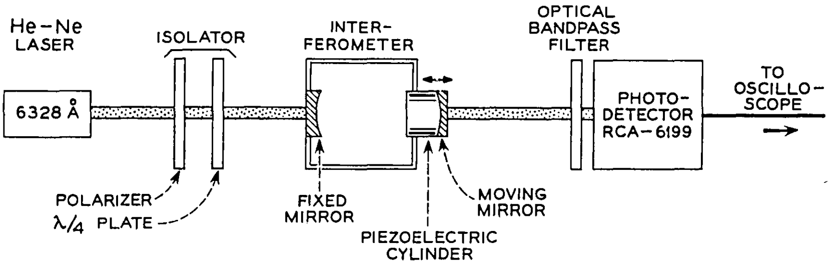


FIG. 2 — Schematic diagram of the interferometer measuring system.

and Brewster angle output windows, and operates at 6328 Å. To reduce the effects of vibrations, the laser and the cavity structures were made very rigid by using a construction similar to that used by Bennett in his magnetostrictively tuned laser.[10] The interferometer mirrors are mounted with suitable tilt controls in 6-inch square steel end blocks, 1 inch thick. The two blocks are tied rigidly together at each of the four corners by 1-inch diameter Invar rods 16 cm long. The interferometer cavity, shown in Fig. 3, uses a piezoelectric transducer to vary the mirror spacing in the cavity. It is a ceramic cylinder $1\frac{1}{4}$ inches ID, $1\frac{1}{2}$ inches OD, and $1\frac{1}{2}$ inches long.[11] The mirror holder is epoxied to one end of the cylinder. The other end is epoxied to a mounting plate which is fastened through suitable tilt controls to the steel end block. The laser, interferometer, and other optical components were fastened rigidly to a heavy steel optical table. This 4 by 8-foot table is supported on six small airplane inner tubes encased in heavy canvas covers. For this type of support, the natural frequency of the table is about four cycles per

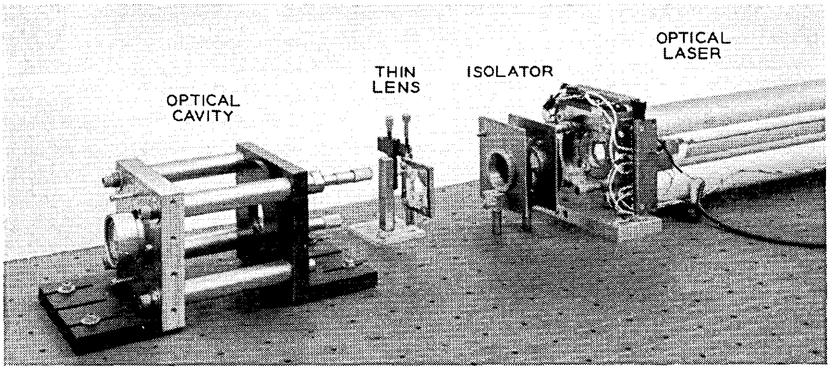


FIG. 3 — Photograph of the interferometer apparatus

second. To avoid air currents, both the laser and the optical cavity are enclosed separately in Plexiglas boxes.

The isolator between the laser and the cavity was found to be essential to prevent interaction between the two optical cavities. It is a circular polarizer consisting of a polaroid analyzer and a quarter-wavelength plate. This circular polarizer will absorb any light reflected back from the interferometer, as the sense of rotation of the circularly polarized light is inverted upon reflection and hence will not be transmitted back through the polaroid analyzer. For this type of isolator, all loss measurement must be made with circularly polarized light.

The photodetector is a standard electron multiplier phototube with an S-20 cathode. A bandpass optical filter, 200 Å wide, centered at 6300 Å, is placed between the source and the detector to eliminate most of the background light.

A 20-cycle triangular wave shape generator delivers 500 volts pp to the piezoelectric mirror drive. For this voltage, the motion of the mirror is about one micron, which is about three half-wavelengths of the 6328 Å laser source. The motion of the mirror is parallel and was checked by using an alignment telescope with a flat mirror placed in the movable mirror holder. The fringe pattern of the alignment telescope reflected back from the moving mirror showed no discernible change when 1000 volts dc or ac was applied to the driver. Therefore any mirror tilt variation must be less than 5 seconds of arc.

IV. EXPERIMENTAL PROCEDURE

When spherical mirrors are used in the interferometer, the ray of incident light must be on a line passing through the centers of curvature of both mirrors. With the aid of a thin optical lens to vary the incident angle, the cavity can readily be aligned to minimize the "off-axis" modes.[2,9]

The visible red gaseous laser, which was one meter long, was so adjusted that it oscillated only in the fundamental transverse mode and at several longitudinal modes.

In spite of all the precautions taken to eliminate building vibrations and air currents, the output pulse pattern on the output scope shows a considerable amount of time position jitter for any one pulse whenever the scope sweep speed is increased to be able to measure the half-power pulse widths. When the sweep speeds are made 10 μsec per cm to view a 30-μsec output pulse, the excursion of the time jitter is about ±50 μsec, and the jitter frequency is about two cycles or less.

To overcome the effects of the time jitter, a photographic method was developed to determine the pulse width. For a single period of the 20-cycle drive on the piezoelectric driver, the time jitter is small. The half height of the output pulse was determined by using the following electronic circuits. In the amplifier following the photocell, a fast transistor switch, operating at a one-megacycle rate, reduces the voltage gain of the network periodically by a factor of two (6 db). With this arrangement, a single photograph shows simultaneously both the full-height and the half-height pulses, as shown in Fig. 4. The time position jitter in the pulse pattern is small enough that the fundamental pulse group spacing can be determined directly from the scope. The measurement of the finesse of the cavity does depend upon the accuracy of the various sweep rate calibrations. According to the manufacturer of the oscilloscope, these sweeps, once calibrated, should remain accurate to several per cent for several months.

In order to obtain repeatable loss measurements, it is important to have the laser operate with a stable mode pattern output. At times, this was found to be difficult because adjacent longitudinal modes would

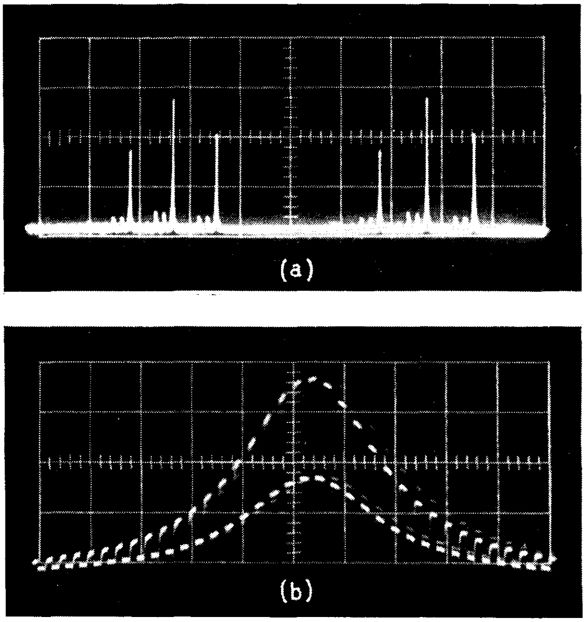


FIG. 4 — (a) Fundamental cavity spacing. Laser source has three output frequencies. Sweep rate is 2 msec per cm. (b) Half width of one of the cavity responses. Sweep rate is 10 μsec per cm. The indicated half width is 34 μsec.

compete with each other, thereby producing an erratic pulse pattern output from the interferometer. This condition was improved by reducing the length of the laser to 70 cm to separate the adjacent modes farther in frequency.

V. EXPERIMENTAL RESULTS

The reflectance was measured for a number of the multiple dielectric coated mirrors whose radius of curvature varied from 2 meters to infinity. The highest value of reflectance was 0.995 and the lowest was 0.940. The average mirror reflectance was about 0.990. The accuracy to which the reflectance of 0.990 can be repeated was about ±0.001. Since the reflectance for an individual mirror was calculated from the three loss measurements for three mirrors taken two at a time, the error in the individual mirror reflectance was probably twice that of the single measurement, or ±0.002 maximum. This leaves something to be desired. The largest source of error is in determining the half-power-height pulse width. The photographic method permits a determination of this width to about ±5 per cent under ideal conditions.

The transmission loss through a fused quartz (Homosil) Brewster angle window with fairly high-quality surfaces was found to be about 0.25 per cent. These and all other transmission loss measurements were taken with two specimens at opposite Brewster angles in the optical cavity, so that the deflection of the light ray passing through the samples canceled out. The transmission loss through the quartz was measured by comparing the loss of a sandwich of three quartz blanks to that of just two windows, where an index of refraction matching liquid was used to overcome the surface irregularities at the interfaces in both cases. This loss was about 0.2 per cent ± 0.05 per cent for a $\frac{3}{4}$ cm optical path length in the Homosil. This would give a transmission loss of about 1.0 db per meter for high-quality optical Homosil. Using the same technique, the transmission loss through cast Plexiglas II, properly annealed, is about 2 db per meter.

VI. CONCLUSIONS

The interferometer method has proven to be capable of measuring very small optical losses. It requires a number of special precautions, such as a stable mode pattern output from the laser source, a careful and correct alignment of the interferometer, all possible reduction of the effects of building vibrations, and a large degree of optical isolation between the cavity and the laser. This method measures the optical loss at only one small spot in the cavity, and this is measured using circularly polarized light. It is believed that the system accuracy can be further increased by developing an improved method of measuring the finesse of the system.

VII. ACKNOWLEDGMENTS

All computer tabulations were programmed by J. F. Ossanna, Jr., and Mrs. L. M. Calamia. The authors would like to thank D. R. Herriott and B. P. Bogert for many interesting discussions and for their invaluable help in developing this small optical loss measurement technique.

APPENDIX

The theory for the loss measuring optical cavity using plane mirrors is that of the Fabry-Perot interferometer, given many times before,[12] modified to include loss within the cavity, different reflectivities for the two mirrors, and the first-order effects of the velocity of the moving mirror. The exact theory for the effect of the moving mirror's velocity upon the interferometer response was developed. However, it was found that after neglecting some of the higher-order velocity terms, the same expression for the first-order velocity effects could be more readily obtained by assuming the mirror spacing to be fixed and by linearly varying the input frequency. Only the simpler theory will be given here. These first-order velocity effects for the interferometer, which has a Lorentzian frequency response, will be shown to be appreciably different from those previously calculated for the Gaussian filter.[13]

We assume that the incident light beam is collimated, monochromatic, and perpendicular to the mirrors. It is also assumed that the index of refraction of the space between the two mirrors is equal to that of free space, and that the mirrors have no loss. The list of all symbols and definition of all terms used in the following theory are given below:

$a = 2\pi(\Delta f/\Delta t)$ = angular sweep rate
$\alpha = (-4\pi vt/\lambda)$
B = half-power bandwidth
$\beta_0 \equiv 2m\pi v/c$
c = velocity of light in free space
d_0 = fixed mirror spacing
E_T = total combined electric field of all the output light rays assuming a unit input
g = power loss per single pass through any material within the cavity expressed as a ratio

$g_0 \equiv [g(1 - R_1)(1 - R_2)]^{\frac{1}{2}}$
I_T = total output light intensity assuming unit light input
I_m = maximum output from the interferometer assuming a unit light input
λ = free-space wavelength of light source
$m \approx (2d_0/\lambda)$ = large integer
R_1, R_2 = power reflectance of the two mirrors expressed as a ratio
t = real time
T_p = pulse width in time of the output pulse at half peak power level
T_c = fundamental pulse group spacing in time
$v = (\Delta d/\Delta t)$ = velocity of moving mirror
$x = g(R_1R_2)^{\frac{1}{2}}$.

Let the instantaneous angular frequency input be

$$\omega = \omega_0 - at. \tag{1}$$

Then the instantaneous phase is $\varphi = \omega_0 t - (at^2/2)$.

If the fixed mirror spacing is d_0 and the input light ray is normal to the mirrors as indicated in Fig. 5, then the total time delay for the nth output ray is

$$\tau_n + \tau_0 = 2n\frac{d_0}{c} + \frac{d_0}{c}$$

where the delay for the initial ray is

$$\tau_0 = \frac{d_0}{c}.$$

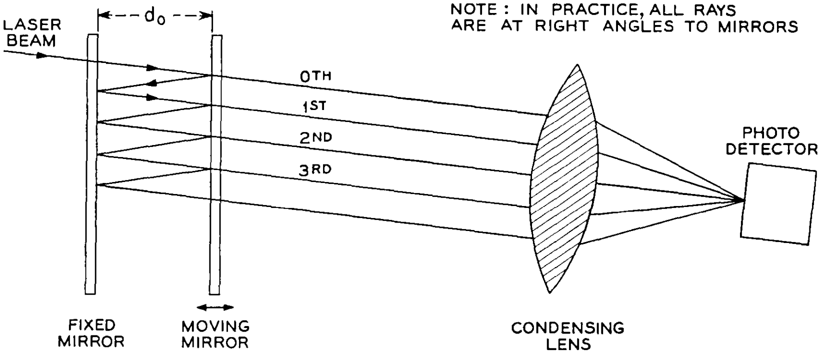


FIG. 5 — Light path through interferometer mirrors to output.

Thus the phase of the nth output ray is

$$\varphi_n = \omega_0 \left[t - \frac{2nd_0}{c} - \frac{d_0}{c} \right] - \frac{a}{2} \left[t - \frac{2nd_0}{c} - \frac{d_0}{c} \right]^2 .$$

The voltage for the nth output ray is

$$E_n = g_0 x^n \exp\left(-j\varphi_n \right).$$

The term independent of n can be neglected since its magnitude is unity. Thus the total output voltage is given by

$$E_T = \sum_{n=0}^{\infty} g_0 x^n \exp \left\{ -j \frac{2nd_0}{c} \left[\omega_0 - a\left(t - \frac{d_0}{c} \right) \right] - j2an^2 \frac{d_0^2}{c^2} \right\}. \quad (2)$$

The relation between the rate of change of the angular frequency, a, and the velocity of the mirror may be found as follows

$$a = 2\pi \frac{\Delta f}{\Delta t} = \frac{2\pi}{\Delta t} \frac{2\Delta d}{\lambda} \frac{c}{2d_0} = \frac{\omega_0 v}{d_0}. \quad (3)$$

Now, the velocity of the mirror is so small that in any reasonable length of time the variation in the mirror spacing is very small compared to the initial spacing. Hence, the exponent of (2) can be written as

$$\frac{2nd_0}{c} \left[\omega_0 - a\left(t - \frac{d_0}{c} \right) \right] + 2an^2 \frac{d_0^2}{c^2} = n[2m\pi + \alpha] + n^2 \beta_0 \quad (4)$$

where α is small,

$$m \approx 2d_0/\lambda = \text{large integer},$$

and

$$\beta_0 = 2\pi mv/c.$$

Thus from (2) and (4), the total output voltage is given by

$$E_T = \sum_{n=0}^{\infty} g_0 [x \exp(-j\alpha)] n \exp(-j\beta_0 n^2).$$

Now

$$I_T = | E_T |^2.$$

Thus

$$I_T \approx \left[\sum_{n=0}^{\infty} g_0 x^n \cos(\alpha n + \beta_0 n^2) \right]^2 + \left[\sum_{n=0}^{\infty} g_0 x^n \sin(\alpha n + \beta_0 n^2) \right]^2 . \quad (5)$$

The above expression is a good approximation to that obtained by the exact theory for $v/c < 10^{-8}$. At the present time, the series in (5) can only be summed numerically, since any other approximation results in a slowly convergent infinite series. The number of terms required by the series may be calculated as follows:

$$\frac{I_T}{g_0^2} \leqq \left[\sum_{n=0}^{\infty} x^n\right]^2 = \left[\sum_{n=0}^{N-1} x^n + \sum_{n=N}^{\infty} x^n\right]^2$$
$$\leqq \frac{(1 - x^N)^2 + 2x^N(1 - x^N) + x^{2N}}{(1 + x)^2}.$$

Thus the ratio of the total error in stopping the series after $N - 1$ terms to the actual value is

$$\epsilon = 2x^N - x^{2N}.$$

The number of terms required is then given by

$$N = \frac{\log\,(2/\epsilon)}{\log\,(1/x)} \approx \frac{\log\,(2/\epsilon)}{1 - x}. \qquad (6)$$

Thus if

$$\epsilon = 1/1000 \quad \text{and} \quad x = 0.9975,$$
$$N = 3040.$$

Hence for $x \approx 1$ a very large number of terms must be used in the series.

The series in (5) was computed for a number of values of x, and β_0, some of which are given in Fig. 1. From these values, it was found that the reduction in the maximum of the output response, I_p/I_m, and the increase in the half-height pulse width, T_m/T_p, produced by the mirror velocity can be expressed as

$$(I_m/I_p)^2 = T_m/T_p \approx \{1 + [13.8\,\beta_0^2/(1 - x)^4]\}^{\frac{1}{2}}. \qquad (7)$$

The above expression matches the computed values only when the first-order velocity effects are the order of a few per cent. For higher values, the computed results are less than those given by (7).

The relation given in (7) may be compared to that developed for a Gaussian filter which might be used in a spectrum analyzer. It has been shown[13] that the loss in sensitivity, S/S_0, and the increase in apparent bandwidth, B_m/B, produced by sweeping the frequency in the spectrum analyzer is given by

$$\left(\frac{S_0}{S}\right) = \frac{B_m}{B} = \left[1 + 0.195\left(\frac{1}{B^2}\frac{\Delta f}{\Delta t}\right)^2\right]^{\frac{1}{2}} \qquad (8)$$

where the above relation was calculated on a power basis.

In the interferometer, the mirror spacing is varied linearly with time. Hence the ratio of any two time intervals may be replaced by the ratio of their appropriate frequency differences. Thus from (3) and (11), it can be readily shown that

$$\frac{1}{B^2}\frac{\Delta f}{\Delta t} = \frac{\pi\beta_0}{(1-x)^2}\,. \tag{8a}$$

From (8a), (7) becomes

$$\left(\frac{I_m}{I_p}\right)^2 = \frac{T_m}{T_p} = \left[1 + 1.34\left(\frac{1}{B^2}\frac{\Delta f}{\Delta t}\right)^2\right]^{\frac{1}{2}}$$

which is an order of magnitude different from the relation given in (8).

When the mirror velocity is small enough to be neglected, the output from the interferometer is given by the first terms of (5). Thus for $\beta_0 = 0$,

$$I_T = \frac{g_0^2}{1 - 2x\cos\alpha + x^2} = \frac{g_0^2}{(1-x)^2 + 4x\sin^2(\alpha/2)}\,. \tag{9}$$

Now the maximum value of I_T occurs at $\alpha = 2\pi k$, $k = 0, 1, 2, 3, \cdots$, hence

$$I_m = (I_T)_{\max} = \frac{g_0^2}{(1-x)^2} \equiv \frac{g(1-R_1)(1-R_2)}{[1 - g(R_1R_2)^{\frac{1}{2}}]^2}\,. \tag{10}$$

From these relations, it can readily be shown that for x nearly unity the finesse of the system is given by[12]

$$F \equiv \frac{T_c}{T_p} = \frac{\pi\sqrt{x}}{1-x}\,.$$

The solution for x is then

$$x \equiv g(R_1R_2) \approx 1 - \frac{\pi T_p}{T_c} + \frac{1}{2}\left(\frac{\pi T_p}{T_c}\right)^2 + \cdots. \tag{11}$$

This equation relates the optical loss to measurable quantities for the pulse response of the interferometer. The time function for the output pulse can be obtained from (9),

$$I_T \approx \frac{I_m}{1 + (4t^2/T_p^2)} \tag{12}$$

for x close to unity.

REFERENCES

1. Bennett, H. E., and Koehler, W. F., J. Opt. Soc. Am., **50,** 1960, No. 1, pp. 1–6.
2. Herriott, D. R., Appl. Opt., **2,** 1963, No. 8, pp. 865–866.
3. Tolansky, S., and Bradley, D. J., Interferometry, Nat. Phys. Labs. Symposium No. 11, Her Majesty's Stationery Office, London, 1960, p. 375.
4. Herriott, D. R., and Gordon, E. I., unpublished work.
5. Fox, A. G., and Li, T., B.S.T.J., **40,** 1961, pp. 453–488.
6. Boyd, G. D., and Gordon, J. P., B.S.T.J., **40,** 1961, pp. 489–508.
7. Boyd, G. D., and Kogelnik, H., B.S.T.J., **41,** 1962, pp. 1347–1369.
8. Fork, R. L., Herriott, D. R., and Kogelnik, H., to be published.
9. Herriott, D. R., Kogelnik, H., and Kompfner, R., Appl. Opt., **3,** 1964, pp. 523–526.
10. Bennett, W. R., and Kindlmann, P. J., Rev. Sci. Instr., **33,** 1962, pp. 601–605.
11. PTZ-4 ceramic manufactured by the Clevite Corporation.
12. Born, M., and Wolf, E., *Principles of Optics*, Pergamon Press, New York, 1959, pp. 322–327.
13. Chang, S. L., Proc. I.R.E., **42,** 1954, pp. 1278–1282.

On the $\mathcal{L}_2$-Boundedness of Solutions of Nonlinear Functional Equations

By I. W. SANDBERG

(Manuscript received April 8, 1964)

Let $\mathcal{E}_N$ *denote the set of N-vector-valued functions of t defined on* $[0, \infty)$ *such that for any real positive number y, the square of the modulus of each component of any element is integrable on* $[0, y]$, *and let* $\mathcal{L}_{2N}(0, \infty)$ *denote the subset of* $\mathcal{E}_N$ *with the property that the square of the modulus of each component of any element is integrable on* $[0, \infty)$.

In the study of nonlinear physical systems, attention is frequently focused on the properties of one of the following two types of functional equations

$$g = f + \mathbf{KQ}f$$

$$g = \mathbf{K}f + \mathbf{Q}f$$

in which **K** *and* **Q** *are causal operators, with* **K** *linear and* **Q** *nonlinear,* $g \; \varepsilon \; \mathcal{E}_N$, *and f is a solution belonging to* $\mathcal{E}_N$. *Typically, f represents the system response and g takes into account both the independent energy sources and the initial conditions at* $t = 0$.

It is often important to determine conditions under which a physical system governed by one of the above equations is stable in the sense that the response to an arbitrary set of initial conditions approaches zero (i.e., the zero vector) as $t \rightarrow \infty$. *In a great many cases of this type, g belongs to* $\mathcal{L}_{2N}(0, \infty)$ *and approaches zero as* $t \rightarrow \infty$ *for all initial conditions, and, in addition, it is possible to show that if* $f \; \varepsilon \; \mathcal{L}_{2N}(0, \infty)$, *then* $f(t) \rightarrow 0$ *as* $t \rightarrow \infty$.

In this paper we attack the stability problem by deriving conditions under which $g \; \varepsilon \; \mathcal{L}_{2N} \; (0, \infty)$ *and* $f \; \varepsilon \; \mathcal{E}_N$ *imply that* $f \; \varepsilon \; \mathcal{L}_{2N}(0, \infty)$. *From an engineering viewpoint, the assumption that* $f \; \varepsilon \; \mathcal{E}_N$ *is almost invariably a trivial restriction.*

As a specific application of the results, we consider a nonlinear integral equation that governs the behavior of a general control system containing linear time-invariant elements and an arbitrary finite number of time-varying nonlinear elements. Conditions are presented under which every solution of this equation belonging to $\mathcal{E}_N$ *in fact belongs to* $\mathcal{L}_{2N}(0, \infty)$ *and approaches zero as* $t \rightarrow \infty$.

I. NOTATION AND DEFINITIONS

Let M denote an arbitrary matrix. We shall denote by M', M^*, and M^{-1}, respectively, the transpose, the complex-conjugate transpose, and the inverse of M. The positive square-root of the largest eigenvalue of M^*M is denoted by $\Lambda\{M\}$.

The set of complex measurable N-vector-valued functions of the real variable t defined on $[0,\infty)[(-\infty,\infty)]$ is denoted by $\mathcal{H}_N(0,\infty)[\mathcal{H}_N(-\infty,\infty)]$, and

$$\mathcal{L}_{2N}(0,\infty) = \left\{ f \mid f \varepsilon \mathcal{H}_N(0,\infty), \int_0^\infty f^*f\, dt < \infty \right\}.$$

In order to be consistent with standard notation, we let $\mathcal{L}_2(0,\infty) = \mathcal{L}_{2N}(0,\infty)$ when $N = 1$. We shall not distinguish between elements of $\mathcal{H}_N(0,\infty)[\mathcal{H}_N(-\infty,\infty)]$ that agree almost everywhere on $[0,\infty)$ $[(-\infty,\infty)]$. The range of any operator considered in this article is assumed to be contained in either $\mathcal{H}_N(0,\infty)$ or $\mathcal{H}_N(-\infty,\infty)$.

The inner product of two elements of $\mathcal{L}_{2N}(0,\infty)$, $f = (f_1, f_2, \cdots, f_N)'$ and $g = (g_1, g_2, \cdots, g_N)'$, is denoted by $\langle f, g\rangle$ and is defined by

$$\langle f, g\rangle = \int_0^\infty f^*g\, dt.$$

The norm of $f \varepsilon \mathcal{L}_{2N}(0,\infty)$ is denoted by $\| f \|$ and is defined by

$$\| f \| = \langle f,f\rangle^{\frac{1}{2}}.$$

The norm of a linear operator $\mathbf{T}$ defined on $\mathcal{L}_{2N}(0,\infty)$ is denoted by $\| \mathbf{T} \|$.

Let $y \varepsilon (0,\infty)$, and define f_y by

$$\begin{aligned} f_y(t) &= f(t) \qquad \text{for } t \varepsilon [0,y] \\ &= 0 \qquad \text{for } t > y \end{aligned}$$

for any $f \varepsilon \mathcal{H}_N(0,\infty)$, and let

$$\mathcal{E}_N = \{ f \mid f \varepsilon \mathcal{H}_N(0,\infty), \qquad f_y \varepsilon \mathcal{L}_{2N}(0,\infty) \text{ for } 0 < y < \infty \}.$$

The set of real vector-valued functions is denoted by $\mathcal{R}$, and $\mathbf{I}$ and 1_N, respectively, denote the identity operator on $\mathcal{L}_{2N}(0,\infty)$ and the identity matrix of order N.

With A an arbitrary measurable $N \times N$ matrix-valued function of t with elements $\{a_{nm}\}$ defined on $[0,\infty)$, let $\mathcal{K}_{pN}$ $(p = 1,2)$ denote

$$\left\{ A \mid \int_0^\infty | a_{nm}(t) |^p\, dt < \infty \qquad (n, m = 1, 2, \cdots, N) \right\}.$$

Let $\psi[f(t),t]$ denote

$$(\psi_1[f_1(t),t],\ \psi_2[f_2(t),t],\ \cdots,\ \psi_N[f_N(t),t])', \qquad f \,\varepsilon\, \mathcal{R} \cap \mathcal{K}_N(0,\infty)$$

where $\psi_1(w,t),\ \psi_2(w,t),\ \cdots,\ \psi_N(w,t)$ are real-valued functions of the real variables w and t for $-\infty < w < \infty$ and $0 \leqq t < \infty$ such that

(*i*) $\psi_n(0,t) = 0$ for $t \,\varepsilon\, [0,\infty)$ and $n = 1, 2, \cdots, N$

(*ii*) there exist real numbers α and β with the property that

$$\alpha \leqq \frac{\psi_n(w,t)}{w} \leqq \beta \qquad (n = 1, 2, \cdots, N)$$

for $t \,\varepsilon\, [0,\infty)$ and all real $w \neq 0$.

(*iii*) $\psi_n[w(t),t](n = 1, 2, \cdots, N)$ is a measurable function of t whenever $w(t)$ is measurable.

The symbol s denotes a scalar complex variable with $\sigma = \mathrm{Re}[s]$ and $\omega = \mathrm{Im}[s]$.

We shall say that a (not necessarily linear) operator $\mathbf{T}$ with domain $\mathfrak{D}(\mathbf{T}) \subset \mathcal{K}_N(0,\infty)$ is *causal* if an only if for an arbitrary $\delta > 0$,

$$(\mathbf{T}f)(t) = (\mathbf{T}g)(t) \quad \text{a.e. on } (0,\delta)$$

whenever $f,g \,\varepsilon\, \mathfrak{D}(\mathbf{T})$ and $f(t) = g(t)$ a.e. on $(0,\delta)$.

II. INTRODUCTION

In the study of nonlinear physical systems, attention is frequently focused on the properties of one of the following two types of functional equations

$$g = f + \mathbf{KQ}f \tag{1}$$

$$g = \mathbf{K}f + \mathbf{Q}f \tag{2}$$

in which $\mathbf{K}$ and $\mathbf{Q}$ are causal operators, with $\mathbf{K}$ linear and $\mathbf{Q}$ nonlinear, $g \,\varepsilon\, \mathcal{E}_N$, and f is a solution belonging to $\mathcal{E}_N$. Typically, f represents the system response and g takes into account both the independent energy sources and the initial conditions at $t = 0$.

It is often important to determine conditions under which a physical system governed by one of the above equations is stable in the sense that the response to an arbitrary set of initial conditions approaches zero (i.e., the zero vector) as $t \rightarrow \infty$. In a great many cases of this type, g belongs to $\mathcal{L}_{2N}(0,\infty)$ and approaches zero as $t \rightarrow \infty$ for all initial conditions, and, in addition, it is possible to show that if $f \,\varepsilon\, \mathcal{L}_{2N}(0,\infty)$, then $f(t) \rightarrow 0$ as $t \rightarrow \infty$.

In this paper we attack the stability problem by deriving conditions under which $g \varepsilon \mathcal{L}_{2N}(0, \infty)$ and $f \varepsilon \mathcal{E}_N$ imply that $f \varepsilon \mathcal{L}_{2N}(0, \infty)$. From an engineering viewpoint, the assumption that $f \varepsilon \mathcal{E}_N$ is almost invariably a trivial restriction.

As a specific application of the abstract results of Section III, we consider, in Section IV, the following integral equation which governs the behavior of a general control system containing linear time-invariant elements and an arbitrary finite number of time-varying nonlinear elements:

$$g(t) = f(t) + \int_0^t k(t - \tau)\psi[f(\tau), \tau]\, d\tau, \qquad t \geqq 0 \tag{3}$$

in which $k \varepsilon \mathcal{K}_{1N} \cap \mathcal{K}_{2N}$, $\psi[\cdot, \cdot]$ is as defined in Section I, and $g \varepsilon \mathcal{L}_{2N}(0, \infty)$. We prove that every solution f of (3) belonging to $\mathcal{R} \cap \mathcal{E}_N$ in fact belongs to $\mathcal{L}_{2N}(0, \infty)$ and approaches zero as $t \rightarrow \infty$ if, with

$$K(s) = \int_0^\infty k(t)e^{-st}\, dt \qquad \text{for} \qquad \sigma \geqq 0,$$

(*i*) $\det [1_N + \frac{1}{2}(\alpha + \beta)K(s)] \neq 0$ for $\sigma \geqq 0$

(*ii*) $\frac{1}{2}(\beta - \alpha) \sup_{\omega} \Lambda\{[1_N + \frac{1}{2}(\alpha + \beta)K(i\omega)]^{-1}K(i\omega)\} < 1.$

An analogous result is proved for the integral equation

$$g(t) = \psi[f(t), t] + \int_0^t k(t - \tau)f(\tau)\, d\tau, \qquad t \geqq 0.$$

For $N = 1$, the key condition (*ii*) possesses a simple geometric interpretation: it is satisfied if and only if the locus of $[K(i\omega)]^{-1}$ for $-\infty < \omega < \infty$ lies outside the circle of radius $\frac{1}{2}(\beta - \alpha)$ centered in the complex plane at $[-\frac{1}{2}(\alpha + \beta), 0]$.†

In Section V we consider two direct applications to nonlinear differential equations. One of our results asserts that if f is any real-valued function of t defined and twice-differentiable on $[0, \infty)$ such that

$$\frac{d^2f}{dt^2} + a\frac{df}{dt} + \psi[f, t] = g$$

for almost all $t \varepsilon [0, \infty)$, where $g \varepsilon \mathcal{R} \cap \mathcal{L}_2(0, \infty)$, $\psi[\cdot, \cdot]$ is as defined in Section I with $N = 1$ and $\alpha > 0$, and a is a real constant such that $a > \sqrt{\beta} - \sqrt{\alpha}$, then $f \varepsilon \mathcal{L}_2(0, \infty)$ and $f(t) \rightarrow 0$ as $t \rightarrow \infty$.

† For some earlier results concerned with frequency-domain conditions for the stability of nonlinear or time-varying systems, see Refs. 1–4.

III. KEY RESULTS

Assumption 1: It is assumed throughout that

(i) $\mathbf{K}$ is a linear causal operator with domain $\mathfrak{D}(\mathbf{K})$ such that $\mathcal{L}_{2N}(0,\infty) \subset \mathfrak{D}(\mathbf{K}) \subset \mathcal{H}_N(0,\infty)$

(ii) $\mathbf{K}$ maps $\mathcal{L}_{2N}(0,\infty)$ into itself, and is bounded on $\mathcal{L}_{2N}(0,\infty)$

(iii) $\mathbf{Q}$ is a (not necessarily linear) causal operator with domain $\mathfrak{D}(\mathbf{Q}) \subset \mathcal{H}_N(0,\infty)$.

The following two theorems are the key results of the paper.

Theorem 1: Let $f \,\varepsilon\, \mathfrak{D}(\mathbf{Q}) \cap \mathcal{E}_N$ such that $\mathbf{Q}f \,\varepsilon\, \mathfrak{D}(\mathbf{K}) \cap \mathcal{E}_N$, $\mathbf{KQ}f \,\varepsilon\, \mathcal{E}_N$, and $g = f + \mathbf{KQ}f$, where $g \,\varepsilon\, \mathcal{L}_{2N}(0,\infty)$. Let f not be the zero-element of $\mathcal{E}_N$, and let $y_0 = inf\,\{y \mid y > 0,\ \|f_y\| \neq 0\}$.

Suppose that $\{f_y\,,\ 0 < y < \infty\} \subset \mathfrak{D}(\mathbf{Q})$ and that there exists a real or complex number x such that

(i) on $\mathcal{L}_{2N}(0,\infty)$, $(\mathbf{I} + x\mathbf{K})^{-1}$ exists and is causal

$$(ii)\ \| (\mathbf{I} + x\mathbf{K})^{-1}\mathbf{K} \| \sup_{y>y_0} \frac{\| (\mathbf{Q}f_y)_y - xf_y \|}{\| f_y \|} < 1.$$

Then $f \,\varepsilon\, \mathcal{L}_{2N}(0,\infty)$ and

$$\| f \| \leqq (1 - r)^{-1} \| (\mathbf{I} + x\mathbf{K})^{-1} g \|,$$

in which

$$r = \| (\mathbf{I} + x\mathbf{K})^{-1}\mathbf{K} \| \sup_{y>y_0} \frac{\| (\mathbf{Q}f_y)_y - xf_y \|}{\| f_y \|}.$$

Theorem 2: Let $f \,\varepsilon\, \mathfrak{D}(\mathbf{Q}) \cap \mathfrak{D}(\mathbf{K}) \cap \mathcal{E}_N$ such that $\mathbf{K}f \,\varepsilon\, \mathcal{E}_N$, $\mathbf{Q}f \,\varepsilon\, \mathcal{E}_N$, and

$$g = \mathbf{K}f + \mathbf{Q}f$$

where $g \,\varepsilon\, \mathcal{L}_{2N}(0,\infty)$. Let f not be the zero-element of $\mathcal{E}_N$, and let $y_0 = inf\,\{y \mid y > 0,\ \|f_y\| \neq 0\}$.

Suppose that $\{f_y\,,\ 0 < y < \infty\} \subset \mathfrak{D}(\mathbf{Q})$ and that there exists a real or complex number x such that

(i) on $\mathcal{L}_{2N}(0,\infty)$, $(x\mathbf{I} + \mathbf{K})^{-1}$ exists and is causal

$$(ii)\ \| (x\mathbf{I} + \mathbf{K})^{-1} \| \sup_{y>y_0} \frac{\| (\mathbf{Q}f_y)_y - xf_y \|}{\| f_y \|} < 1.$$

Then $f \,\varepsilon\, \mathcal{L}_{2N}(0,\infty)$ and

$$\| f \| \leqq (1 - q)^{-1} \| (x\mathbf{I} + \mathbf{K})^{-1} g \|,$$

in which

$$q = \| (x\mathbf{I} + \mathbf{K})^{-1} \| \sup_{y>y_0} \frac{\| (\mathbf{Q}f_y)_y - xf_y \|}{\| f_y \|}.$$

3.1 *Proof of Theorem 1*

It is convenient to introduce the operator **P** defined on $\mathcal{H}_N(0,\infty)$ by $\mathbf{P}f = f_y$, where y is an arbitrary real positive number.

From $g = f + \mathbf{KQ}f$, we clearly have

$$g_y = f_y + \mathbf{PKQ}f.$$

Since **K** is causal,

$$g_y = f_y + \mathbf{PKPQ}f.$$

Similarly, since **Q** is causal,

$$\begin{aligned} g_y &= f_y + \mathbf{PKPQP}f \\ &= f_y + \mathbf{PKPQ}f_y . \end{aligned}$$

Thus,

$$g_y = \mathbf{P}(\mathbf{I} + x\mathbf{K})f_y + \mathbf{PKP}(\mathbf{Q} - x\mathbf{I})f_y .$$

Since on $\mathcal{L}_{2N}(0,\infty)$, $(\mathbf{I} + x\mathbf{K})^{-1}$ exists and is causal,

$$\mathbf{P}(\mathbf{I} + x\mathbf{K})^{-1}\mathbf{P}(\mathbf{I} + x\mathbf{K})f_y = f_y ,$$

and hence,

$$f_y = -\mathbf{P}(\mathbf{I} + x\mathbf{K})^{-1}\mathbf{PKP}(\mathbf{Q} - x\mathbf{I})f_y + \mathbf{P}(\mathbf{I} + x\mathbf{K})^{-1}g_y .$$

It follows that

$$\| f_y \| \leqq \| \mathbf{P}(\mathbf{I} + x\mathbf{K})^{-1}\mathbf{PK} \| \cdot \| \mathbf{PQ}f_y - xf_y \| + \| \mathbf{P}(\mathbf{I} + x\mathbf{K})^{-1}g_y \|.$$

Moreover, in view of the causality of $(\mathbf{I} + x\mathbf{K})^{-1}$,

$$\mathbf{P}(\mathbf{I} + x\mathbf{K})^{-1}\,\mathbf{PK} = \mathbf{P}(\mathbf{I} + x\mathbf{K})^{-1}\mathbf{K},$$

and hence, using the fact that **P** is a projection on $\mathcal{L}_{2N}(0,\infty)$,

$$\| \mathbf{P}(\mathbf{I} + x\mathbf{K})^{-1}\mathbf{PK} \| \leqq \| (\mathbf{I} + x\mathbf{K})^{-1}\mathbf{K} \|.$$

Similarly,

$$\| \mathbf{P}(\mathbf{I} + x\mathbf{K})^{-1}g_y \| \leqq \| (\mathbf{I} + x\mathbf{K})^{-1}g \|.$$

Thus, with r as defined in the statement of the theorem,

$$\| f_y \| \leqq r \| f_y \| + \| (\mathbf{I} + x\mathbf{K})^{-1}g \|$$

or

$$\| f_y \| \leqq (1 - r)^{-1} \| (\mathbf{I} + x\mathbf{K})^{-1}g \|.$$

Since this inequality is valid for arbitrary positive y, it follows that $f \varepsilon \mathcal{L}_{2N}(0, \infty)$ and

$$\| f \| \leqq (1 - r)^{-1} \| (\mathbf{I} + x\mathbf{K})^{-1} g \|.$$

3.2 *Proof of Theorem 2*

The argument is essentially the same as the one used in the proof of Theorem 1.

We have, with **P** as defined in the proof of Theorem 1,

$$\begin{aligned} g_y &= \mathbf{PK}f + \mathbf{PQ}f \\ &= \mathbf{PKP}f + \mathbf{PQP}f = \mathbf{PK}f_y + \mathbf{PQ}f_y \\ &= \mathbf{P}(x\mathbf{I} + \mathbf{K})f_y + \mathbf{P}(\mathbf{Q} - x\mathbf{I})f_y \,. \end{aligned}$$

Using the fact that on $\mathcal{L}_{2N}(0, \infty)$, $(x\mathbf{I} + \mathbf{K})^{-1}$ exists and is causal

$$f_y = -\mathbf{P}(x\mathbf{I} + \mathbf{K})^{-1}\mathbf{P}(\mathbf{Q} - x\mathbf{I})f_y + \mathbf{P}(x\mathbf{I} + \mathbf{K})^{-1}g_y \,.$$

Thus, with q as defined in the statement of the theorem,

$$\| f_y \| \leqq q \| f_y \| + \| (x\mathbf{I} + \mathbf{K})^{-1} g \|,$$

or

$$\| f_y \| \leqq (1 - q)^{-1} \| (x\mathbf{I} + \mathbf{K})^{-1} g \|.$$

This inequality is valid for arbitrary positive y. Hence $f \varepsilon \mathcal{L}_{2N}(0, \infty)$ and

$$\| f \| \leqq (1 - q)^{-1} \| (x\mathbf{I} + \mathbf{K})^{-1} g \|.$$

Remark: A moment's reflection concerning the proofs of Theorems 1 and 2 will show that by simply reinterpreting the symbols, analogous results can be obtained for other function spaces.

3.3 *Conditions under Which the Hypotheses of Theorems 1 or 2 Concerning x Are Satisfied*

The following theorem asserts that the hypotheses of Theorems 1 or 2 concerning x are satisfied in certain special but very important cases. The implications of the theorem are of direct interest in the theory of passive nonlinear electrical networks.

Theorem 3: Let f be as defined in Theorem 1 [*Theorem 2*]. *Suppose that*

(*i*) *there exist a nonnegative constant* k_1 *and a positive constant* k_2 *such that*

$$Re \langle \mathbf{Q}f_y, f_y \rangle \geqq k_1 \| f_y \|^2, \qquad \| \mathbf{Q}f_y \|^2 \leqq k_2 \| f_y \|^2 \qquad \text{for } 0 < y < \infty$$

(ii) **K** *maps* $\mathcal{L}_{2N}(0,\infty)$ *into itself such that there exists a nonnegative constant c with the property that*

$$Re\,\langle \mathbf{K}h,h\rangle \geqq c\,\|\,h\,\|^2$$

for all $h\ \varepsilon\ \mathcal{L}_{2N}(0,\infty)$.

Then the hypotheses concerning x of Theorem 1 [Theorem 2] are satisfied if either:

$$k_1 > 0 \quad and \quad c \geqq 0,$$

or

$$k_1 = 0 \quad and \quad c > 0.$$

3.4 *Proof of Theorem 3*

Lemmas 1, 2, and 3 (below) imply that for real positive x the operators $(\mathbf{I} + x\mathbf{K})$ and $(x\mathbf{I} + \mathbf{K})$ possess causal inverses on $\mathcal{L}_{2N}(0,\infty)$ and

$$\|\,(\mathbf{I} + x\mathbf{K})^{-1}\mathbf{K}\,\|^2 \leqq \frac{2(\|\,\mathbf{K}\,\| - c) + x\,\|\,\mathbf{K}\,\|^2}{x(1 + x\,\|\,\mathbf{K}\,\|)^2}$$

$$\|\,(x\mathbf{I} + \mathbf{K})^{-1}\,\|^2 \leqq \frac{1}{x(x + 2c)}\,.$$

With x real and positive,

$$\begin{aligned}\|\,(\mathbf{Q}f_y)_y - xf_y\,\|^2 &\leqq \|\,\mathbf{Q}f_y - xf_y\,\|^2\\ &\leqq \|\,\mathbf{Q}f_y\,\|^2 - 2x\,\mathrm{Re}\,\langle \mathbf{Q}f_y\,,f_y\rangle + x^2\,\|\,f_y\,\|^2\\ &\leqq (k_2 - 2xk_1 + x^2)\,\|\,f_y\|^2.\end{aligned}$$

It is a simple matter to verify that if $(k_1 + c) > 0$, there exist positive values of x such that

$$\frac{2(\|\,\mathbf{K}\,\| - c) + x\,\|\,\mathbf{K}\,\|^2}{x(1 + x\,\|\,\mathbf{K}\,\|)^2}\,(k_2 - 2xk_1 + x^2) < 1,$$

and there exist positive values of x such that

$$\frac{k_2 - 2xk_1 + x^2}{x(x + 2c)} < 1.$$

Hence, it remains to prove Lemmas 1, 2, and 3.

Lemma 1: Let **T** *be a bounded linear mapping of* $\mathcal{L}_{2N}(0,\infty)$ *into itself* *such that there exists a constant* $c_1 > -1$ *with the property that*

$$Re \langle \mathbf{T}f,f \rangle \geqq c_1 \| f \|^2$$

for all $f \; \varepsilon \; \mathfrak{L}_{2N}(0,\infty)$. *Then* $(\mathbf{I} + \mathbf{T})$ *possesses an inverse on* $\mathfrak{L}_{2N}(0,\infty)$.
Proof:

Since $c_1 > -1$, it is evident that there exists a positive constant k_1 such that

$$\text{Re} \langle (\mathbf{I} + \mathbf{T})f,f \rangle \geqq k_1 \| f \|^2$$

for all $f \; \varepsilon \; \mathfrak{L}_{2N}(0, \infty)$. This, together with the boundedness of $\mathbf{T}$, implies that $(\mathbf{I} + \mathbf{T})^{-1}$ exists (see Ref. 5, for example).

Lemma 2: Let $\mathbf{T}$ *be an invertible bounded linear mapping of* $\mathfrak{L}_{2N}(0,\infty)$ *into itself such that* $\mathbf{T}$ *is causal and* $Re \langle \mathbf{T}h,h \rangle \geqq 0$ *for all* $h \; \varepsilon \; \mathfrak{L}_{2N}(0,\infty)$. *Then* $\mathbf{T}^{-1}$ *is causal.*
Proof:

A bounded linear mapping $\mathbf{A}$ of $\mathfrak{L}_{2N}(0,\infty)$ into itself is causal if and only if[6]

$$\text{Re} \int_0^y (\mathbf{A}f)^* f \, dt \geqq - \| \mathbf{A} \| \int_0^y f^* f \, dt$$

for all real $y \geqq 0$ and all $f \; \varepsilon \; \mathfrak{L}_{2N}(0,\infty)$. Thus, to prove the lemma it suffices to point out that the causality of $\mathbf{T}$ implies that

$$\text{Re} \int_0^y (\mathbf{T}g)^* g \, dt = \text{Re} \langle \mathbf{T}g_y , g_y \rangle \geqq 0,$$

for all real $y \geqq 0$ and all $g \; \varepsilon \; \mathfrak{L}_{2N}(0,\infty)$, and hence that

$$\text{Re} \int_0^y h^* \mathbf{T}^{-1} h \, dt \geqq 0$$

for all real $y \geqq 0$ and all $h \; \varepsilon \; \mathfrak{L}_{2N}(0,\infty)$.

Lemma 3:† *Let* $\mathbf{T}$ *be a bounded linear mapping of* $\mathfrak{L}_{2N}(0,\infty)$ *into itself such that* $(\mathbf{I} + \mathbf{T})$ *is invertible and there exists a real constant* c_2 *with the property that*

$$Re \langle \mathbf{T}f,f \rangle \geqq c_2 \| f \|^2$$

for all $f \; \varepsilon \; \mathfrak{L}_{2N}(0,\infty)$. *Then, for* $c_2 \geqq -\frac{1}{2}$,

† Lemmas 1 and 3, and their proofs, remain valid if $\mathfrak{L}_{2N}(0,\infty)$ is replaced with an arbitrary Hilbert space with inner product $\langle \cdot,\cdot \rangle$ and norm $\| \cdot \|$.

$$\| (\mathbf{I} + \mathbf{T})^{-1}\mathbf{T} \| \leqq [1 - (2c_2 + 1)(1 + \| \mathbf{T} \|)^{-2}]^{\frac{1}{2}}$$

and, for $c_2 > -\frac{1}{2}$,

$$\| (\mathbf{I} + \mathbf{T})^{-1} \| \leqq (1 + 2c_2)^{-\frac{1}{2}}.$$

Proof:

In order to establish the first inequality, let $g = (\mathbf{I} + \mathbf{T})^{-1}\mathbf{T}f$ and, using the fact that $g = f - (\mathbf{I} + \mathbf{T})^{-1}f$, observe that

$$\langle g,g \rangle = \langle f,f \rangle - 2 \operatorname{Re} \langle \mathbf{T}z,z \rangle - \langle z,z \rangle,$$

where $z = (\mathbf{I} + \mathbf{T})^{-1}f$. Since

$$2 \operatorname{Re} \langle \mathbf{T}z,z \rangle + \langle z,z \rangle \geqq (2c_2 + 1) \| z \|^2$$

and

$$\| z \| \geqq \| (\mathbf{I} + \mathbf{T}) \|^{-1} \| f \| \geqq (1 + \| T \|)^{-1} \| f \|,$$

it follows that

$$\langle g,g \rangle \leqq [1 - (2c_2 + 1)(1 + \| \mathbf{T} \|)^{-2}]\langle f,f \rangle$$

for all f, $g \; \varepsilon \; \mathcal{L}_{2N}(0, \infty)$ such that $g = (\mathbf{I} + \mathbf{T})^{-1}\mathbf{T}f$. Thus

$$\| (\mathbf{I} + \mathbf{T})^{-1}\mathbf{T} \| \leqq [1 - (2c_2 + 1)(1 + \| \mathbf{T} \|)^{-2}]^{\frac{1}{2}}.$$

The second inequality follows directly from the fact that if $g = (\mathbf{I} + \mathbf{T})^{-1}f$,

$$\| f \|^2 = \| g \|^2 + 2 \operatorname{Re} \langle \mathbf{T}g,g \rangle + \| \mathbf{T}g \|^2 \geqq (1 + 2c_2) \| g \|^2.$$

IV. APPLICATIONS TO NONLINEAR INTEGRAL EQUATIONS

In this section our primary objective is to prove the following two theorems.

Theorem 4: Let $k \; \varepsilon \; \mathcal{K}_{1N}$ *and let*

$$g(t) = f(t) + \int_0^t k(t - \tau)\psi[f(\tau), \tau] \, d\tau, \qquad t \geqq 0$$

where $g \; \varepsilon \; \mathcal{L}_{2N}(0, \infty)$ *and* $f \; \varepsilon \; \mathcal{R} \cap \mathcal{E}_N$. *Let*

$$K(s) = \int_0^\infty k(t)e^{-st} \, dt, \qquad \sigma \geqq 0.$$

Suppose that

(*i*) *det* $[1_N + \frac{1}{2}(\alpha + \beta)K(s)] \neq 0$ *for* $\sigma \geqq 0$

(*ii*) $\frac{1}{2}(\beta - \alpha) \; \sup_\omega \; \Lambda\{[1_N + \frac{1}{2}(\alpha + \beta)K(i\omega)]^{-1}K(i\omega)\} < 1.$

Then $f \varepsilon \mathcal{L}_{2N}(0, \infty)$.

Theorem 5: Let $k \varepsilon \mathcal{K}_{1N}$ and let

$$g(t) = \int_0^t k(t - \tau)f(\tau)\, d\tau + \psi[f(t), t], \qquad t \geqq 0$$

where $g \varepsilon \mathcal{L}_{2N}(0, \infty)$ and $f \varepsilon \mathcal{R} \cap \mathcal{E}_N$. Let

$$K(s) = \int_0^\infty k(t)e^{-st}\, dt, \qquad \sigma \geqq 0.$$

Suppose that

(i) det $[\frac{1}{2}(\alpha + \beta)1_N + K(s)] \neq 0$ for $\sigma \geqq 0$

(ii) $\frac{1}{2}(\beta - \alpha) \sup_\omega \Lambda\{[\frac{1}{2}(\alpha + \beta)1_N + K(i\omega)]^{-1}\} < 1$.

Then $f \varepsilon \mathcal{L}_{2N}(0, \infty)$.

4.1 *Proof of Theorems 4 and 5*

In Theorems 1 and 2 let **Q** denote the operator defined by

$$(\mathbf{Q}g)(t) = \psi[g(t), t], \qquad 0 \leqq t < \infty$$

where g is an arbitrary element of $\mathcal{R} \cap \mathcal{H}_N(0, \infty)$. This operator maps $\mathcal{R} \cap \mathcal{L}_{2N}(0, \infty)$ into itself and possesses the property that for any real x

$$\| \mathbf{Q}h - xh \| \leqq \eta(x) \; \| h \|, \qquad h \varepsilon \mathcal{R} \cap \mathcal{L}_{2N}(0, \infty)$$

where

$$\eta(x) = \max\, [(x - \alpha), (\beta - x)].$$

Thus, with **K** defined on $\mathcal{L}_{2N}(0, \infty)$ by[7]

$$\mathbf{K}h = \int_0^t k(t - \tau)h(\tau)\, d\tau, \qquad h \varepsilon \mathcal{L}_{2N}(0, \infty),$$

condition (*ii*) of Theorem 1 and the corresponding condition of Theorem 2, respectively, are satisfied if there exists a real x such that

$$\| (\mathbf{I} + x\mathbf{K})^{-1}\mathbf{K} \| \; \eta(x) < 1,$$

and

$$\| (x\mathbf{I} + \mathbf{K})^{-1} \| \; \eta(x) < 1.$$

Lemmas 4 and 5 (below) imply at once that if the assumptions of Theorem 4 (Theorem 5) are met, then hypotheses (*i*) and (*ii*) of Theorem 1 (Theorem 2) are satisfied with $x = \frac{1}{2}(\alpha + \beta)$. It can be shown[8] (with the aid of Lemma 4) that this choice of x is optimal in the sense that if there exists a real x such that $(\mathbf{I} + x\mathbf{K})$ possesses a causal inverse on $\mathcal{L}_{2N}(0, \infty)$ and

$$\| (\mathbf{I} + x\mathbf{K})^{-1}\mathbf{K} \| \, \eta(x) < 1,$$

then $[\mathbf{I} + \frac{1}{2}(\alpha + \beta)\mathbf{K}]$ possesses a causal inverse on $\mathcal{L}_{2N}(0, \infty)$ and

$$\| (\mathbf{I} + x\mathbf{K})^{-1}\mathbf{K} \| \, \eta(x) \geqq \| [\mathbf{I} + \tfrac{1}{2}(\alpha + \beta)\mathbf{K}]^{-1}\mathbf{K} \| \, \eta[\tfrac{1}{2}(\alpha + \beta)].$$

This choice of x is similarly optimal with regard to the statement of the conditions in Theorem 5.

Before proceeding to the statement and proofs of the lemmas, it is convenient to introduce a few definitions.

4.2 *Definitions*

With τ an arbitrary positive constant, let $\mathbf{S}_\tau$ denote the mapping of $\mathcal{L}_{2N}(0, \infty)$ into itself defined by

$$\begin{aligned} (\mathbf{S}_\tau f)(t) &= 0, & t \,\varepsilon\, [0, \tau) \\ &= f(t - \tau), & t \,\varepsilon\, [\tau, \infty) \end{aligned}$$

for any $f \,\varepsilon\, \mathcal{L}_{2N}\,(0, \infty)$.

Let

$$\mathcal{L}_{2N}(-\infty, \infty) = \left\{ f \mid f \,\varepsilon\, \mathcal{H}_N(-\infty, \infty), \int_{-\infty}^{\infty} f^* f \, dt < \infty \right\}.$$

We take as the definition of the Fourier transform of $f \,\varepsilon\, \mathcal{L}_{2N}(-\infty, \infty)$:

$$\hat{f} = \text{l.i.m.} \int_{-\infty}^{\infty} f(t) e^{-i\omega t} \, dt,$$

and consequently,

$$2\pi \int_{-\infty}^{\infty} f^* f \, dt = \int_{-\infty}^{\infty} \hat{f}^* \hat{f} \, d\omega.$$

By the Fourier transform of an $f \,\varepsilon\, \mathcal{L}_{2N}(0, \infty)$ we mean simply

$$\text{l.i.m.} \int_0^{\infty} f(t) e^{-i\omega t} \, dt.$$

4.3 *Lemmas 4 and 5*

Lemma 4: Let $\mathbf{A}$ *be an invertible linear mapping of* $\mathcal{L}_{2N}(0, \infty)$ *into itself* *such that for an arbitrary $f \varepsilon \mathfrak{L}_{2N}(0,\infty)$*

$$\mathbf{A}\mathbf{S}_\tau f = \mathbf{S}_\tau \mathbf{A} f, \qquad \tau > 0.$$

Then $\mathbf{A}^{-1}$ *is causal.*

Proof:

Suppose that, on the contrary, $\mathbf{A}$ possesses an inverse on $\mathfrak{L}_{2N}(0,\infty)$, but that it is not causal. Then there exist elements z_1, $z_2 \varepsilon \mathfrak{L}_{2N}(0,\infty)$, and a $\delta > 0$ such that $z_1(t) \neq 0$ on some positive-measure subset of $(0,\delta)$, and $\mathbf{A}z_1 = \mathbf{S}_\delta z_2$. Since $\mathbf{A}$ is assumed to possess an inverse, there exists a unique $z_3 \varepsilon \mathfrak{L}_{2N}(0,\infty)$ such that $z_2 = \mathbf{A}z_3$. Thus,

$$\mathbf{A}\mathbf{S}_\delta z_3 = \mathbf{S}_\delta \mathbf{A} z_3 = \mathbf{S}_\delta z_2 .$$

Clearly, $\mathbf{S}_\delta z_3 \neq z_1$, which contradicts the assumption that $\mathbf{A}$ possesses an inverse. This proves the lemma.

Lemma 5: Let $u \varepsilon \mathfrak{K}_{1N}$ and let $\mathbf{U}$ *be the mapping of $\mathfrak{L}_{2N}(0,\infty)$ into itself defined by*

$$\mathbf{U}f = \int_0^t u(t-\tau)f(\tau)\,d\tau, \qquad f \varepsilon \mathfrak{L}_{2N}(0,\infty).$$

Let

$$U(s) = \int_0^\infty u(t)e^{-st}\,dt, \qquad \sigma \geqq 0.$$

Suppose that det $[1_N + U(s)] \neq 0$ for $\sigma \geqq 0$. Then

(i) $(\mathbf{I} + \mathbf{U})$ *possesses an inverse on $\mathfrak{L}_{2N}(0,\infty)$*

(ii) $$\|(\mathbf{I}+\mathbf{U})^{-1}\mathbf{U}\| \leqq \sup_\omega \Lambda\{[1_N + U(i\omega)]^{-1}U(i\omega)\}$$
$$\|(\mathbf{I}+\mathbf{U})^{-1}\| \leqq \sup_\omega \Lambda\{[1_N + U(i\omega)]^{-1}\}.$$

Proof:

Consider first the invertibility of the operator $(\bar{\mathbf{I}} + \bar{\mathbf{U}})$ defined on $\mathfrak{L}_{2N}(-\infty,\infty)$ by

$$(\bar{\mathbf{I}} + \bar{\mathbf{U}})f = f + \int_{-\infty}^t u(t-\tau)f(\tau)\,d\tau, \qquad f \varepsilon \mathfrak{L}_{2N}(-\infty,\infty).$$

The assumption that $u \varepsilon \mathfrak{K}_{1N}$ implies that the elements of $U(i\omega)$ approach zero as $|\omega| \rightarrow \infty$, and that they are uniformly bounded and uniformly continuous for $\omega \varepsilon (-\infty,\infty)$. Thus, det $[1_N + U(i\omega)]$ approaches unity as $|\omega| \rightarrow \infty$, and is uniformly continuous for

$\omega \,\varepsilon\, (-\infty,\infty)$. It follows that $\det\,[1_N + U(i\omega)] \neq 0$ for all ω implies that

$$\inf_{\omega} |\det\,[1_N + U(i\omega)]| > 0,$$

and hence that

$$\sup_{\omega} \Lambda\{[1_N + U(i\omega)]^{-1}\} < \infty.$$

Let $\hat{g}$ denote the Fourier transform of an arbitrary $g \,\varepsilon\, \mathfrak{L}_{2N}(-\infty,\infty)$. Then,

$$\int_{-\infty}^{\infty} \hat{g}^*[1_N + U(i\omega)]^{-1*}[1_N + U(i\omega)]^{-1}\hat{g}\, d\omega$$

$$\leqq \int_{-\infty}^{\infty} \Lambda^2\{[1_N + U(i\omega)]^{-1}\}\hat{g}^*\hat{g}\, d\omega$$

$$\leqq \sup_{\omega} \Lambda^2\{[1_N + U(i\omega)]^{-1}\} \int_{-\infty}^{\infty} \hat{g}^*\hat{g}\, d\omega < \infty,$$

and hence, by the Riesz-Fischer theorem, there exists an $f \,\varepsilon\, \mathfrak{L}_{2N}(-\infty,\infty)$ with Fourier transform

$$\hat{f} = [1_N + U(i\omega)]^{-1}\hat{g}.$$

This establishes the existence of $(\bar{\mathbf{I}} + \bar{\mathbf{U}})^{-1}$.

Since $\det\,[1_N + U(s)] \neq 0$ for $\sigma \geqq 0$, and $U(s) \to 0$ as $|s| \to \infty$ uniformly in the closed right-half plane, every element of $[1_N + U(s)]^{-1}$ is analytic and uniformly bounded for $\sigma > 0$. Thus, $(\bar{\mathbf{I}} + \bar{\mathbf{U}})^{-1}$ maps

$$\{f \mid f \,\varepsilon\, \mathfrak{L}_{2N}(-\infty,\infty), f(t) = 0 \quad \text{for} \quad t < 0\}$$

into itself,[9,10] and hence the operator $(\mathbf{I} + \mathbf{U})$ defined on $\mathfrak{L}_{2N}(0,\infty)$ possesses an inverse.

To establish the first of the inequalities stated in the lemma, let $f \,\varepsilon\, \mathfrak{L}_{2N}(0,\infty)$ and let

$$g = (\mathbf{I} + \mathbf{U})^{-1}\mathbf{U}f.$$

Then, with $\hat{g}$ and $\hat{f}$, respectively, the Fourier transforms of g and f,

$$\hat{g} = [1_N + U(i\omega)]^{-1}U(i\omega)\hat{f}.$$

Thus,

$$\int_{-\infty}^{\infty} \hat{g}^*\hat{g}\, d\omega = \int_{-\infty}^{\infty} \hat{f}^*U(i\omega)^*[1_N + U(i\omega)]^{-1*}[1_N + U(i\omega)]^{-1}U(i\omega)\hat{f}\, d\omega$$

$$\leqq \int_{-\infty}^{\infty} \Lambda^2\{[1_N + U(i\omega)]^{-1}U(i\omega)\}\hat{f}^*\hat{f}\, d\omega$$

$$\leqq \sup_{\omega} \Lambda^2\{[1_N + U(i\omega)]^{-1}U(i\omega)\} \int_{-\infty}^{\infty} \hat{f}^*\hat{f}\, d\omega,$$

from which, using Plancherel's identity,

$$\| g \| \leqq \sup_{\omega} \Lambda\{[1_N + U(i\omega)]^{-1}U(i\omega)\} \ \| f \|.$$

Thus,

$$\| (\mathbf{I} + \mathbf{U})^{-1}\mathbf{U} \| \leqq \sup_{\omega} \Lambda\{[1_N + U(i\omega)]^{-1}U(i\omega)\}.$$

By simply repeating this argument with $(\mathbf{I} + \mathbf{U})^{-1}\mathbf{U}$ and $[1_N + U(i\omega)]^{-1} \cdot U(i\omega)$, respectively, replaced with $(\mathbf{I} + \mathbf{U})^{-1}$ and $[1_N + U(i\omega)]^{-1}$, we find that

$$\| (\mathbf{I} + \mathbf{U})^{-1} \| \leqq \sup_{\omega} \Lambda\{[1_N + U(i\omega)]^{-1}\}.$$

This proves the lemma.

4.4 *Remarks*

It can easily be shown that conditions (*i*) and (*ii*) of Theorems 4 and 5 are satisfied if $\alpha > 0$ and

$$K(i\omega) + K(i\omega)^*$$

is nonnegative definite for all ω.

A moment's reflection concerning the proof of Theorems 4 and 5 will show that those theorems remain valid if $\psi[f(t),t]$ and f, respectively, are replaced with $(\mathbf{Q}f)(t)$ and $f \,\varepsilon\, \mathfrak{D}(\mathbf{Q})$, with the understanding that

(a) $\mathbf{Q}$ is a causal mapping of a subset $\mathfrak{D}(\mathbf{Q})$ of $\mathcal{E}_N$ into $\mathcal{E}_N$ such that $\mathbf{Q}h \,\varepsilon\, \mathfrak{L}_{2N}(0,\infty)$ whenever $h \,\varepsilon\, \mathfrak{D}(\mathbf{Q}) \cap \mathfrak{L}_{2N}(0,\infty)$, and there exist real constants α and β ($\beta > \alpha$) with the property that

$$\| \mathbf{Q}h - \tfrac{1}{2}(\alpha + \beta)h \| \leqq \tfrac{1}{2}(\beta - \alpha) \ \| h \|$$

for all $h \,\varepsilon\, \mathfrak{D}(\mathbf{Q}) \cap \mathfrak{L}_{2N}(0,\infty)$.

(b) if $h \,\varepsilon\, \mathfrak{D}(\mathbf{Q})$, $\{h_y\,,\, 0 < y < \infty\} \subset \mathfrak{D}(\mathbf{Q})$.

4.5 *Conditions under Which $f(t) \to 0$ as $t \to \infty$*

Theorem 6: Suppose that the hypotheses of Theorem 4 are satisfied, that

$g(t) \rightarrow 0$ as $t \rightarrow \infty$, and that $k \; \varepsilon \; \mathcal{K}_{2N}$. Then $f(t) \rightarrow 0$ as $t \rightarrow \infty$.

Proof:

Observe first that the N-vector-valued function with values

$$\psi[f(t),t], \qquad 0 \leqq t < \infty,$$

is an element of $\mathcal{R} \cap \mathcal{L}_{2N}(0,\infty)$. Thus it suffices to show that if $h \; \varepsilon \; \mathcal{L}_{2N}(0,\infty)$,

$$\int_0^t k(t-\tau)h(\tau)\,d\tau \rightarrow 0 \quad \text{as} \quad t \rightarrow \infty.$$

In terms of $K(i\omega)$ and $\hat{h}(i\omega)$, respectively, the Fourier transforms of k and h,

$$\int_0^t k(t-\tau)h(\tau)\,d\tau = \frac{1}{2\pi}\int_{-\infty}^{\infty} K(i\omega)\hat{h}(i\omega)e^{i\omega t}\,d\omega.$$

Since $k \; \varepsilon \; \mathcal{K}_{2N}$, it follows that the modulus of each element of the N-vector $K(i\omega)\hat{h}(i\omega)$ is integrable on the ω-set $(-\infty,\infty)$. Thus, by the Riemann-Lebesgue lemma

$$\int_0^t k(t-\tau)h(\tau)\,d\tau \rightarrow 0 \quad \text{as} \quad t \rightarrow \infty.$$

This proves Theorem 6.

It is obvious that essentially the same argument suffices to prove the following corresponding result relating to Theorem 5.

Theorem 7: Suppose that the hypotheses of Theorem 5 are satisfied, that $g(t) \rightarrow 0$ as $t \rightarrow \infty$, and that $k \; \varepsilon \; \mathcal{K}_{2N}$. Then $\psi[f(t),t] \rightarrow 0$ as $t \rightarrow \infty$.

V. APPLICATIONS TO NONLINEAR DIFFERENTIAL EQUATIONS

Theorem 8: Let A be an $N \times N$ matrix of real constants, let $\psi[\cdot,\cdot]$ be as defined in Section I with α and β, respectively, replaced with $\hat{\alpha}$ and $\hat{\beta}$, and let f denote a real N-vector-valued function of t defined and differentiable on $[0,\infty)$ such that

$$\frac{df}{dt} + Af + \psi[f,t] = g$$

for almost all $t \; \varepsilon \; [0,\infty)$, where $g \; \varepsilon \; \mathcal{R} \cap \mathcal{L}_{2N}(0,\infty)$.

Suppose that

(*i*) *det* $[s1_N + \frac{1}{2}(\hat{\alpha} + \hat{\beta})1_N + A] \neq 0$ for $\sigma \geqq 0$

(*ii*) $\frac{1}{2}(\hat{\beta} - \hat{\alpha}) \sup\limits_{\omega} \Lambda\{[(i\omega)1_N + \frac{1}{2}(\hat{\alpha} + \hat{\beta})1_N + A]^{-1}\} < 1.$

Then $f \varepsilon \mathcal{L}_{2N}(0, \infty)$ and $f(t) \to 0$ as $t \to \infty$.
Proof:

Clearly, $f \varepsilon \mathcal{R} \cap \mathcal{E}_N$. Using the well-known expression for the solution of an inhomogeneous system of linear first-order differential equations in terms of the solution of the corresponding matrix homgeneous differential equation, and regarding

$$g + \tfrac{1}{2}(\hat{\alpha} + \hat{\beta})f - \psi[f,t]$$

as the "forcing function," we find that f satisfies

$$e^{-Bt}c + \int_0^t e^{-B(t-\tau)}g(\tau)d\tau$$

$$= f(t) + \int_0^t e^{-B(t-\tau)}\left\{\psi[f(\tau), \tau] - \frac{1}{2}(\hat{\alpha} + \hat{\beta})f(\tau)\right\} d\tau$$

for $t \varepsilon [0, \infty)$, in which $B = \frac{1}{2}(\hat{\alpha} + \hat{\beta})1_N + A$, and c is a real constant N-vector.

In view of (i), the matrix e^{-Bt} is an element of $\mathcal{K}_{1N} \cap \mathcal{K}_{2N}$. By the argument used in the proof of Theorem 6,

$$\int_0^t e^{-B(t-\tau)}g(\tau)\, d\tau \to 0 \quad \text{as} \quad t \to \infty,$$

a property which is obviously shared by $e^{-Bt}c$. Thus, using the fact that

$$-\frac{1}{2}(\hat{\beta} - \hat{\alpha}) \leqq \frac{\psi_n(w, t) - \frac{1}{2}(\hat{\alpha} + \hat{\beta})w}{w} \leqq \frac{1}{2}(\hat{\beta} - \hat{\alpha})$$

$$(n = 1, 2, \cdots, N)$$

for all $t \varepsilon [0, \infty)$ and all real $w \neq 0$, the theorem follows from a direct application of Theorems 4 and 6 [$-\frac{1}{2}(\hat{\beta} - \hat{\alpha})$ and $\frac{1}{2}(\hat{\beta} - \hat{\alpha})$, respectively, play the roles of α and β in Theorem 4].

5.1 *Generalization of an Earlier Theorem Concerning a Linear Differential Equation with Periodic Coefficients*[11]

Theorem 9: Let $\psi[\cdot,\cdot]$ be as defined in Section I with $N = 1$ and α and β, respectively, replaced with $\hat{\alpha}$ and $\hat{\beta}$. Let f denote a real-valued function of t defined and twice-differentiable on $[0, \infty)$ such that

$$\frac{d^2f}{dt^2} + a\frac{df}{dt} + \psi[f, t] = g$$

for almost all $t \varepsilon [0, \infty)$, where $g \varepsilon \mathcal{R} \cap \mathcal{L}_2(0, \infty)$ and a is a real constant. Then if $\hat{\alpha} > 0$ and $a > \sqrt{\hat{\beta}} - \sqrt{\hat{\alpha}}$, $f \varepsilon \mathcal{L}_2(0, \infty)$ and $f(t) \to 0$ as $t \to \infty$.

Proof:

Proceeding as in the proof of Theorem 8, we find that f satisfies

$$h(t) + \int_0^t k(t-\tau)g(\tau)\,d\tau$$

$$= f(t) + \int_0^t k(t-\tau)\{\psi[f(\tau),\tau] - \frac{1}{2}(\hat{\alpha} + \hat{\beta})f(\tau)\}\,d\tau$$

for $t \varepsilon [0, \infty)$, in which h is a solution of

$$\frac{d^2h}{dt^2} + a\frac{dh}{dt} + \frac{1}{2}(\hat{\alpha} + \hat{\beta})h = 0,$$

and $k \varepsilon \mathcal{K}_{11} \cap \mathcal{K}_{21}$ with

$$K(s) = \int_0^\infty k(t)e^{-st}\,dt = \left[s^2 + as + \frac{1}{2}(\hat{\alpha} + \hat{\beta})\right]^{-1}, \qquad \sigma \geqq 0.$$

With $\alpha = -\frac{1}{2}(\hat{\beta} - \hat{\alpha})$ and $\beta = \frac{1}{2}(\hat{\beta} - \hat{\alpha})$, condition ($i$) of Theorem 4 is obviously satisfied, while condition (ii) reduces to

$$\tfrac{1}{2}(\hat{\beta} - \hat{\alpha}) < \inf_{\omega} |\tfrac{1}{2}(\hat{\alpha} + \hat{\beta}) - \omega^2 + ia\omega|.$$

It is a simple matter to show that this inequality is satisfied if $\hat{\alpha} > 0$ and $a > \sqrt{\hat{\beta}} - \sqrt{\hat{\alpha}}$. Hence $f \varepsilon \mathcal{L}_2(0, \infty)$.

Since $h(t) \rightarrow 0$ as $t \rightarrow \infty$ and, by the argument used to prove Theorem 6,

$$\int_0^t k(t-\tau)g(\tau)\,d\tau \rightarrow 0 \quad \text{as} \quad t \rightarrow \infty,$$

Theorem 6 implies that $f(t) \rightarrow 0$ as $t \rightarrow \infty$. This completes the proof of Theorem 9.

VI. FINAL REMARK

Some of the results and techniques of this paper are useful in establishing sufficient conditions for the existence and uniqueness of solutions of functional equations of the type that we have considered. The reader familiar with the contraction-mapping fixed-point theorem has probably recognized this fact.

REFERENCES

1. Popov, N. M., Absolute Stability of Nonlinear Systems of Automatic Control, Avtomatika i Telemekhanika, **22**, Aug., 1961, pp. 961–978.

2. Kalman, R. E., Lyapunov Functions for the Problem of Lur'e in Automatic Control, Proc. Natl. Acad. Sci., **49,** Feb., 1963, pp. 201–205.
3. Bongiorno, Jr., J. J., An Extension of the Nyquist-Barkhausen Stability Criterion to Linear Lumped Parameter Systems, IEEE-PTGAC, **AC-8,** No. 2, April, 1963, pp. 166–170.
4. Youla, D. C., Some Results in the Theory of Active Networks, Polytechnic Institute of Brooklyn Research Report No. 1063-62, Aug., 1962.
5. Sandberg, I. W., On the Properties of Some Systems That Distort Signals—I, B.S.T.J., **42,** September, 1963, p. 2033.
6. Sandberg, I. W., Conditions for the Causality of Nonlinear Operators Defined on a Function Space, to be published.
7. Bochner, S., and Chandrasekharan, *Fourier Transforms,* Princeton University Press, Princeton, N. J., 1949, p. 99.
8. Sandberg, I. W., A Note on the Application of the Contraction-Mapping Fixed-Point Theorem to a Class of Nonlinear Functional Equations, to be published.
9. Paley, R. E., and Wiener, N., *Fourier Transforms in the Complex Domain,* published by the American Mathematical Society, Providence, Rhode Island, p. 8.
10. Titchmarsh, E. C., *Introduction to the Theory of Fourier Integrals,* Clarendon Press, Oxford, 2nd ed., 1948, pp. 125 and 128.
11. Sandberg, I. W., On the Stability of Solutions of Linear Differential Equations with Periodic Coefficients, to be published in the SIAM Journal.

ERRATA

On the Theory of Linear Multi-Loop Feedback Systems, I. W. Sandberg, B.S.T.J., **42,** March, 1963, pp. 355–382.

On page 361, the expression $(y_1 + y_2)$, which appears twice, should be replaced in both positions with $(y_1 + \bar{y}_2)$, in which $\bar{y}_2$ denotes the value of y_2 when $y_1 = 0$.

On page 377, the left side of the first equation of Section 9.4 should be det $\mathbf{F}_{\bar{\mathfrak{y}}_1}$, not det $\mathbf{F}\ {}_1$.

A Frequency-Domain Condition for the Stability of Feedback Systems Containing a Single Time-Varying Nonlinear Element

By I. W. SANDBERG

(Manuscript received May 6, 1964)

It is proved that a condition similar to the Nyquist criterion guarantees the stability (in an important sense) of a large class of feedback systems containing a single time-varying nonlinear element. In the case of principal interest, the condition is satisfied if the locus of a certain complex-valued function (a) is bounded away from a particular disk located in the complex plane, and (b) does not encircle the disk.

I. INTRODUCTION

The now well-known techniques introduced by Lyapunov have led to many very interesting results concerning the stability of time-varying nonlinear feedback systems governed by systems of differential equations. However, these methods have by no means led to a definitive theory of stability for even the simplest nontrivial time-varying nonlinear feedback systems. The general problem is, of course, one of considerable difficulty.

The unparalleled utility of the Nyquist stability criterion for single-loop, linear, time-invariant feedback systems is directly attributable to the fact that it is an explicit frequency-domain condition. The Nyquist locus not only indicates the stability or instability of a system, it presents the information in such a way as to aid the designer in arriving at a suitable design. The criterion is useful even in cases in which the system is so complicated that a sufficiently accurate analysis is not feasible, since experimental measurements can be used to construct the loop-gain locus.

The primary purpose of this article is to point out that some recently obtained mathematical results,[1] not involving the theory of Lyapunov, imply that a condition similar to, and possessing the advantages of, the Nyquist criterion guarantees the stability (in an important sense) of feedback systems containing a single time-varying nonlinear element.*,†

II. THE PHYSICAL SYSTEM AND DEFINITION OF $\mathcal{L}_2$-STABILITY

Consider the feedback system of Fig. 1. We shall restrict our discussion throughout to cases in which g_1, f, u, and v denote real-valued measurable functions of t defined for $t \geqq 0$.

The block labeled ψ is assumed to represent a memoryless time-varying (not necessarily linear) element that introduces the constraint $u(t) = \psi[f(t),t]$, in which $\psi(x,t)$ is a function of x and t with the properties that $\psi(0,t) = 0$ for $t \geqq 0$ and there exist a positive constant β and a real constant α such that

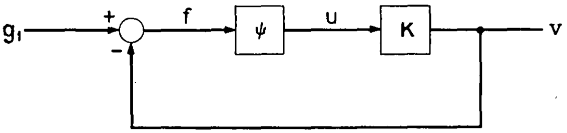


Fig. 1 — Nonlinear feedback system.

$$\alpha \leqq \frac{\psi(x,t)}{x} \leqq \beta, \qquad t \geqq 0$$

for all real $x \neq 0$. In particular, we permit the extreme cases in which $\psi(x,t)$ is either independent of t or linear in x [i.e., $\psi(x,t) = \psi(1,t)x$].

The block labeled **K** represents the linear time-invariant portion of the forward path. It is assumed to introduce the constraint

$$v(t) = \int_0^t k(t-\tau)u(\tau)d\tau - g_2(t), \qquad t \geqq 0$$

in which k and g_2 are real-valued functions such that

$$\int_0^\infty |\, k(t) \,|\, dt < \infty, \qquad \int_0^\infty |\, g_2(t) \,|^2 dt < \infty. \tag{1}$$

* The results of Ref. 1 relate to feedback systems containing an arbitrary finite number of time-varying nonlinear elements, but, with the exception of the case discussed here, they do not admit of a simple geometric interpretation.

† For results concerned with frequency-domain conditions for the global asymptotic stability (a sense of stability that is different from the one considered here) of nonlinear systems, see, for example, Refs. 2–4.

The function g_2 takes into account the initial conditions at $t = 0$. Our assumptions regarding **K** are satisfied, for example, if, as is often the case, u and v are related by a differential equation of the form

$$\sum_{n=0}^{N} a_n \frac{d^n v}{dt^n} = \sum_{n=0}^{N-1} b_n \frac{d^n u}{dt^n}, \qquad t \geqq 0$$

in which the a_n and the b_n are constants with $a_N \neq 0$, and

$$\sum_{n=0}^{N} a_n s^n \neq 0 \quad \text{for} \quad \operatorname{Re}[s] \geqq 0.$$

However, we *do not* require that u and v be related by a differential equation (or by a system of differential equations).

Assumption: We shall assume throughout that the response v is well defined and satisfies the inequality

$$\int_0^t \mid v(\tau) \mid^2 d\tau < \infty \tag{2}$$

for all *finite* $t > 0$, for each initial-condition function g_2 that meets the conditions stated above and each input g_1 such that

$$\int_0^\infty \mid g_1(t) \mid^2 dt < \infty.$$

Although this assumption plays an important role in the proof of the theorem to be presented, from an engineering viewpoint it is a trivial restriction (see Ref. 5).

Definition: We shall say that the feedback system of Fig. 1 is "$\mathfrak{L}_2$-stable" if and only if there exists a positive constant ρ with the property that the response v satisfies

$$\left(\int_0^\infty \mid v(t) \mid^2 dt\right)^{\frac{1}{2}} \leqq \rho \left(\int_0^\infty \mid g_1(t) + g_2(t) \mid^2 dt\right)^{\frac{1}{2}} + \left(\int_0^\infty \mid g_2(t) \mid^2 dt\right)^{\frac{1}{2}}$$

for every initial-condition function g_2 that meets the conditions stated above, and every input g_1 such that

$$\int_0^\infty \mid g_1(t) \mid^2 dt < \infty.$$

In particular, if the system is $\mathfrak{L}_2$-stable, then the response is square-integrable whenever the input is square-integrable.

It can be shown* that the response $v(t)$ approaches zero as $t \to \infty$ for any square-integrable input g_1, provided that the system is $\mathfrak{L}_2$-stable,

* See the proof of Theorem 6 of Ref. 1.

$g_2(t) \to 0$ as $t \to \infty$, and

$$\int_0^\infty |\, k(t) \,|^2 \, dt < \infty. \tag{3}$$

In addition, it follows at once from the Schwarz inequality that the response $v(t)$ is uniformly bounded on $[0, \infty)$ for any square-integrable input g_1, provided that the system is $\mathcal{L}_2$-stable, $g_2(t)$ is uniformly bounded on $[0, \infty)$, and (3) is satisfied.

III. SUFFICIENT CONDITIONS FOR $\mathcal{L}_2$-STABILITY

Theorem: Let

$$K(i\omega) = \int_0^\infty k(t) e^{-i\omega t} \, dt, \qquad -\infty < \omega < \infty.$$

The feedback system of Fig. 1 is $\mathcal{L}_2$-stable if one of the following three conditions is satisfied:

(i) $\alpha > 0$; *and the locus of $K(i\omega)$ for $-\infty < \omega < \infty$ (a) lies outside the circle C_1 of radius $\frac{1}{2}(\alpha^{-1} - \beta^{-1})$ centered on the real axis of the complex plane at $[-\frac{1}{2}(\alpha^{-1} + \beta^{-1}),0]$, and (b) does not encircle C_1 (see Fig. 2)*

(ii) $\alpha = 0$, *and $Re[K(i\omega)] > -\beta^{-1}$ for all real ω*

(iii) $\alpha < 0$, *and the locus of $K(i\omega)$ for $-\infty < \omega < \infty$ is contained within the circle C_2 of radius $\frac{1}{2}(\beta^{-1} - \alpha^{-1})$ centered on the real axis of the complex plane at $[-\frac{1}{2}(\alpha^{-1} + \beta^{-1}),0]$ (see Fig. 3).*

Proof: Note first that

$$\int_0^\infty |\, u(t) \,|^2 \, dt \leqq \max(\beta^2, |\, \alpha \,|^2) \int_0^\infty |\, f(t) \,|^2 \, dt,$$

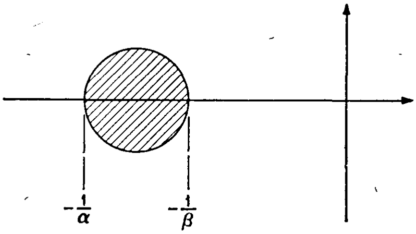


Fig. 2 — Location of the "critical circle" C_1 in the complex plane ($\alpha > 0$). The feedback system is $\mathcal{L}_2$-stable if the locus of $K(i\omega)$ for $-\infty < \omega < \infty$ lies outside C_1 and does not encircle C_1.

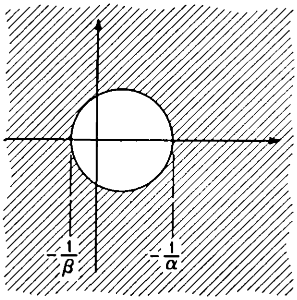


Fig. 3 — Location of the "critical circle" C_2 in the complex plane ($\alpha < 0$). The feedback system is $\mathcal{L}_2$-stable if the locus of $K(i\omega)$ for $-\infty < \omega < \infty$ is contained within C_2.

and hence, by a well-known result,

$$\int_0^\infty \left| \int_0^t k(t - \tau)u(\tau)d\tau \right|^2 dt \leqq \left(\int_0^\infty | k(t) | dt\right)^2 \int_0^\infty | u(t) |^2 dt$$

$$\leqq \max\left(\beta^2, | \alpha |^2\right)\left(\int_0^\infty | k(t) | dt\right)^2 \int_0^\infty | f(t) |^2 dt.$$

Using Minkowski's inequality,

$$\left(\int_0^\infty | v(t) |^2 dt\right)^{\frac{1}{2}} \leqq \left(\int_0^\infty \left| \int_0^t k(t - \tau)u(\tau)d\tau \right|^2 dt\right)^{\frac{1}{2}}$$

$$+ \left(\int_0^\infty | g_2(t) |^2 dt\right)^{\frac{1}{2}} \leqq \max(\beta, | \alpha |) \int_0^\infty | k(t) | dt$$

$$\cdot\left(\int_0^\infty | f(t) |^2 dt\right)^{\frac{1}{2}} + \left(\int_0^\infty | g_2(t) |^2 dt\right)^{\frac{1}{2}}.$$

Consider now the relation between $(g_1 + g_2)$ and f:

$$g_1(t) + g_2(t) = f(t) + \int_0^t k(t - \tau)\psi[f(\tau),\tau]d\tau, \qquad t \geqq 0$$

and suppose that

$$\int_0^\infty | g_1(t) + g_2(t) |^2 dt < \infty.$$

According to the results of Ref. 1, our assumptions* imply that there

* In Ref. 1 it is assumed that

$$\int_0^t | f(\tau) |^2 d\tau < \infty$$

exists a positive constant ρ_1 (which does not depend upon g_1 or g_2) such that

$$\int_0^\infty |f(t)|^2\,dt < \rho_1 \int_0^\infty |g_1(t) + g_2(t)|^2\,dt$$

provided that, with

$$K(s) = \int_0^\infty k(t)e^{-st}\,dt$$

and $\omega = \mathrm{Im}[s]$,

(i) $1 + \frac{1}{2}(\alpha + \beta)K(s) \neq 0$ for $\mathrm{Re}[s] \geqq 0$, and

(ii) $\frac{1}{2}(\beta - \alpha) \max\limits_{-\infty<\omega<\infty} |K(i\omega)[1 + \frac{1}{2}(\alpha + \beta)K(i\omega)]^{-1}| < 1.$

Thus the feedback system of Fig. 1 is $\mathfrak{L}_2$-stable if conditions (i) and (ii) are satisfied.

According to the well-known theorem of complex-function theory that leads to the Nyquist criterion, condition (i) is satisfied if (and only if) the polar plot of $K(i\omega)$ for $-\infty < \omega < \infty$ does not encircle or pass through the point $[-2(\alpha + \beta)^{-1},0]$. It can easily be verified that condition (ii) is met if one of the following three conditions is satisfied.

(a) $\alpha > 0$, and the locus of $K(i\omega)$ for $-\infty < \omega < \infty$ lies outside the circle C_1 of radius $\frac{1}{2}(\alpha^{-1} - \beta^{-1})$ centered in the complex plane at $[-\frac{1}{2}(\alpha^{-1} + \beta^{-1}),0]$.

(b) $\alpha = 0$, and $\mathrm{Re}[K(i\omega)] > -\beta^{-1}$ for all real ω.

(c) $\alpha < 0$, and the locus of $K(i\omega)$ for $-\infty < \omega < \infty$ is contained within the circle C_2 of radius $\frac{1}{2}(\beta^{-1} - \alpha^{-1})$ centered in the complex plane at $[-\frac{1}{2}(\alpha^{-1} + \beta^{-1}),0]$.

If $\alpha > 0$, the point $[-2(\alpha + \beta)^{-1},0]$ lies on the real-axis diameter of C_1, while if condition (b) or (c) is met, it is impossible for the polar plot of $K(i\omega)$ to encircle the point $[-2(\alpha + \beta)^{-1},0]$. Therefore, the conditions of the theorem guarantee that the feedback system is $\mathfrak{L}_2$-stable.

Remarks

With regard to the necessity of our sufficient conditions for $\mathfrak{L}_2$-stability, consider, for example, the case in which $\alpha > 0$ and suppose, for simplicity, that v and u are related by a differential equation of the type mentioned in Section II. Then, a moment's reflection shows that there exists a $\psi(x,t)$, in fact a $\psi(x,t)$ which is independent of t and linear in x,

for all finite $t > 0$. Our assumption that (2) is satisfied for all finite $t > 0$ implies that this condition is met.

that satisfies our assumptions and for which the feedback system is *not* $\mathfrak{L}_2$-stable, provided that for some value of ω, $K(i\omega)$ is a point on the real-axis diameter of C_1. This clearly shows that the condition is in the correct "ball park." Similar remarks can be made concerning our conditions for the cases in which $\alpha < 0$ and $\alpha = 0$.

IV. FURTHER PROPERTIES OF THE FEEDBACK SYSTEM OF FIG. 1

It is possible to say much more about the properties of the feedback system on the basis of frequency-domain information if our assumptions regarding $\psi(x,t)$ are strengthened.

For example, suppose that

$$\alpha \leqq \frac{\psi(x_1,t) - \psi(x_2,t)}{x_1 - x_2} \leqq \beta, \qquad \psi(0,t) = 0 \tag{4}$$

for $t \geqq 0$ and all real $x_1 \neq x_2$, and that one of the three conditions of our theorem is met. Let g_1 and $\hat{g}_1$ denote two arbitrary input functions such that

$$\int_0^t |g_1(\tau)|^2 \, d\tau < \infty \quad \text{and} \quad \int_0^t |\hat{g}_1(\tau)|^2 \, d\tau < \infty$$

for all finite $t > 0$, and

$$\int_0^\infty |g_1(\tau) - \hat{g}_1(\tau)|^2 \, d\tau < \infty.$$

Let v and $\hat{v}$, respectively, denote the (assumed well defined) responses due to g_1 and $\hat{g}_1$. Then if

$$\int_0^t |v(\tau)|^2 \, d\tau < \infty \quad \text{and} \quad \int_0^t |\hat{v}(\tau)|^2 \, d\tau < \infty$$

for all finite $t > 0$, and the assumptions of Section II are met, it follows* that

$$\int_0^\infty |v(\tau) - \hat{v}(\tau)|^2 \, d\tau < \infty$$

and that there exists a positive constant λ (which does not depend upon g_1 or $\hat{g}_1$) such that

$$\int_0^\infty |v(\tau) - \hat{v}(\tau)|^2 \, d\tau \leqq \lambda \int_0^\infty |g_1(\tau) - \hat{g}_1(\tau)|^2 \, d\tau.$$

* Consider Theorem 1 of Ref. 6 with $h_1(t) = f_1(t) = 0$ for $t < 0$.

Suppose now that $\psi(x,t)$ satisfies (4) and is either independent of t or periodic in t with period T for each x, and that one of the three conditions of our theorem is met. Assume that the initial-condition function $g_2(t)$ approaches zero as $t \to \infty$, and that the input $g_1(t)$ applied at $t = 0$ is a bounded periodic function with period T. Then it can be shown* that there exists a bounded periodic function p, with period T, which is independent of g_2 and such that the (assumed well defined) response $v(t)$ approaches $p(t)$ as $t \to \infty$, provided that the conditions of Section II are met, (2) is satisfied for all finite $t > 0$, and

$$\int_0^\infty \left| \int_t^\infty | k(\tau) | \, d\tau \right|^2 dt < \infty, \qquad \int_0^\infty | (1 + t)k(t) |^2 \, dt < \infty. \tag{5}$$

Observe that the conditions of (5) are satisfied if u and v are related by a differential equation of the form described in Section II.

REFERENCES

1. Sandberg, I. W., On the $\mathfrak{L}_2$-Boundedness of Solutions of Nonlinear Functional Equations, B.S.T.J., this issue, p. 1581.
2. Popov, V. M., Absolute Stability of Nonlinear Systems of Automatic Control, Avtomatika i Telemekhanika, **22**, Aug., 1961, pp. 961–978.
3. Kalman, R. E., Lyapunov Functions For the Problem of Lur'e in Automatic Control, Proc. Natl. Acad. Sci., **49**, Feb., 1963, pp. 201–205.
4. Rekasius, Z. V., A Stability Criterion for Feedback Systems with One Nonlinear Element, Trans. IEEE-PTGAC, **AC9**, Jan., 1964, pp. 46–50.
5. Tricomi, F. G., *Integral Equations*, Interscience Publishing, Inc., New York, 1957, p. 46.
6. Sandberg, I. W., and Beneš, V. E., On the Properties of Nonlinear Integral Equations That Arise in the Theory of Dynamical Systems, to be published.

* See Theorem 3 of Ref. 6.

The Resistance of an Infinite Slab with a Disk Electrode

By G. F. FOXHALL and J. A. LEWIS

(Manuscript received April 17, 1964)

We consider the resistance of an infinite slab measured between an electrode covering one face and a circular electrode attached to the other face by a uniform resistive film representing a contact resistance. Upper and lower bounds are found on the difference between the total resistance and the resistance of the film alone. The bounds are obtained by a combination of analysis and experiment, using an electrolytic tank. The results may be applied to determine contact resistance from a measured value of total resistance and a knowledge of the bulk resistivity of the slab material.

I. INTRODUCTION

We consider the resistance of an infinite conducting slab as measured between an electrode entirely covering one face and a circular electrode affixed to the other face by a resistive film. This resistance can be imagined to be made up of two resistances in series: namely, the film contact resistance and a resistance which is due to the body to which the electrode is attached, but which depends on the film resistance.

Lewis[1] has derived general upper and lower bounds on this body resistance. In the present case the upper bound may be calculated analytically. The lower bound is the resistance which would exist between the electrodes in the absence of the film. Calculation of the last-mentioned resistance involves a classical potential problem treated by Weber in 1873, but still not completely solved today.

We treat this problem by a combination of analysis and experiment, the latter in effect being an analog computation using an electrolytic tank. An asymptotic solution is found which converges rapidly for slab thicknesses as small as one disk radius, while for smaller thicknesses experimentally determined values of resistance are used.

The upper and lower bounds for the body resistance, which differ only by 8 per cent for a thick slab and tend to the same value for a thin slab, provide a convenient estimate of slab resistance in the presence of a contact film. More important from the practical point of view, they provide a useful estimate of contact resistance for measured total resistance.

II. THE PROBLEM

We wish to determine the electrical resistance of the slab electrode configuration shown in Fig. 1. The entire base of the slab is in perfect contact with a highly conducting, grounded electrode, while the upper electrode, a circular disk of radius a, is separated from the slab by a thin film of surface conductance c. The potential $V(R,Z)$ in the slab then satisfies Laplace's equation

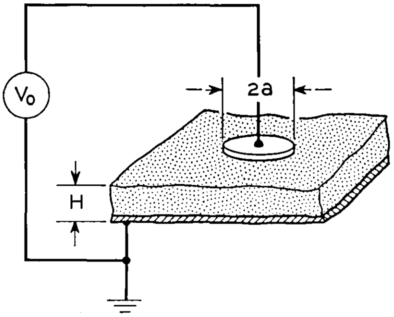


Fig. 1 — Infinite slab with disk electrode.

$$\partial^2 V/\partial R^2 + \partial V/R\partial R + \partial^2 V/\partial Z^2 = 0, \tag{1}$$

for all R and $0 < Z < H$, and the boundary conditions

$$V(R,0) = 0, \tag{2}$$

$$\sigma \partial V(R,H)/\partial Z = \begin{cases} c[V_0 - V(R,H)], & \text{for } R < a \\ 0, & \text{for } R > a, \end{cases} \tag{3}$$

where σ is the conductivity of the slab.

The total input current I is given by

$$I = \int_0^a \sigma \frac{\partial V}{\partial Z} (R,H) 2\pi R \, dR, \tag{4}$$

so that the resistance measured between the electrodes is

$$R_m = \frac{V_0}{I} = V_0 \Big/ \int_0^a \sigma \frac{\partial V}{\partial Z} (R,H) 2\pi R \, dR. \tag{5}$$

If we set

$$r = R/a, \qquad z = Z/a, \qquad h = H/a, \qquad w(r,z) = V(R,Z)/V_0,$$

this becomes

$$\sigma a R_m = 1 \bigg/ \int_0^1 \frac{\partial w}{\partial z}\,(r,h) 2\pi r\,dr, \tag{6}$$

where

$$\nabla^2 w = \partial^2 w/\partial r^2 + \partial w/r\partial r + \partial^2 w/\partial z^2 = 0, \tag{7}$$

in $0 < z < h$,

$$w(r,0) = 0, \tag{8}$$

for all r, and

$$\partial w(r,h)/\partial z = \begin{cases} (ca/\sigma)[1 - w(r,h)], & \text{for } r < 1 \\ 0, & \text{for } r > 1. \end{cases} \tag{9}$$

It is easy to show that R_m may be written in the form

$$R_m = R_c + R(w), \tag{10}$$

where R_c is the film resistance $(1/\pi a^2 c$, in the present case) and $R(w)$ is the ratio of average potential difference between electrodes to total current, i.e.,

$$\sigma a R(w) = 2 \int_0^1 w(r,h) r\,dr \bigg/ \int_0^1 \frac{\partial w}{\partial z}\,(r,h) 2\pi r\,dr. \tag{11}$$

Thus, if $R(w)$ has been calculated and R_m measured, the film resistance R_c may be determined.

However, the calculation of $R(w)$ in general involves the solution of a difficult mixed boundary value problem. Furthermore, since w depends on the film conductance c, which is essentially the quantity to be determined by combined calculation and measurement, $R(w)$ must be calculated for a large number of values of c to make certain that the experimental range is covered. These difficulties can be circumvented, with only a moderate loss in accuracy in the present case, by the use of certain upper and lower bounds on $R(w)$.

III. UPPER AND LOWER BOUNDS

The bounds on $R(w)$ have the form[1]

$$R(u) \leqq R(w) \leqq R(v), \tag{12}$$

where

$$\sigma a R(u) = 1 \Big/ \int_0^1 \frac{\partial u}{\partial z} (r,h) 2\pi r \, dr, \tag{13}$$

$$\sigma a R(v) = (2/\pi) \int_0^1 v(r,h) r \, dr, \tag{14}$$

u and v satisfy (7) and (8), and

$$u(r,h) = \partial v(r,h)/\partial z = 1, \qquad \text{for } r < 1 \tag{15}$$

$$\partial u(r,h)/\partial z = \partial v(r,h)/\partial z = 0, \qquad \text{for } r > 1. \tag{16}$$

The lower bound $R(u)$ is the resistance as usually defined, i.e., the reciprocal of the total current for a unit potential difference applied uniformly between the electrodes. On the other hand, $R(v)$ is the *average* potential difference required to give unit total current, distributed uniformly over the input electrode. The former case may be realized by letting the film conductance c become very large; the latter by letting c tend to zero and V_0 tend to infinity in such a way that cV_0 tends to a finite value.

The calculation of $R(v)$ is straightforward, involving only the solution of an unmixed boundary value problem ($\partial v/\partial z$ specified all over $z = h$), but $R(u)$ involves the solution of a *mixed* boundary value problem and ultimately the solution of dual integral equations. An asymptotic solution of these integral equations, valid for large thickness ($h \gg 1$), has been obtained by Tranter.[2] The corresponding expression for $R(u)$ is rapidly convergent, so that it appears to be usable down to values of h of order unity. The range $h < 1$ is covered by measurements on an electrolytic tank analog.

IV. THE UPPER BOUND

If we set

$$v(r,z) = \int_0^\infty f(p) \frac{\sinh pz}{\cosh ph} J_0(pr) dp, \tag{17}$$

the conditions

$$\nabla^2 v = v(r,0) = 0$$

are satisfied and the remaining conditions become

$$\partial v(rh)/\partial z = \int_0^\infty p f(p) J_0(pr) dp = \begin{cases} 1, & r < 1 \\ 0, & r > 1. \end{cases} \tag{18}$$

Now

$$\int_0^\infty J_0(pr)J_1(p)dp = \begin{cases} 1, & r < 1 \\ 0, & r > 1, \end{cases} \tag{19}$$

so that, if we set

$$f(p) = J_1(p)/p,$$

we obtain the complete solution. In particular,

$$v(r,h) = \int_0^\infty \frac{\tanh ph}{p} J_1(p)J_0(pr)dp \tag{20}$$

and

$$\sigma a R(v) = (2/\pi) \int_0^\infty [J_1(p)/p]^2 \tanh ph\, dp. \tag{21}$$

For small h, $\sigma aR(v) \approx h/\pi$ or $R(v) \approx H/\pi a^2\sigma$, the resistance of a circular cylinder, while for large h, $\sigma aR(v) \approx 8/3\pi^2$, a result derived in Carslaw and Jaeger.[3]

Miss M. C. Gray has obtained an expansion in powers of h^{-1} whose first two terms, i.e.,

$$\sigma aR(v) \approx \frac{8}{3\pi^2} - \frac{\log 2}{2\pi h} \tag{22}$$

give reasonable accuracy down to $h = 1$. She has also evaluated the integral for $R(v)$ numerically for $0.1 < h < 10$. This is the curve labeled $R(v)$ in Fig. 2.

V. THE ASYMPTOTIC VALUE OF $R(u)$

If we assume $u(r,z)$ to have the same form as $v(r,z)$ in the previous section, the function $f(p)$ must now satisfy the dual integral equations

$$u(r,h) = \int_0^\infty f(p) \tanh phJ_0(pr)dp = 1, \qquad r < 1, \tag{23}$$

$$\partial u(r,h)/\partial z = \int_0^\infty pf(p)J_0(pr)dp = 0, \qquad r > 1. \tag{24}$$

We can no longer solve these equations by inspection, but for large h an approximate solution, due to Tranter,[2] is available. Tranter gives

$$f(p) \approx \left(\frac{2}{\pi}\right) A(h) \frac{\sin p}{p}, \tag{25}$$

where

$$A(h) = 1 + \frac{2 \log 2}{\pi h} + \left(\frac{2 \log 2}{\pi h}\right)^2 + O(h^{-3}). \tag{26}$$

For $A = 1$ $(h = \infty)$ we obtain the classical result

$$[\sigma a R(u)]_\infty = \tfrac{1}{4},$$

and for large h

$$\sigma a R(u) \approx 1/4A(h). \tag{27}$$

This function is shown in Fig. 2 for $h \geqq 1$.

VI. THE EXPERIMENT

In order to verify and supplement the computed values of resistance, an experiment using an electrolytic solution as the conducting slab was devised. The apparatus, shown in Fig. 3, consisted of a 10 × 14 inch plastic tank into which a gold-plated brass plate was fitted. A 0.01 normal KCl solution was used to simulate the conducting slab, while the end of a 0.564-inch diameter gold-plated brass rod served as the disk electrode.

The experiment consisted of two parts. First resistance measurements were made for the slab configuration at various solution levels. Then a glass sleeve, closely fitted to the upper electrode, was used to constrain the current in the solution to a simple cylindrical geometry. This provided a measurement of solution conductivity and also of contact resist-

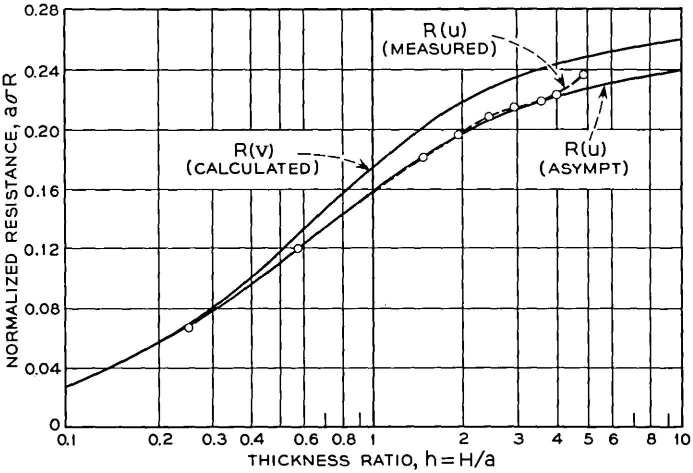


Fig. 2 — The resistance of an infinite slab.

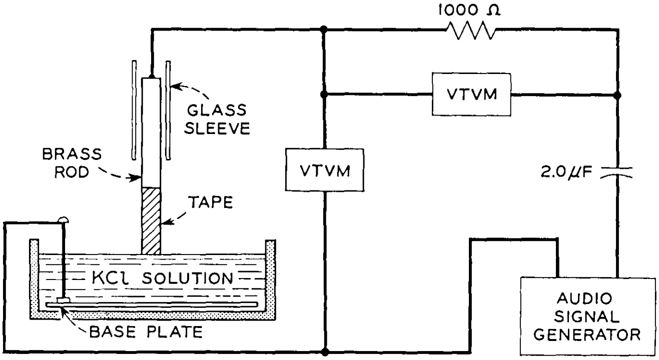


Fig. 3 — Electrolytic tank analog.

ance. The resulting measured values are shown in Fig. 4 as a function of solution depth. The slope gives a conductivity of 1.33 × 10^{-3} (ohm-cm)$^{-1}$, in close agreement with the tabulated value[4] for the solution temperature of 22°C. Extrapolation to zero solution depth indicates a negligible contact resistance.

As seen in Fig. 2, the measured values of resistance for the slab fall

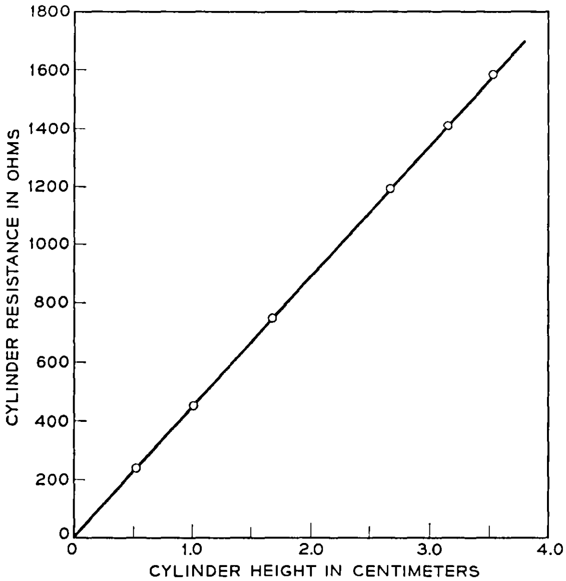


Fig. 4 — The resistance of a cylinder of electrolyte.

very close to the computed values in the range $2 \leqq h \leqq 4$, where the asymptotic form for $R(u)$ can be expected to be accurate. At smaller values of h, the accuracy of the asymptotic form decreases and the measured values may be taken as a good approximation to $R(u)$. At large values of h, on the other hand, the asymptotic form becomes very accurate, while the experimental model becomes less so. This may be seen in Fig. 2 for $h \approx 5$.

This divergence, which at first was attributed to the finite diameter of the tank, is now believed to be due to polarization effects produced by the nonuniform field near the upper electrode. The abrupt upturn of resistance near $h = 5$ cannot be due to finite tank diameter, which would yield a resistance-depth curve with a slope decreasing from a value for a cylinder having the same diameter as the upper electrode to one for a long cylinder having the same diameter as the tank. On the other hand, the current density distribution over the upper electrode is nonuniform for any depth, being infinite at the electrode edge in the mathematical idealization. This nonuniformity increases with depth, for more current is drawn from the electrode center at small depth than at large depth. Thus for fixed total current (the experimental condition) the current density at the edge increases with increasing depth until a depth is reached ($h \approx 5$ in the experiments) at which local polarization effects become appreciable. This dependence on depth also accounts for the different behavior of measured values for the cylinder (Fig. 4) and the slab (Fig. 2).

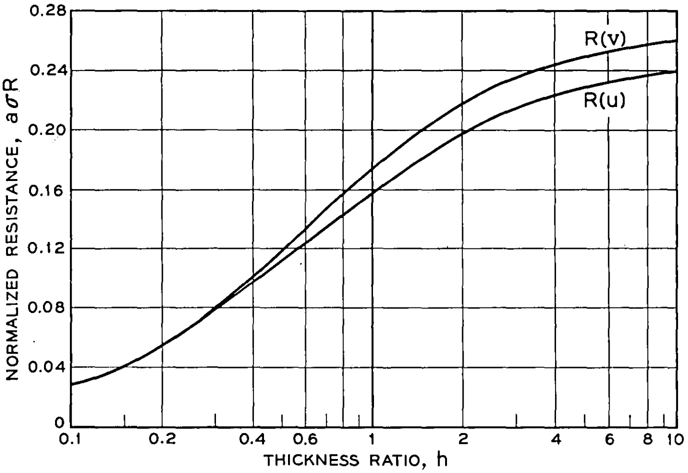


Fig. 5 — Upper and lower bounds on the resistance of an infinite slab.

VII. SUMMARY OF RESULTS

Fig. 5 summarizes the results of the analysis and the experiment. The lower curve, labeled $R(u)$, gives the resistance between the electrodes for zero contact resistance. It also gives a lower bound on the difference between the total resistance R_m with a resistive film between the upper electrode and the slab and the resistance $R_c = 1/\pi a^2 c$ of the film itself, while the upper curve, labeled $R(v)$, gives an upper bound on the same quantity. Thus, for all slab thicknesses and (constant) film conductances c,

$$R(u) \leqq R_m - R_c \leqq R(v), \tag{28}$$

or, assuming R_m to be determined by measurement,

$$R_m - R(v) \leqq R_c \leqq R_m - R(u) \tag{29}$$

giving an estimate of the contact resistance itself.

To supplement Fig. 3, the asymptotic forms

$$\sigma a R(u) \sim (0.250)/[1 + (0.441/h) + (0.441/h)^2], \tag{30}$$

$$\sigma a R(v) \sim (0.270) + (0.110/h), \tag{31}$$

may be used for thickness ratio $h \geqq 1$, while, for $h \leqq 0.1$, both $\sigma a R(u)$ and $\sigma a R(v)$ are closely approximated by the cylinder resistance h/π.

VIII. ACKNOWLEDGMENTS

The authors wish to acknowledge the help of Miss M. C. Gray with the evaluation of the integral for $R(v)$ and the advice of D. L. Klein on the experiment.

REFERENCES

1. Lewis, J. A., Generalized Thermal Resistance, Quart. Appl. Math., **19**, April, 1961, pp. 76–80.
2. Tranter, C. J., *Integral Transforms in Mathematical Physics*, 2nd Ed., Methuen, New York, 1956, pp. 117–120.
3. Carslaw, H. S., and Jaeger, J. C., *Conduction of Heat in Solids*, 2nd. Ed., Oxford, 1959, pp. 214–217.
4. *Handbook of Chemistry and Physics*, 32nd Ed., Chemical Rubber Publishing Company, Cleveland, 1950, p. 2147.

Permutation Groups, Complexes, and Rearrangeable Connecting Networks

By V. E. BENEŠ

(Manuscript received March 12, 1964)

In the interest of providing good telephone service with efficient connecting networks, it is desirable to have at hand a knowledge of some of the combinatorial properties of such networks. One of these properties is rearrangeability: a connecting network is rearrangeable if its permitted states realize every assignment of inlets to outlets, or alternatively, if given any state x of the network, any inlet idle in x, and any outlet idle in x, there is a way of assigning new routes (if necessary) to the calls in progress in x so that the idle inlet can be connected to the idle outlet.

A natural algebraic and combinatorial approach to the study of rearrangeable networks is described, with attention centered principally on two-sided networks built of stages of square crossbar switches, each stage having N inlets and N outlets. The approach is based in part on the elementary theory of permutation groups. The principal problem posed (and partly answered) is this: What connecting networks built of stages are rearrangeable? Sufficient conditions, including all previously known results, are formulated and exemplified.

I. INTRODUCTION

A connecting network is an arrangement of switches and transmission links through which certain terminals can be connected together in many combinations. Typical examples of connecting networks can be found in telephone central offices, where they are used to complete calls among the customers themselves, and between customers and outgoing trunks leading to other offices.

In the interest of providing good service with efficient connecting networks, it is desirable to have a thorough understanding of some of the combinatorial properties of such networks. In a previous paper,[1] we singled out three such combinatory properties as useful in assessing the performance of connecting networks. The weakest of these properties was that of *rearrangeability*. A connecting network is rearrangeable if its permitted states realize every assignment of inlets to outlets, or alternatively, if given any state x of the network, any inlet idle in x, and any outlet idle in x, there is a way of assigning new routes (if necessary) to the calls in progress in x so as to lead to a new state of the network in which the idle inlet can be connected to the idle outlet.

Figs. 1 and 2 show the structure of two connecting networks built out of square crossbar switches, with each switch capable of connecting any subset of its inlets to an equinumerous subset of its outlets in any desired one-one combination. The network of Fig. 1 is often found in telephone central offices; we may call it the No. 5 crossbar network. It is *not* rearrangeable. The network of Fig. 2 *is* rearrangeable, but so far it has not found extensive practical use.

We shall describe a natural algebraic and combinatorial approach to the study of rearrangeability. For the most part we restrict attention to two-sided connecting networks that are built of stages of crossbar switches, and have the same number N of inlets as outlets. The approach is based in part on the elementary theory of permutation groups. The way the connection with group theory arises can be summarized as follows: a maximal state of the network is one in which no additional

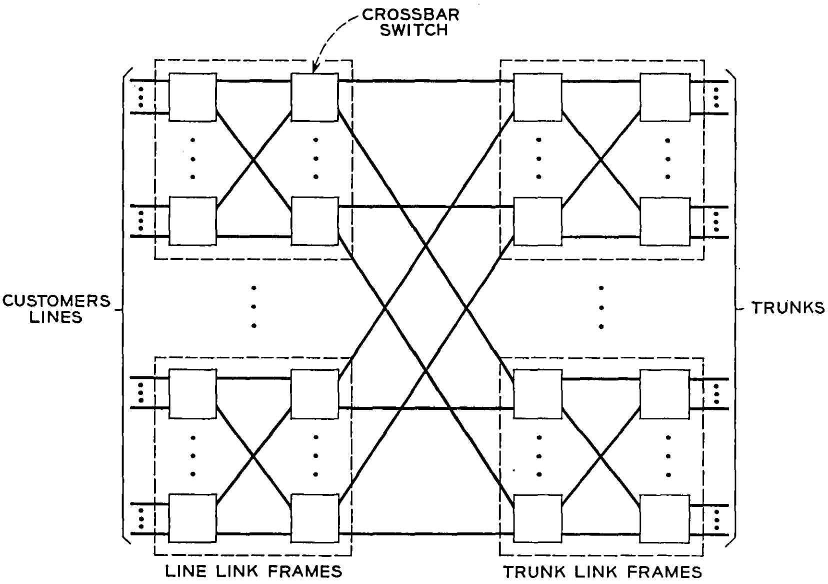


Fig. 1 — Structure of No. 5 crossbar network.

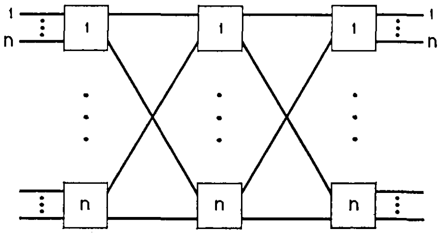


Fig. 2 — Rearrangeable network.

calls can be completed in the network; suppose that both the inlets and the outlets are numbered in an arbitrary way from 1 to N; each maximal state realizes some submap of a permutation on $\{1, \cdots, N\}$; the network is rearrangeable if and only if the whole group of all permutations of $\{1, \cdots, N\}$ is generated in this way by the maximal states of the network. Details are worked out in the main body of the paper.

It is not possible to explore in one paper all the possible uses of group theory in the study of connecting networks. Indeed, we shall restrict ourselves to formulating the fundamental problem of rearrangeable networks in terms of complexes of permutations, and to giving a partial answer. One of the difficulties with the approach is that it always seems to be easier to obtain results about groups by the few available methods known for rearrangeable networks, than *vice versa.*

A sequel[2] to the present paper is concerned with the problem of synthesizing a rearrangeable network (for N inlets and outlets), subject to certain structural conditions and to the condition that it have a minimum number of crosspoints.

II. SUMMARY

In Section III we define a precise general notion of a "stage" of switching in a connecting network, and, after describing how the networks which will be of interest are built out of stages by joining them together by patterns of links, we pose two problems: first, to discover what networks built in this way are rearrangeable; and second, to synthesize optimal rearrangeable networks of given size, optimal in the sense of having fewest crosspoints (among those in some class of networks having practical interest). (See Ref. 2.)

Section IV is devoted to giving a formulation of the first problem (discovering rearrangeable networks) in terms of partitions and permutation groups, using the notion of stage. In Section V we discuss how stages generate complexes (in the group theory sense, i.e., sets of group elements). It is shown that a stage can generate a subgroup only if it contains a substage made of square switches, a result that indicates to some extent the "best possible" nature of stages made of square switches.

A known example, discussed in Section VI, indicates how a particular symmetric group S is generated by a rearrangeable network in the form

$$S = \left\{ \prod_{i=1}^{3} \varphi_i \right\}$$

with φ_1, φ_3 in a subgroup H and φ_2 in a certain subgroup $\varphi^{-1}H\varphi$ conjugate to H.

The remainder of the paper is devoted to proving two "rearrangeability" theorems for connecting networks built of stages of square switches. The first theorem gives sufficient conditions under which a set of stages of square switches connected by link patterns will give rise to a rearrangeable network. The second theorem indicates a simple way of describing link patterns and stages that satisfy the hypotheses of the first theorem, and so yield many specific rearrangeable networks, generalizations of those given by Paull.[3]

III. STAGES AND LINK PATTERNS

The switches in Figs. 1 and 2 are arranged in columns which we shall call *stages*, the switches in these stages being identical. Two adjacent stages are connected by a pattern of *links* or *junctors*. Along with the switches, the link patterns are responsible for the distributive characteristics of the network. They afford an inlet ways of reaching many outlets. Obviously, each outlet on a switch in a given stage is some inlet of the next stage, if there is one. Suppose that the N inlets are numbered in an arbitrary way, and that the N outlets are also numbered in an arbitrary way, both from 1 to N. Then it is clear that each link pattern, and each permitted way of closing the largest possible number of crosspoints in a stage, viz. N, can be viewed abstractly as a permutation on $\{1, \cdots, N\}$. Both the networks in Figs. 1 and 2 have the property that all maximal* states have the same number N of calls in progress, and any such maximal state realizes a permutation which is a *product* of certain of the permutations represented by the link patterns and the stages.

It will be convenient to generalize the usual notion of a "stage" of switching in a connecting network. By a stage (of switching) we shall

* I.e., states in which no additional calls can be completed.

mean a connecting network constructed as follows: with I the set of inlets, and Ω that of outlets, we choose an arbitrary subset $\mathcal{S}$ of $I \times \Omega$, and we place a crosspoint between all and only those inlets $u \in I$ and outlets $v \in \Omega$ such that $(u,v) \in \mathcal{S}$. We shall also speak of $\mathcal{S}$ itself as the "stage." Thus we make

Definition 1: A stage is a subset of $I \times \Omega$.

This terminology is easily seen to be an extension of the usual one, according to which, e.g., a column of switches in Fig. 1 forms a stage, the network having four stages separated (or joined) by three link patterns. Actually, a link pattern may be associated with one or the other (but usually not both) of the stages it connects, to define a new stage; we do not usually do this.

Definition 2: A stage $\mathcal{S}$ is made of square switches if and only if there is a partition Π of $\{1, \cdots, N\}$ such that

$$\mathcal{S} = \bigcup_{A \in \Pi} (A \times A).$$

Definition 3: A substage $\mathcal{S}'$ of $\mathcal{S}$ is a subset of $\mathcal{S}$.

Except in the trivial case in which $\mathcal{S}$ is actually a square N-by-N switch (i.e., $\mathcal{S} = I \times \Omega$), a stage $\mathcal{S}$ will not by itself give rise to a rearrangeable network. Still, it is known that several stages joined end to end by suitable link patterns can together give rise to such a network, e.g., that of Fig. 2. We can thus formulate two fundamental questions about connecting networks built out of stages:

(1) What stages and link patterns can be used to construct a rearrangeable network?

(2) What stages, and how many of them, should be used to construct a rearrangeable network that has a minimum number of crosspoints (switches) for a given number of terminals on a side?

Question (1) is studied in the present work, while question (2) is treated in another paper.[2]

IV. GROUP THEORY FORMULATION

We shall adopt some notational conventions from group theory to simplify our presentation. Let G be a group. It is customary to speak of a subset $K \subseteq G$ as a *complex*. If $x \in G$, then xK denotes the set of products xy with $y \in K$, Kx denotes the set of products yx with $y \in K$. Similarly, if K_1 and K_2 are complexes, K_1K_2 denotes the set of products yz with $y \in K_1$ and $z \in K_2$.

A group G of permutations is called *imprimitive*[4] if the objects acted on by the permutations of G can be partitioned into mutually disjoint sets, called the sets of *imprimitivity*, such that every $\varphi \in G$ either permutes the elements of a set among themselves, or carries that set onto another. That is, there is a nontrivial partition Π of the set X of objects acted on such that $\varphi \in G$ and $A \in \Pi$ imply $\varphi(A) \in \Pi$. We shall extend this terminology as follows:

Definition 4: G is called *strictly imprimitive* if it is imprimitive, and each set A of imprimitivity is carried into itself by elements of G, i.e., there is a nontrivial partition Π of X such that $A \in \Pi$ implies $\varphi(A) = A$ for all $\varphi \in G$, so that $\varphi \in G$ is "nonmixing" on Π.

Consider a stage of switching that has N inlets and N outlets. It is evident that such a stage provides ways of connecting some of the inlets to some of the outlets. If the stage contains enough crosspoints it can be used to connect every inlet to some outlet in a one-to-one fashion, i.e., with no inlet connected to more than one outlet and vice versa. With the inlets and outlets both numbered $1, 2, \cdots, N$, such a setting of the switches corresponds to a permutation on $\{1, \cdots, N\}$. Indeed, there may be many ways of doing this, differing in what inlets are connected to what outlets, that is, corresponding to different permutations.

Definition 5: A stage $\mathcal{S}$ generates the permutation φ if there is a setting of N switches of $\mathcal{S}$ which connects each inlet to one and only one outlet in such a way that i is connected to $\varphi(i) i = 1, \cdots, N$, that is, if

$$(i,\varphi(i)) \in \mathcal{S}.$$

Definition 6: The set of permutations generated by a stage $\mathcal{S}$ is denoted by $P(\mathcal{S})$.

Definition 7: A network (with N inlets and N outlets) generates a permutation φ if there is a setting of the switches in the network which connects, by mutually disjoint paths, each inlet to one and only one outlet in such a way that i is connected to $\varphi(i)$, $i = 1, \cdots, N$.

If two stages $\mathcal{S}_1$, $\mathcal{S}_2$ are connected by a link pattern corresponding to a permutation φ_2, then the permutations that they generate together are those of the form

$$\varphi_1\varphi_2\varphi_3, \qquad \varphi_i \in P(\mathcal{S}_i), \qquad i = 1 \text{ or } 3.$$

If a network consists of two stages $\mathcal{S}_1$, $\mathcal{S}_2$ joined by a link pattern corresponding to a permutation φ, then it can be seen that it generates exactly the permutations in the set

$$P(\mathcal{S}_1)\varphi P(\mathcal{S}_2).$$

A network of s stages $\mathcal{S}_i$, $i = 1, \cdots, s$, with a link pattern corresponding to φ_i, $i = 1, \cdots, s - 1$, between the ith and the $(i + 1)$th stages, generates the complex

$$P(\mathcal{S}_1)\varphi_1 P(\mathcal{S}_2) \cdots \varphi_{s-1}P(\mathcal{S}_s).$$

We shall occasionally use the suggestive notation

$$\mathcal{S}_1\varphi_1\mathcal{S}_2 \cdots \varphi_{s-1}\mathcal{S}_s$$

to refer to or indicate such a network.

It is now possible to formulate a group-theoretic approach to the analysis and synthesis of rearrangeable connecting networks made of stages of switching joined by link patterns. Consider such a network, generating the complex

$$P(\mathcal{S}_1)\varphi_1 \cdots \varphi_{s-1}P(\mathcal{S}_s).$$

The factors $\varphi_i P(\mathcal{S}_{i+1})$, $i = 1, \cdots$, s $- 1$, occurring herein are themselves again just complexes. Thus, given any product of complexes

$$\prod_{i=1}^{s} K_i$$

we seek to know whether the product is the whole symmetric group, and whether the factor complexes K_i can be written in the form

$$\varphi P(\mathcal{S})$$

where φ is a permutation and $\mathcal{S}$ is a stage. In this general form the problem is largely unsolved; however, special cases are worked out in the sequel.

V. THE GENERATION OF COMPLEXES BY STAGES

We start with this elementary result:

Remark 1: Let M be a complex (i.e., a set) of permutations. Define a stage $\mathcal{S}$ by

$$\mathcal{S} = \{(x,y) : \varphi(x) = y \text{ for some } \varphi \in M\}.$$

Then $P(\mathcal{S}) \supseteq M$ and no smaller stage has this property.

In cases of practical importance, such as shown in Figs. 1 and 2, the stages are made of square switches, and it is clear that a stage $\mathcal{S}$ (with N inlets and N outlets) is capable of effecting certain special permutations on $X = \{1, \cdots, N\}$, and of course, all submaps thereof. (Indeed, for each switch there are numbers m and n with $m < n$ such that the switch is capable of performing all the $(n - m + 1)!$ permutations of the numbers k in the range $m \leqq k \leqq n$ among themselves.) Since no inlet [outlet] is on more than one switch, these permutations form a *subgroup* of the symmetric group $S(X)$ of *all* permutations on $\{1, \cdots, N\}$. This subgroup has a property which might be described intuitively by saying that there exist sets on which the subgroup elements can mix "strongly," but which they keep separate. It is apparent, indeed, that the subgroup generated by a stage made of square switches is strictly imprimitive, the sets of imprimitivity being just the elements of the partition Π of $\{1, \cdots, N\}$ according to what switch an inlet [outlet] is on. This situation might also be described by saying that a permutation φ from the subgroup is *nonmixing* on Π.

Our second observation is

Remark 2: Let H be a strictly imprimitive group of permutations on $X = \{1, \cdots, N\}$, with sets of imprimitivity forming the partition Π. Let $\mathcal{S}$ be the smallest stage with $P(\mathcal{S}) \supseteq H$. Then

$$\mathcal{S} = \bigcup_{A \in \Pi} A \times A,$$

i.e., $\mathcal{S}$ is made of square switches.

The main result of this section states that a stage can generate a subgroup only if it contains a substage made of square switches. This suggests that stages made of square switches necessarily arise in the generation of the symmetric group by products of complexes some of which are subgroups.

Theorem 1: Let $\mathcal{S}$ be a stage, and let $P(\mathcal{S})$ contain a subgroup H of $S(X)$. Then there is a substage $\mathcal{R}$ of $\mathcal{S}$ which is made of square switches.

Proof: Define a relation $\mathcal{R}$ on $\{1, \cdots, N\}$ by the condition that $i\mathcal{R}j$ if and only if $j = \varphi(i)$ for some $\varphi \in H$. Since H is a subgroup, it must contain the identity permutation, i.e., $i\mathcal{R}i$ for all $i = 1, \cdots, N$. Let i, j, k be numbers in $\{1, \cdots, N\}$ such that $j = \varphi(i)$ and $k = \psi(j)$ for some permutations $\varphi, \psi \in H$. Then $\psi\varphi \in H$ and $k = \psi\varphi(i)$, that is, $i\mathcal{R}k$; hence $\mathcal{R}$ is transitive. Finally, if $j = \varphi(i)$ with $\varphi \in H$, we have $i = \varphi^{-1}(j)$ with $\varphi^{-1} \in H$, since H is a group. Hence $\mathcal{R}$ is an equivalence relation, and there is a partition Π such that

$$\mathcal{R} = \bigcup_{A \in \Pi} (A \times A).$$

Since $i\mathcal{R}j$ obviously implies $(i,j) \in \mathcal{S}$, we have

$$\mathcal{R} \subseteq \mathcal{S}.$$

$\mathcal{R}$ is clearly a substage of $\mathcal{S}$ made of square switches.

VI. AN EXAMPLE

As is well-known, elementary group theory contains many results that allow one to write a group as a product of complexes. These results often involve a *subgroup* of the group in question. We shall quote an elementary result of this kind, and interpret it in terms of a network that is known to be rearrangeable.

Let G be a group, and let H_1 and H_2 be subgroups of G, not necessarily distinct. A *double coset* is a complex of the form

$$H_1\varphi H_2 , \qquad \varphi \in G$$

It is a known result[5] that two double cosets are either identical or disjoint. Thus there is at least one complex M with the properties

$$\bigcup_{\varphi \in M} H_1\varphi H_2 = G$$

$$H_1\varphi H_2 \cap H_1\psi H_2 = \theta \qquad \text{if } \varphi \neq \psi, \quad \text{with } \varphi, \psi \in M.$$

In particular

$$G = H_1MH_2 ,$$

and we have factored G into a product of three complexes, two of which are subgroups. Now suppose that G is actually $S(X)$, the symmetric group of all permutations of N objects, and that m and n are positive integers such that $mn = N$. Let Π be a partition of $X = \{1, \cdots , N\}$ into m sets of n elements each, and let H be the largest strictly imprimitive subgroup of $S(X)$ whose sets of imprimitivity form Π. Also let φ be a self-inverse permutation, and Π_φ a partition, such that $A \in \Pi$, $B \in \Pi_\varphi$ imply*

$$| \varphi(A) \cap B | = 1.$$

Let K be the largest strictly imprimitive subgroup of $S(X)$ whose sets of imprimitivity form Π_φ.

By Remark 2, Section IV, H and K can each be generated by stages of square switches.

Returning to the earlier discussion leading to the factorization $G = H_1MH_2$, we let $H_1 = H_2 = H$. Now it can be seen that the complex

$$H\varphi K\varphi H$$

is generated by a network of the form shown in Fig. 2. By the Slepian-Duguid theorem (Beneš, Ref. 1, p. 1484) this network is rearrangeable, so that

$$H\varphi K\varphi H = S(X) .$$

* $| A |$ denotes the number of elements of a set A.

Since $\varphi = \varphi^{-1}$, the complex $\varphi K \varphi$ is itself actually the subgroup $\varphi^{-1} K \varphi$ conjugate to K. Thus for $G = S(X)$ and $H_1 = H_2 = K$, the factor M in

$$S(X) = HMH$$

can be chosen to be $\varphi^{-1} K \varphi$.

VII. SOME DEFINITIONS

The number of elements of a set A is denoted $| A |$. Let X, Y be arbitrary finite sets with $| X | = | Y |$, let B be a subset of Y, let Π_1 ,Π_2 be partitions of X, Y respectively, and let φ be a one-to-one map of Y onto X. Let θ be the null set.

Definition 8: $\varphi(B) = \{x \epsilon X: \varphi^{-1}(x) \epsilon B\}$.

Definition 9: φ *hits* Π_1 from B if and only if $A \epsilon \Pi_1$ implies $\varphi(B) \cap A \neq \theta$.

Definition 10: φ *covers* Π_1 from Π_2 if and only if $B \epsilon \Pi_2$ implies φ hits Π_1 from B.

Definition 11: $\varphi(\Pi_2)$ is the partition of X induced by φ acting on elements of Π_2 , i.e.,

$$\varphi(\Pi_2) = \{\varphi(B): B \epsilon \Pi_2\}.$$

Definition 12: ${}_B\varphi$ is the restriction of φ to B.

Let A be a subset of X.

Definition 13: ${}_A\Pi_1$ is the partition of A induced by Π_1 , i.e.,

$${}_A\Pi_1 = \{C \cap A: C \epsilon \Pi_1\}.$$

Definition 14: φ B-covers Π_1 from Π_2 if and only if ${}_B\varphi$ covers ${}_{\varphi(B)}\Pi_1$ from ${}_B\Pi_2$.

Definition 15: Let Π_0 , Π_1 be partitions of X. Then $\Pi_1 > \Pi_0$ (read "pi-one refines pi-zero") if and only if every set in Π_0 is a union of sets in Π_1 , and $\Pi_1 \neq \Pi_0$.

VIII. PRELIMINARY RESULTS

Lemma 1: Let X and Y be any sets with $| X | = | Y | < \infty$, let $\Pi_1 = \{A_1 , \cdots , A_n\}$ and $\Pi_2 = \{B_1 , \cdots , B_m\}$ be partitions of X and Y respectively, and suppose that for $k = 1, \cdots , n$ the union of any k elements of Π_1 has more elements than the union of any $k - 1$ elements of Π_2 . Then

(i) $m \geqq n$.

(*ii*) *For each one-to-one map f of X onto Y there exists a set of n distinct integers $k(1), \cdots, k(n)$ with $1 \leqq k(i) \leqq m, i = 1, \cdots, n$, and*

$$f(A_i) \cap B_{k(i)} \neq \theta \qquad i = 1, \cdots, n.$$

Proof: Since Π_1 and Π_2 are partitions, and $| X | = | Y |$,

$$\sum_{j=1}^{n} | A_i | = \sum_{j=1}^{m} | B_j |.$$

If m were less than n, then the union of m B's has as many elements as the union of n A's, for $m < n$; this contradicts the hypothesis. Let

$$K_i = \{j: f(l) \in B_j \text{ for some } l \in A_i\}, \qquad i = 1, \cdots, n.$$

Also let $A_{i(1)}, \cdots, A_{i(k)}$ be any k elements of Π_1, $1 \leqq k \leqq n$, and set

$$T = \bigcup_{j=1}^{k} K_{i(j)}.$$

All of the

$$\sum_{j=1}^{k} | A_{i(j)} |$$

elements of

$$\bigcup_{j=1}^{k} A_{i(j)}$$

are mapped by f into $| T |$ sets of Π_2. Since no union of $k - 1$ sets of Π_2 has

$$\sum_{j=1}^{k} | A_{i(j)} |$$

elements, it follows that $| T | \geqq k$. Thus the union of any k of the sets $\{K_i, i = 1, \cdots, n\}$ has at least k members. Hence by P. Hall's theorem[6] there is a set of n distinct representatives $k(1), \cdots, k(n)$ with

$$k(i) \in K_i \qquad i = 1, \cdots, n$$

$$k(i) \neq k(j) \qquad \text{for } i \neq j.$$

But clearly $k(i) \in K_i$ if and only if

$$f(l) \in B_{k(i)} \qquad \text{for some } l \in A_i,$$

that is, if and only if

$$f(A_i) \cap B_{k(i)} \neq \theta.$$

Lemma 2: Let Π_1, Π_2 *be partitions of sets* X, Y *respectively with the properties* $|X| = |Y|$ *and*

$$C_1, C_2 \in \Pi_1 \cup \Pi_2 \textit{ implies } |C_1| = |C_2|.$$

Then for every one-to-one map φ *of* X *onto* Y *there is a set* $D \subseteq X$ *such that* φ *hits* Π_2 *from* D *and* φ^{-1} *hits* Π_1 *from* $\varphi(D)$.

Proof: We observe that $|\Pi_1| = |\Pi_2|$, and that the conditions of Lemma 1 are satisfied, with $m = n$. For each onto map φ there is a set $D \subseteq X$ with $|D| = |\Pi_1|$, such that

$$A \in \Pi_1 \text{ implies } D \cap A \neq \theta, \qquad (\varphi^{-1} \text{ hits } \Pi_1 \text{ from } \varphi(D))$$

$$B \in \Pi_2 \text{ implies } \varphi(D) \cap B \neq \theta, \qquad (\varphi \text{ hits } \Pi_2 \text{ from } D).$$

Lemma 3: Let X *be any set and let* Π_1 *and* Π_2 *be partitions of* X *such that* $A,B \in \Pi_1 \cup \Pi_2$ *implies* $|A| = |B|$. *Then for every permutation* φ *on* X *there is a partition* Π_φ *of* X *such that* (*i*) φ *covers* Π_2 *from* Π_φ, (*ii*) φ^{-1} *covers* Π_1 *from* $\varphi(\Pi_\varphi)$,

$$(iii) \quad |\Pi_\varphi| = |A|, \qquad A \in \Pi_1 \cup \Pi_2$$

$$|B| = |\Pi_1| = |\Pi_2|, \qquad B \in \Pi_\varphi .$$

Proof: Let φ be a permutation on X. The hypothesis implies that the union of any k elements of Π_1 has more elements than the union of any k-1 elements of Π_2, for $k = 1, \cdots, |\Pi_1|$. Hence by Lemma 2, with $X = Y$ and $m = n$, there is a set $D_1 \subseteq X$ such that

$$\varphi \text{ hits } \Pi_2 \text{ from } D_1$$

$$\varphi^{-1} \text{ hits } \Pi_1 \text{ from } \varphi(D_1).$$

Now we consider the sets $X_1 = X - D_1$ and $Y_1 = Y - \varphi(D_1)$ partitioned by

$${}_{X_1}\Pi_1 \quad \text{and} \quad {}_{Y_1}\Pi_2 \text{ respectively}$$

and we apply Lemma 2 to ${}_{X_1}\varphi$, i.e., to the restriction of φ to $X - D_1$. This gives a new set $D_2 \subseteq X$ such that again

$$\varphi \text{ hits } \Pi_2 \text{ from } D_2$$

$$\varphi^{-1} \text{ hits } \Pi_1 \text{ from } \varphi(D_2).$$

We proceed in this manner till X and Y are exhausted, and set $\Pi_\varphi = \{D_1, \cdots, D_n\}$, where $n = |A|$ for all $A \in \Pi_1 \cup \Pi_2$. It is clear that Π_φ has the stated properties.

Let φ be a permutation on X, and let Π_1, Π_2 be partitions of X.

Lemma 4: If φ *covers* Π_1 *from* Π_2, *and* $B \in \Pi_2$ *implies* $|B| = |\Pi_1|$, *then* $A \in \Pi_1$ *implies* $|A| = |\Pi_2|$.

Proof: Since φ covers Π_1 from Π_2, then $A \in \Pi_1$ and $B \in \Pi_2$ imply that there is an $x \in B$ with $\varphi^{-1}(x) \in A$. Thus $\varphi^{-1}(x) \in A$ for at least $|\Pi_2|$ distinct values of x, and so $|A| \geqq |\Pi_2|$. Since $B \in \Pi_2$ implies $|B| = |\Pi_1|$, it follows that $|\Pi_1|$ divides $|X|$ and

$$|\Pi_2| = \frac{|X|}{|\Pi_1|}.$$

Clearly

$$\sum_{A \in \Pi_1} |A| = |X|.$$

Since there are $|\Pi_1|$ sets in Π_1 each with at least $|\Pi_2|$ elements,

$$\sum_{A \in \Pi_1} |A| \geqq |\Pi_1| \cdot |\Pi_2| = |X|.$$

If any $A \in \Pi_1$ had more than $|\Pi_2|$ elements the sum would exceed $|X|$, which cannot be. Thus $A \in \Pi_1$ implies $|A| = |\Pi_2|$.

IX. GENERATING THE PERMUTATION GROUP

In this section we exhibit a sufficient condition on permutations $\varphi_1, \cdots, \varphi_{s-1}$ and stages $\mathcal{S}_1, \cdots, \mathcal{S}_s$ under which the complex

$$P(\mathcal{S}_1)\varphi_1 \cdots \varphi_{s-1}P(\mathcal{S}_s)$$

is actually the whole symmetric group, and the corresponding network (obtained by linking $\mathcal{S}_i$ and $\mathcal{S}_{i+1}$ by φ_i, $i = 1, \cdots, s - 1$) is rearrangeable.

In order to focus on the mathematical character of the results, on their purely formal aspects divorced from physical considerations having to do with switches, etc., the conditions on the φ's and the $\mathcal{S}$'s purposely are phrased in a quite abstract way. Consequently, the practical implications and applications of the result may be unclear and require discussion. This discussion is given after the theorem has been stated, and is followed by its proof.

Theorem 2: Let $s > 3$ *be an odd integer, let* $\varphi_1, \cdots, \varphi_{s-1}$ *be permutations on* $X = \{1, \cdots, N\}$, *let* Π_k, $k = 1, \cdots, s$, *and* Π^k, $k = 1, \cdots, \frac{1}{2}(s - 1)$, *be partitions of* X, *and let*

$$\Psi^k = \begin{array}{ll} \{X\} & k = 0 \\ \varphi_k^{-1} \cdots \varphi_{\frac{1}{2}(s-1)}^{-1}(\Pi^k) & k = 1, \cdots, \tfrac{1}{2}(s-1) \\ \varphi_k \cdots \varphi_{\frac{1}{2}(s+1)}(\Pi^{s-k}) & k = \tfrac{1}{2}(s+1), \cdots, s-1, \\ \{X\} & k = s. \end{array}$$

Suppose that

(i) *If* $s \geqq 3$, *then* $\Pi^k < \Pi^{k+1}$, $k = 0, \cdots, \frac{1}{2}(s-3)$.

(ii) $\Pi_{\frac{1}{2}(s+1)} = \Pi^{\frac{1}{2}(s-1)}$

(iii) *For* $k = 1, \cdots, \frac{1}{2}(s-1)$, *and every* $B \in \Psi^{k-1}$, φ_k^{-1} *B-covers* Π_k *from* φ_k *from* $\varphi_k(\Psi^k)$.

(iv) *For* $k = \frac{1}{2}(s+1), \cdots, s-1$, *and every* $B \in \Psi^{k+1}$, φ_k *B-covers* Π_{k+1} *from* $\varphi_k^{-1}(\Psi^k)$.

(v) *If* $A \in \Pi^k$ *and* $B \in \Psi^{k-1} \cup \Psi^{k+\frac{1}{2}(s+1)}$ *then* $|{}_B\Pi_k| = |{}_B\Pi_{s-k+1}| = |A|$, $k = 1, \cdots, \frac{1}{2}(s-1)$.

Let H_k, $k = 1, \cdots, s$ *be the largest strictly imprimitive subgroup of* $S(X)$ *whose sets of imprimitivity are exactly the elements of* Π_k (*i.e.*, $A \in \Pi_k$ *implies* $\{{}_A\varphi\colon \varphi \in H_k\} = S(A)$). *Then the complex K defined by*

$$K = H_1\varphi_1H_2 \cdots H_{s-1}\varphi_{s-1}H_s$$

has the property $K = S(X)$, *and any network generating this complex is rearrangeable.*

The theorem given above does not provide any new designs of rearrangeable networks that are not already implicit in the work of M. C. Paull[3] and D. Slepian;[1] thus no new principle is involved. Rather, in formulating the result, we have sought insight by stating a generalized, purely combinatory form of these previous results. The theorem exhibits this generalization, first as providing a way of generating the symmetric group in a fixed number of multiplications of certain restricted group elements, and second as based on some purely abstract properties of some partitions and permutations.

As in A. M. Duguid's proof of the Slepian-Duguid theorem,[1] the basic combinatory theorem of P. Hall on distinct representatives of subsets is used repeatedly. This means (roughly) that the proof proceeds by showing that an arbitrary permutation (to be realized in the network) can be decomposed into submaps each of which can be realized in a disjoint part of the network, thereby not interfering with the realization of the other submaps. A significant departure from Ref. 3 is that we try to obtain rearrangeability directly from conditions that are stated for the network as a whole, as well as by building it up from rearrangeable subnetworks.

The following intuitive guides should be useful in understanding

Theorem 2. The permutations $\varphi_1, \cdots, \varphi_{s-1}$ are of course intended to be those corresponding to the link patterns between the stages of a network. The partition Π_k corresponds to the assignment of the terminals entering the kth stage to various square switches, all $u \in A$ for $A \in \Pi_k$ being on the same switch. The partitions Π^k are used in defining the submaps mentioned above.

The "covering" properties (*iii*) and (*iv*) of the φ_k in Theorem 2 ensure (roughly) that the φ_k are sufficiently mixing or distributive to be able to generate *all* permutations in the restricted ways permitted in the definition of K. They are generalizations of the property, exhibited in Fig. 2, that every middle switch is connected to every side switch by a link. The property (*v*), finally, implies that various sets of switches all have the same cardinality; this ensures (again, roughly) that if a crosspoint is not being used for a connection between one inlet-outlet pair, then it can be used for a connection between some other pair.

Proof of Theorem 2: We use induction on odd $s \geqq 3$. If $s = 3$, there is only one Π^k, viz. Π^1. Let φ be a permutation; we show that

$$\varphi \in H_1\varphi_1H_2\varphi_2H_3 .$$

The argument to be given is constructive, in that we do not use proof by contradiction, but actually give a kind of recipe for finding three permutations $\eta_i \in H_i$, $i = 1, 2, 3$, with

$$\varphi = \eta_1\varphi_1\eta_2\varphi_2\eta_3 .$$

To prove the theorem for $s = 3$, it is enough for $i = 1, 2, 3$ to exhibit a partition $\Pi(i)$ and to define η_i on $A \in \Pi(i)$, i.e., to give

$$_A\eta_i , \qquad A \in \Pi(i), \qquad i = 1, 2, 3.$$

Condition (*v*) for $k = 1$ $(=\frac{1}{2}(s - 1)$ here) tells us that for $A \in \Pi^1$

$$| \Pi_1 | = | \Pi_3 | = | A |.$$

However, the "middle-stage" condition (*ii*) states that $\Pi_2 = \Pi^1$. Hence $| \Pi_1 | \cdot | \Pi_2 | = N$. Since [condition (*iii*) now] φ_1^{-1} covers Π_1 from $\varphi_1(\Psi^1) = \varphi_1\varphi_1^{-1}(\Pi^1) = \Pi_2$, it follows that $B \in \Pi_1$ implies $| B | \geqq | \Pi_2 |$. If for some $B \in \Pi_1$ it was true that $| B | < | \Pi_2 |$, then

$$\sum_{B \in \Pi_1} | B | < | \Pi_1 | \cdot | \Pi_2 | = N$$

which is impossible. Thus $| B | = | \Pi_2 |$ for $B \in \Pi_1$. In exactly the same way, using condition (*iv*), we find that $C \in \Pi_3$ implies $| C | = | \Pi_2 |$. Therefore $B, C \in \Pi_1 \cup \Pi_3$ implies $| B | = | C |$.

Returning now to the chosen permutation φ, we apply Lemma 3 to conclude that there exists a partition Π_φ of X such that φ covers Π_3 from Π_φ, φ^{-1} covers Π_1 from $\varphi(\Pi_\varphi)$, $|\Pi_\varphi| = |A|$ for $A \in \Pi_1 \cup \Pi_3$, and $|B| = |\Pi_1| = |\Pi_3|$ for $B \in \Pi_\varphi$. Hence also $|\Pi_\varphi| = |\Pi_2|$.

Let μ: $\Pi_\varphi \leftrightarrow \Pi_2$ be *any* map of Π_φ onto Π_2. The desired partitions $\Pi(i)$, $i = 1, 2, 3$ will be taken to be

$$\begin{aligned} \Pi(1) &= \varphi_1(\Pi_2) \\ \Pi(2) &= \{\mu(D) : D \in \Pi_\varphi\} = \Pi_2 \\ \Pi(3) &= \Pi_\varphi \end{aligned}$$

and the desired permutations η_i, $i = 1, 2, 3$, are defined so as to have these properties: for $D \in \Pi_\varphi$

$$\begin{aligned} \eta_3 &: \varphi(D) \leftrightarrow \varphi_2^{-1}(\mu(D)) \\ \eta_1 &: \varphi_1(\mu(D)) \leftrightarrow D \\ \eta_2 &: \varphi_2\eta_3(D) \leftrightarrow \varphi_1^{-1}\eta_1^{-1}\varphi(D). \end{aligned}$$

That this can be done (uniquely, indeed) can be seen as follows: Let $D \in \Pi_\varphi$, and $\mu(D) = B \in \Pi_2$. Since φ^{-1} covers Π_1 from $\varphi(\Pi_\varphi)$, and φ covers Π_3 from Π_φ, it must be true that

(1) φ^{-1} hits Π_1 from $\varphi(D)$

(2) φ hits Π_3 from D.

But at the same time, by conditions (*iii*) and (*iv*), and the fact that $\Pi^1 = \Pi_2$, φ_1^{-1} covers Π_1 from Π_2 and φ_2 covers Π_3 from Π_2, and so

(3) φ_1^{-1} hits Π_1 from B

(4) φ_2 hits Π_3 from B.

Thus if $u \in D \cap A$ and $A \in \Pi_3$, there is a unique $v \in A$ such that $\varphi_2(v) \in B$, and we take $\eta_3(u) = v$. Similarly, if $z = \varphi(u) \in C$ and $C \in \Pi_1$, there is a unique $w \in C$ such that $\varphi_1^{-1}(w) \in B$, and we take $\eta_1(w) = z$. Finally, define η_2 so that $\eta_2(\varphi_2(v)) = \varphi_1^{-1}(w)$. Since $\mu(\cdot)$ is onto, each $D \in \Pi_\varphi$ deals with a unique $B = \mu(D) \in \Pi_2$, and the definition of η_i, $i = 1, 2, 3$ can be made for each D and its associated B, independently of the others. It is apparent that

$$\eta_1\varphi_1\eta_2\varphi_2\eta_3(D) = \varphi(D), \qquad D \in \Pi_\varphi,$$

or

$$\varphi = \eta_1\varphi_1\eta_2\varphi_2\eta_3 .$$

Since η_i for $i = 1, 2, 3$ is onto, and is a subset of

$$\bigcup_{E \in \Pi_i} E \times E,$$

it follows that $\eta_i \in H_i$, $i = 1, 2, 3$, and thus that

$$K = S(X).$$

We now assume, as an hypothesis of induction, that the theorem is true for a given odd $s - 2 \geqq 3$, and that we are given permutations φ_k, and partitions $\{\Pi_k\}$ and $\{\Pi^k\}$, satisfying the conditions of the theorem.

No loss of generality is sustained if it is assumed that each $A \in \Pi^2$ is invariant under $\varphi_2, \cdots, \varphi_{s-2}$. This invariance can always be achieved by redefining the φ_i, without loss of properties (*iii*) to (*v*). It can now be seen that for $k = 2, \cdots, s - 2$ the restrictions

$$_A\Pi_k, \qquad _A\Pi^k, \qquad \text{with } A \in \Pi^2,$$

satisfy all the conditions on Π_k, Π^k (respectively) used in Theorem 2. Hence by the hypothesis of induction, for each $A \in \Pi^2$, the restriction of the complex

$$H_2\varphi_2 \cdots \varphi_{s-2}H_{s-1}$$

to A generates $S(A)$. The argument used for the case $s = 3$ can now be used to complete the induction, $\Psi^2(= \Psi^{s-1}$ here) playing the role of Π_2.

X. CONSTRUCTION OF A CLASS OF REARRANGEABLE NETWORKS

We consider a network ν built of an odd number $s \geqq 3$ of stages,

$$\nu = \mathcal{S}_1\varphi_1 \cdots \varphi_{s-1}\mathcal{S}_{s-1},$$

satisfying the symmetry conditions

$$\left.\begin{aligned} \varphi_k &= \varphi_{s-k}^{-1} \\ \mathcal{S}_k &= \mathcal{S}_{s-k+1} \end{aligned}\right\} k = 1, \cdots, \tfrac{1}{2}(s - 1),$$

with each stage $\mathcal{S}_k$ made of identical square switches. The φ_k will be chosen in the following way: order the switches of each stage; to define φ_k for a given $1 \leqq k \leqq \frac{1}{2}(s - 1)$ take the first switch of $\mathcal{S}_k$, say with n outlets and n a divisor of N, and connect these outlets one to each of the first n switches of $\mathcal{S}_{k+1}$; go on to the second switch of $\mathcal{S}_k$ and connect its n outlets one to each of the *next* n switches of $\mathcal{S}_{k+1}$; when all the switches of $\mathcal{S}_{k+1}$ have one link on the inlet side, start again with the first switch; proceed cyclically in this way till all the outlets of $\mathcal{S}_k$ are assigned. (See Fig. 3.)

We shall show that a network ν constructed in this way is always rearrangeable.

Theorem 3: Let $s \geqq 3$ be an odd integer. Let n_k, $k = 1, \cdots, (s + 1)/2$, be any positive integers such that

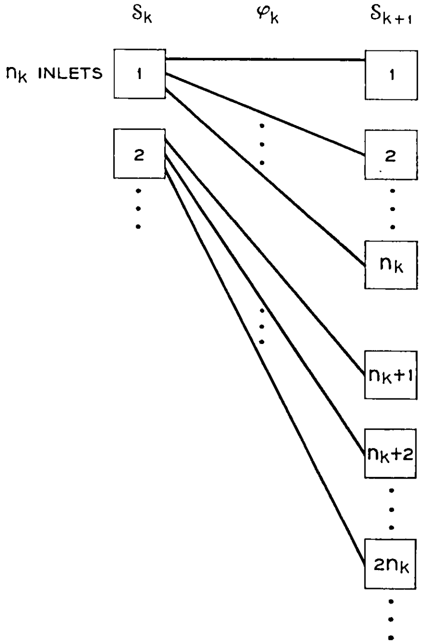


Fig. 3 — Assignment of inlets and outlets of rearrangeable network.

$$\prod_{k=1}^{\frac{1}{2}(s+1)} n_i = N \quad \text{and} \quad n_i \geqq 2.$$

For each $k = 1, \cdots, (s+1)/2$, *let* Π_k *be the partition of* $X = \{1, \cdots, N\}$ *into the* N/n_k *sets of the form*

$$A_{ki} = \{t\colon (i-1)n_k < t \leqq in_k\} \qquad i = 1, \cdots, N/n_k .$$

Let φ_k, $k = 1, \cdots, (s-1)/2$, *be permutations with the property that* $n \equiv t$ (*mod* N/n_{k+1}) *if and only if*

$$\varphi_k(n) \in A_{k+1,t} \qquad t = 0, \cdots, (N/n_{k+1}) - 1.$$

Define

$$\varphi_k = \varphi_{s-k}^{-1} \quad \text{for} \quad (s-1)/2 < k \leqq s-1.$$

Let $\mathcal{S}_k$, $k = 1, \cdots, s$, *be the stages made of square switches defined by*

$$\mathcal{S}_k = \mathcal{S}_{s-k+1} \qquad k = 1, \cdots, (s-1)/2$$

$$\mathcal{S}_k = \bigcup_{A \in \Pi_k} A \times A$$

and ν be the network constructed by putting a link pattern corresponding to φ_k, $k = 1, \cdots, s - 1$ between stages $\mathcal{S}_k$ and $\mathcal{S}_{k+1}$. Then ν is rearrangeable.

Proof: It is readily seen that for $k = 1, \cdots, s$

$$\begin{aligned} P(\mathcal{S}_k) &= \text{the largest strictly imprimitive subgroup with sets of imprimitivity } \Pi_k, \\ &= H_k, \end{aligned}$$

in the notation of Theorem 2. Thus to prove the theorem by appeal to Theorem 2 it is enough to exhibit suitable partitions Π^k, $k = 1, \cdots, \frac{1}{2}(s - 1)$, and to show that these, together with $\varphi_1, \cdots, \varphi_{s-1}$, satisfy the conditions of Theorem 2. For $k = 1, \cdots, \frac{1}{2}(s - 1)$, set

$$\Pi^k = \text{class of all } \{j\colon (i - 1)N/n_1 \cdots n_k < j \leqq iN/n_1 \cdots n_k, \quad i = 1, \cdots, n_1 n_2 \cdots n_k\}.$$

It is evident that the Π^k are successively finer partitions, i.e., that

$$\Pi^k < \Pi^{k+1} \qquad k = 1, \cdots, \tfrac{1}{2}(s - 3).$$

Also, since $\Pi^{\frac{1}{2}(s-1)}$ consists of the $n_1 n_2 \cdots n_{\frac{1}{2}(s-1)}$ sets

$$\{j\colon (k - 1)n_{\frac{1}{2}(s+1)} < j \leqq kn_{\frac{1}{2}(s+1)}, \; k = 1, \cdots, n_1 n_2 \cdots n_{\frac{1}{2}(s-1)}\}$$

it can be seen that

$$\Pi^{\frac{1}{2}(s-1)} = \{A_{\frac{1}{2}(s+1),i} : 1 \leqq i < N/n_{\frac{1}{2}(s+1)}\},$$

and hence that the middle stage condition (*ii*) in Theorem 2,

$$\Pi_{\frac{1}{2}(s+1)} = \Pi^{\frac{1}{2}(s-1)},$$

is satisfied.

The remainder of the proof, in which the requisite covering properties of the φ_k are demonstrated, is based on some auxiliary results.

Lemma 5: For $k = 2, \cdots, \frac{1}{2}(s - 1)$, and $1 \leqq i \leqq N/n_n n_{n+1}$ the following identity holds

$$\bigcup_{\substack{t \equiv i-1 \\ \left(mod \frac{N}{n_k n_{k+1}}\right)}} A_{k,t} = \varphi_k^{-1}\Big(\bigcup_{i-1 < \frac{t}{n_k} \leqq i} A_{k+1,t}\Big)$$

and the sets on the right are disjoint for different i.

Proof: Since

$$\varphi_k^{-1}(A_{k+1,t}) = \{n\colon n \equiv t \;(\text{mod } N/n_{k+1}), \quad 1 \leqq n \leqq N\}$$

the union on the right in the lemma is the set of all n that are $\equiv t$ (mod

$N/n_{k+1})$ for some t with $(i-1)n_k < t \leqq in_k$. Consider such an n, with say

$$n = \frac{lN}{n_{k+1}} + t, \qquad \begin{array}{l} i-1 < \dfrac{t}{n_k} \leqq i \\ 0 \leqq l < n_{k+1}. \end{array}$$

Then

$$n = n_k \frac{lN}{n_k n_{k+1}} + (i-1)n_k + u,$$

with $0 < u \leqq n_k$, and so

$$n \in A_{k,(lN/(n_k n_{k+1}) + i - 1)}$$

or

$$n \in A_{k,t} \quad \text{with} \quad t \equiv (i-1) \pmod{N/n_k n_{k+1}}.$$

Since the representation of n in terms in l and u is unique, the lemma follows.

The practical or physical import of the lemma is this: In any stage $k+1$, $1 \leqq k \leqq \frac{1}{2}(s-1)$, the ith block of n_k switches is connected by the link pattern φ_k to exactly those n_{k+1} switches (in the kth stage) whose number $t \equiv (i-1) \pmod{N/n_n n_{n+1}}$.

Definition: For $1 \leqq i \leqq n$, and $2 \leqq k \leqq m = \frac{1}{2}(s+1)$

$$B_{ik} = \bigcup_t \left\{ A_{k,t} : t \equiv r (\operatorname{mod} n_1 n_2 \cdots n_{k-1}) \text{ for some } r \text{ with } (i-1) < \frac{r}{n_2 n_3 \cdots n_{k-1}} \leqq i \right\}$$

where $n_2 n_3 \cdots n_{k-1}$ is taken $= 1$ if $k = 2$.

Lemma 6: For $1 \leqq i \leqq n_1$ *and* $2 \leqq k < m$

$$B_{ik} = \varphi_k^{-1}(B_{i,k+1}).$$

Proof: We show that the right-hand side contains the left. Equality then follows from Lemma 5, since $B_{i,k+1}$ will always be a union of sets of the form

$$\bigcup_{j-1 < \frac{\tau}{n_k} \leqq j} A_{k+1,\tau}.$$

This is because $t \equiv r \pmod{u}$ if and only if $t + 1 \equiv (r + 1) \pmod{u}$, if $0 \leqq r < u$. Consider then an $n \epsilon B_{ik}$. There is a $t \geqq 1$ congruent to an $r \pmod{n_1 n_2 \cdots n_{k-1}}$, with

$$i - 1 < \frac{r}{n_2 n_3 \cdots n_{k-1}} \leqq i,$$

such that $n \epsilon A_{k,t}$. The latter fact implies that

$$(t - 1)n_k < n \leqq tn_k .$$

Now $\varphi_k(n) \epsilon A_{k+1,\gamma}$, where n is congruent to $\tau \pmod{N/n_{k+1}}$, so we can represent n in the form

$$n = \frac{aN}{n_{k+1}} + \tau.$$

Hence

$$(t - 1)n_k < \frac{aN}{n_{k+1}} + \tau \leqq tn_k .$$

Writing $t = ln_1 n_2 \cdots n_{k-1} + r$, with

$$(i - 1)n_2 n_3 \cdots n_{k-1} < r \leqq in_2 n_3 \cdots n_{k-1} ,$$

we see that

$$qn_1 \cdots n_k + (r - 1)n_k < \tau \leqq qn_1 \cdots n_k + rn_k$$

where

$$q = l - \frac{aN}{n_1 n_2 \cdots n_{k+1}} .$$

It follows that τ is congruent $(\mathrm{mod}\ n_1 n_2 \cdots n_k)$ to some integer p in the region

$$(i - 1)n_2 \cdots n_k < p \leqq in_2 \cdots n_k ,$$

and thus $\varphi_k(n) \epsilon B_{i,k+1}$, completing the proof of Lemma 6.

Now if $k = m$, the defining condition that $t \equiv r \pmod{n_1 \cdots n_{m-1}}$ for some r with

$$(i - 1)n_2 \cdots n_{m-1} < r \leqq in_2 \cdots n_{m-1} ,$$

used in the definition of B_{ik}, can be put into a slightly different form. In this case we must have

$$1 \leqq t \leqq N/n_m = n_1 n_2 \cdots n_{k-1}$$

and so t can only be congruent to r by being equal to r, that is

$$(i - 1)n_2 \cdots n_{m-1} < t \leqq in_2 \cdots n_{m-1} .$$

Hence it can be seen that

$$\{B_{im} , i = 1, \cdots , n_1\} = \Pi^1.$$

Applying Lemma 6 $(m - 2)$ times we find that

$$\begin{aligned}\{B_{i2} , i = 1, \cdots , n_1\} &= \varphi_2^{-1} \cdots \varphi_m^{-1}(\Pi^1)\\ &= \varphi_1(\Psi^1).\end{aligned}$$

Now $n \equiv t \pmod{N/n_2}$ if and only if $\varphi_1(n) \in A_{2t}$, $t = 1, \cdots , N/n_2$. Also, by definition of B_{i2},

$$B_{i2} = \bigcup_{t \equiv i (\mathrm{mod}\ n_1)} A_{2t}.$$

Let $\varphi_1(n) \in B_{i2}$, $\varphi_1(n) \in A_{2t}$. Then n has the form

$$n = \frac{aN}{n_1 n_2} n_1 + t = \left(\frac{aN}{n_1 n_2} + b\right) n_1 + i,$$

so $n \equiv i \pmod{n_1}$. Since $| B_{i2} | = | B_{im} | = n_2 \cdots n_m = N/n_1$, it follows that B_{12} is the φ_1 image of N/n_1 integers each of which is congruent to i (mod n_1). Since each such integer must be in a different A_{1t}, it follows that φ_1^{-1} covers Π_1 from $\varphi_1(\Psi^1)$.

The remaining conditions in Theorem 3 can be demonstrated in essentially the same way; one has merely to identify the sets in question, and use Lemma 5 and an analog of Lemma 6. The details will be omitted.

REFERENCES

1. Beneš, V. E., On Rearrangeable Three-Stage Connecting Networks, B.S.T.J., **41,** 1962, pp. 1481–1492.
2. Beneš, V. E., Optimal Multistage Rearrangeable Connecting Networks, B.S.T.J., this issue, p. 1641.
3. Paull, M. C., Reswitching of Connection Networks, B.S.T.J., **41,** May, 1962, pp. 833–855.
4. Hall, M., Jr., *The Theory of Groups*, Macmillan, New York, 1959, p. 64.
5. *Ibid.*, p. 10.
6. Hall, P., On Representatives of Subsets, J. London Math. Soc., **10** (1935), pp. 26–30.

Optimal Rearrangeable Multistage Connecting Networks

By V. E. BENEŠ

(Manuscript received March 12, 1964)

A rearrangeable connecting network is one whose permitted states realize every assignment of inlets to outlets—that is, one in which it is possible to rearrange existing calls so as to put in any new call. In the effort to provide adequate telephone service with efficient networks it is of interest to be able to select rearrangeable networks (from suitable classes) having a minimum number of crosspoints. This problem is fully resolved for the class of connecting networks built of stages of identical square switches arranged symmetrically around a center stage: roughly, the optimal network should have as many stages as possible, with switches that are as small as possible, the largest switches being in the center stage; the cost (in crosspoints per inlet) of an optimal network of N inlets and N outlets is nearly twice the sum of the prime divisors of N, while the number of its stages is 2x − 1, where x is the number of prime divisors of N, in each case counted according to their multiplicity. By using a large number of stages, these designs achieve a far greater combinatorial efficiency than has been attained heretofore.

I. INTRODUCTION

A study of rearrangeable connecting networks, begun in a previous paper,[1] is here continued; the object of the present work is to solve the synthesis problem of choosing, from a class of networks that are built of stages of square switches and satisfy some reasonable conditions on uniformity of switch size, a rearrangeable connecting network having a *minimum* number of crosspoints. Some of the terminology, notation, and results of Ref. 1 are used, and familiarity with it will be assumed from Section IV on.

Naturally, we do not pretend that minimizing the number of crosspoints (used to achieve a given end) is the only consideration relevant to the design of a connecting network. Other factors, like the number of memory elements, the amount and placing of terminal equipment, the ease with which a network is controlled (e.g., the possibility of reliable end-marking), etc., may be of overriding significance, depending on the technology used. Still, it is important to know the limits of the region of possible designs, and these are obtained by optimizing on one variable without attention to others.

The problem of designing a good rearrangeable network was (probably first) considered in a paper of C. E. Shannon[2] investigating memory requirements in a telephone exchange. On the networks that he considered he imposed the realistic "separate memory condition" to the effect that in operation a separate part of the memory can be assigned to each call in progress. This means that completion of a new call or termination of an old call will not disturb the state of memory elements associated with any call in progress. Shannon showed that under this assumption a two-sided rearrangeable network, with N inlets and N outlets, and N a power of 2, requires at least

$$2N \log_2 N$$

memory elements (e.g., relays). He gave a design which actually realized this lower bound using

$$4(2^N - 1)\log_2 N$$

crosspoints (e.g., relay contacts). His design had the disadvantage of having very large numbers of contacts on certain relays. It is to be noted that Shannon was concerned with minimizing the number of memory elements, without regard to the number of crosspoints.

Shannon's separate memory condition is actually met by modern connecting networks that are of current practical interest, viz., by the networks made of stages of crossbar switches, considered here. For indeed, an inlet relay on an $n \times n$ crossbar switch is used to close any and each of n crosspoints: the exact one that closes depends on what outlet relay is simultaneously activated.

In this paper we consider the problem of minimizing the number of crosspoints in a network built of square switches, without attention to the number of relays. The following result (a consequence of Theorem 8) then complements Shannon's: For N a power of 2 it is possible to design a rearrangeable network with N inlets and N outlets using $4N \log_2 N - 6N$ relays and $4N(\log_2 N - 2)$ crosspoints. The figure for relays is roughly twice Shannon's while that for crosspoints is much smaller than his, for N large. In our design, no relay controls more than 4 contacts.

II. SUMMARY AND DISCUSSION

In Section III we discuss the notion of the combinatory power or efficiency of a connecting network, and propose to define it as the fraction r of permutations it can realize. According to this definition the four-stage No. 5 crossbar type of network with 10 × 10 switches has efficiency r close to zero, although it turns out that for the same number ($\approx$1000) of terminals there are networks that achieve $r = 1$ with a smaller number of crosspoints.[1] This greater efficiency is obtained by using many more stages than four.

Preliminaries are treated in Section IV. Particular attention is drawn to the class C_N of all two-sided networks having N inlets and N outlets, and built of stages of identical square switches symmetrically arranged around a center stage. The cost $c(\nu)$ of such a network ν is defined as the total number of crosspoints, divided by N. It is proposed to select rearrangeable networks ν from C_N that have minimal cost $c(\nu)$. This problem is attacked in Section V by defining (*i*) a map T from C_N to a special set A such that $c(\nu)$ is a function of $T(\nu) \epsilon A$, and (*ii*) a partial ordering of A. It is then shown (Section VI) that (roughly) a network ν is optimal if and only if $T(\nu)$ is at the bottom of the partial ordering of A. This result allows one to identify (Section VII) the optimal networks in C_N. Their general characteristics are these: Except in some easily enumerated cases, the optimal network should have as many stages as possible, and switches that are as small as possible, the largest switches being in the middle stage; the cost $c(\nu)$ of an optimal network ν is very nearly twice the sum of the prime divisors of N, while the number of its stages is $2x - 1$, where x is the number of prime divisors of N.

Our chief conclusion is that by using many stages of small switches it is possible to design networks that are rearrangeable and cost less (in crosspoints per terminal) than networks in current use, which are far from being rearrangeable. The price paid for this great increase in combinatory power is the current difficulty of controlling networks of many stages. This difficulty is technological, though, and will decrease as improved circuits are developed.

III. THE COMBINATORY POWER OF A NETWORK

A principal reason why *rearrangeable* networks are of practical interest is (of course) that they can be operated as nonblocking networks. If the control unit of the connecting system using the rearrangeable network is made complex enough, it is in principle possible to rearrange calls in progress, repeatedly, in such a way that no call is ever blocked.

At present this possibility is being exploited in only a few special-purpose systems, because of the large amount of searching and data-processing it requires.

However, there is another reason why rearrangeable networks should evoke current interest. Even if we do not care to exploit it, the property of rearrangeability in a connecting network is an indication of its combinatory power or reach, and so can be used as a qualitative "figure of merit" for comparing networks. Other things being equal, a rearrangeable network is better than one which cannot realize all assignments of inlets to outlets. Rearrangeability expresses to some extent the efficiency with which crosspoints have been utilized in designing a connecting network for *distribution*, that is, for reaching many outlets from inlets.

If a numerical measure is called for, one can use the fraction of realizable maximal assignments. For a network ν with the same number N of inlets as outlets, and with inlets disjoint from outlets, this is just

$$\begin{aligned} r &= \frac{\text{number of permutations realizable by } \nu}{N!} \\ &= \text{combinatorial power of } \nu. \end{aligned}$$

It is apparent that $0 \leqq r \leqq 1$, and that for a rearrangeable network $r = 1$. Also, r may be viewed as the chance that a permutation chosen at random will be realizable.

We shall calculate a bound on the combinatorial power r of the kind of connecting network most commonly found in modern telephone central offices. This is the network illustrated in Fig. 1. We choose the switch size $n = 10$ as a representative value; the network then has $N = 1000$ inlets, as many outlets, and 4×10^4 crosspoints. Clearly, the network can realize at most all the permutations that take exactly n terminals from each frame on the inlet side into each frame on the outlet side. Now a frame has n^2 inlets (outlets), and there are

$$\frac{n^2!}{(n!)^n}$$

ways of partitioning n^2 things into n groups of n each. Since there are $2n$ frames, there are

$$\left(\frac{n^2!}{(n!)^n}\right)^{2n}$$

ways of choosing n groups of n each on each frame, and assigning inlet groups to outlet groups (one-to-one and onto) in such a way that for

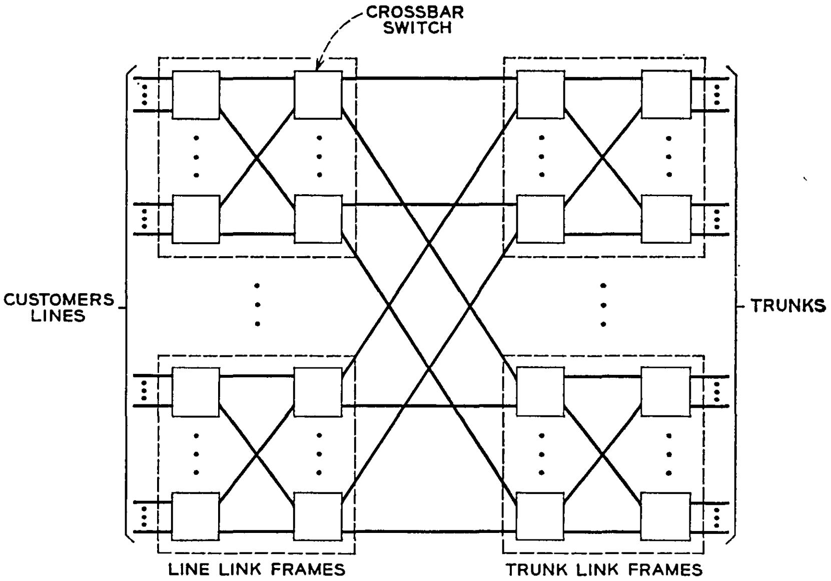


Fig. 1 — Structure of No. 5 crossbar network.

every inlet frame and every outlet frame exactly one group on the inlet frame is assigned to a group on the outlet frame. There are n^2 groups on a side (inlet or outlet), and within each group (at most) $n!$ permutations can be made, i.e., each inlet group can be mapped, terminal by terminal, in at most $n!$ ways onto its assigned outlet group. Hence at most

$$\left(\frac{(n^2!)}{(n!)^n}\right)^{2n} (n!)^{n^2}$$

permutations can be realized. There are $N = n^3$ terminals on a side, and a total of $n^3!$ possible permutations in all. Thus

$$r \leqq \frac{(n^2!)^{2n}}{n!^{n^2} n^3!} \, .$$

For $n = 10$, with

$$20 \log (100!) = 3159.4000$$

$$100 \log (10!) = 655.976$$

$$\tfrac{1}{2} \log 2\pi = 0.39959$$

$$\log (x!) \sim \tfrac{1}{2} \log 2\pi + (x + \tfrac{1}{2}) \log x - x \log_{10} e$$

we find roughly

$$r \leqq 10^{-64}.$$

Thus only a vanishingly small fraction of all possible permutations can actually be achieved by the No. 5 crossbar network (illustrated in Fig. 1) for $n = 10$, a reasonable switch size.

In the example calculated, the network has a "cost" of 40 crosspoints per terminal on a side. Much of the force of the example would be lost if it were in fact impossible to achieve high values of r (i.e., near 1) without incurring a great increase in the cost in crosspoints per terminal. This, however, is not the case. It follows from our Theorem 8 that a *rearrangeable* network ($r = 1$) can be designed for $N = 1024$ terminals on a side using only

$$4(\log_2 N - 2) = 32$$

crosspoints per terminal. Thus it is actually possible to *achieve* $r = 1$ for *more* than 1000 lines with *fewer* than 40 crosspoints per line. The network that does this turns out to have 17 stages instead of 4, an illustration of the way that allowing many stages can lead to vastly more combinatorially efficient network designs. The middle stage of this network consists of a column of 256 4×4 switches, and each of the other 16 stages, arranged symmetrically, consists of a column of 512 2×2 switches. For $k = 1, \cdots, 8$, the kth stage is connected to the $(k + 1)$th as follows: the first outlet of the first switch of stage k goes to the first switch of stage $(k + 1)$, the second outlet of the first switch of stage k goes to the second switch of stage $(k + 1)$, the first outlet of the second switch of stage k goes to the third switch of stage $(k + 1)$, etc., as in Fig. 2 with $1 \leqq k \leqq 7$; when each switch of stage $(k + 1)$ has 1 link on it the process starts over again with the first switch, and continues cyclically until all the links from stage k are assigned. The connections between stages k and $k + 1$ for $k = 9, \cdots, 17$ are the inverses of those for $k = 1, \cdots, 8$, so that the network is symmetric about the middle stage.[1]

IV. PRELIMINARIES

The symbol C_N, $N \geqq 2$, is used to denote the class of all connecting networks ν with the following properties:*

(1) ν is two-sided, with N terminals on each side

(2) ν is built of an odd number s of stages $\mathcal{S}_k$, $k = 1, \cdots, s$, of

* Familiarity with Ref. 1 is assumed henceforth.

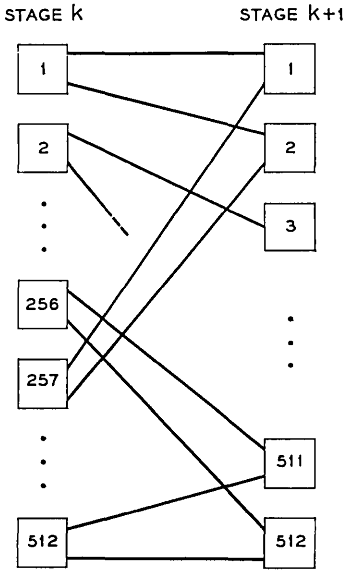


Fig. 2 — Link assignment.

square switches, i.e., there are permutations $\varphi_1, \cdots, \varphi_{s-1}$ such that

$$\nu = \mathcal{S}_1\varphi_1\mathcal{S}_2, \cdots, \varphi_{s-1}\mathcal{S}_s$$

(3) ν is *symmetric* in the sense that

$$\mathcal{S}_k = \mathcal{S}_{s-k+1} \qquad \text{for} \qquad k = 1, \cdots, \tfrac{1}{2}(s-1)$$

(4) With the notation

$$s = s(\nu) = \text{number of stages of } \nu$$

$$n_k = n_k(\nu) = \text{switch size in the } k\text{th stage of } \nu,$$

ν has N/n_k identical switches in stage k, i.e., each stage $\mathcal{S}_k$ is of the form

$$\bigcup_{A \in \Pi} A \times A$$

for some partition Π with $|A| = |B|$ for all $A, B \in \Pi$.

The defining conditions of C_N imply that

$$n_k = n_{s-k+1} \qquad \text{for} \qquad k = 1, \cdots, (s-1)/2$$

and that

$$\prod_{k=1}^{\frac{1}{2}(s+1)} n_k = N.$$

It is assumed throughout that $n_k(\nu) \geqq 2$ for all ν and all $k = 1, \cdots, s(\nu)$.

The *cost per terminal* (on a side) $c(\nu)$ of a network $\nu \in C_N$ is defined to be the total number of crosspoints of ν divided by the number N of terminals on a side. Since there are N/n_k $n_k \times n_k$ switches in stage k, the total number of crosspoints is (using the symmetry condition)

$$\begin{aligned}\sum_{k=1}^{s} (N/n_k)\cdot n_k^2 &= N\sum_{k=1}^{s} n_k \\ &= N\left(n_{\frac{1}{2}(s+1)} + 2\sum_{k=1}^{\frac{1}{2}(s-1)} n_k\right)\end{aligned}$$

and so

$$c(\nu) = n_{\frac{1}{2}(s+1)} + 2\sum_{k=1}^{\frac{1}{2}(s-1)} n_k .$$

A network ν is called *optimal* if

$$c(\nu) = \min \{c(\mu) : \mu \in C_N\}.$$

It is clear that the cost per terminal of a network $\nu \in C_N$ depends only on the switch sizes, and not at all on the permutations that define the link patterns between stages.

Also, it is apparent from Theorem 3 of Ref. 1 that given any network $\nu_1 \in C_N$ there is another network $\nu_2 \in C_N$ that is rearrangeable and differs from ν_1 only in the fixed permutations that are used to connect the stages; in particular, ν_1 and ν_2 have the same number of crosspoints. Thus the problem of selecting an optimal rearrangeable network from C_N is equivalent to that of choosing an optimal network from C_N, rearrangeable or not. A network in C_N can be made rearrangeable by changing its link patterns at no increase in cost.

We make

Definition 1: $m = m(\nu) = [s(\nu) + 1]/2$ = numerical index of the middle stage
$n = n(\nu) = n_{m(\nu)}$ = size of middle stage switches

Definition 2: $O(\nu) = \{n_1, \cdots, n_{m-1}\}$ = the set of switch sizes (with repetitions) in *outer* (i.e., nonmiddle) stages

Definition 3: $\omega(N) = \{O(\nu) : \nu \in C_N\}$.

Remark 1: $c(\nu) = n(\nu) + 2 \sum_{x \epsilon O(\nu)} x.$

Theorem 1: Let (A,n) be a point (element) of

$$\omega(N) \times X$$

with

$$n \prod_{y \epsilon A} y = N.$$

Then there exists a nonempty set $Y \subseteq C_N$ such that

$$T(\nu) = (A,n), \qquad \nu \epsilon Y.$$

The ν's in Y differ only in the permutations between the stages and in the placing of the outer stages, and at least one of them is rearrangeable. This result follows from the definition of C_N and from Theorem 2 of Ref. 1.

V. CONSTRUCTION OF THE BASIC PARTIAL ORDERING

The solution to the problem of synthesizing an optimal rearrangeable network from C_N will be accomplished as follows: we shall define a mapping T of C_N into $\omega(N) \times X$, with $X = \{1, \cdots, N\}$, and a *partial ordering* $\leqq$ of $T(C_N)$; the map T will have the property that $c(\nu)$ is a function of $T(\nu)$; then we shall prove that (roughly speaking) a network ν is optimal if and only if $T(\nu)$ is at the "bottom" of the partial ordering, i.e., that $c(\nu)$ is almost an isotone function of $T(\nu)$.

To define a partial ordering of a finite set, it is enough to specify consistently which elements *cover* which others. Let $Z, Z_0, Z_1, \cdots$ be sets of positive integers $\leqq N$ possibly containing repetitions.

Definition 4: Z_1 covers Z_2 if and only if there are positive integers j and k such that k occurs in Z_1, j divides k, and Z_2 is obtained from Z_1 by replacing an occurrence of k with one occurrence each of j and k/j.

Definition 5: $Z_0 \leqq Z$ if and only if there is an integer n and sets Z_1, $Z_2, \cdots, Z_n$ such that Z_{i+1} covers Z_i, $i = 0, 1, \cdots, n-1$ and $Z_n = Z$.

Definition 6: T: $\nu \longrightarrow O(\nu), n(\nu)$.
A partial ordering $\leqq$ of $T(C_N)$ is defined by the following definition of covering:

Definition 7: Let μ, ν be elements of $C(N)$.
$T(\mu)$ covers $T(\nu)$ if and only if either

(*i*) $n(\nu) < n(\mu)$, $n(\nu)$ divides $n(\mu)$, and $O(\nu)$ results from $O(\mu)$ by adding an occurrence of $n(\mu)/n(\nu)$, or

(*ii*) $n(\nu) = n(\mu)$ and $O(\mu)$ covers $O(\nu)$.

VI. COST IS NEARLY ISOTONE ON $T(C_N)$

Theorem 2: If $T(\nu) \leqq T(\mu)$, and $n(\mu) > 6$, then

$$c(\nu) \leqq c(\mu).$$

Proof: It is enough to prove the result for μ and ν such that $T(\mu)$ covers $T(\nu)$.

Case (*i*)*:* $n(\nu) < n(\mu)$, $n(\nu)$ divides $n(\mu)$, $n(\nu) \geqq 2$, and $O(\nu)$ results from $O(\mu)$ by adding an occurrence of $n(\mu)/n(\nu)$. Then

$$\begin{aligned} c(\nu) &= n(\nu) + 2 \sum_{x \epsilon O(\nu)} x \\ &= n(\nu) + \frac{2n(\mu)}{n(\nu)} + 2 \sum_{x \epsilon O(\mu)} x \\ &= c(\mu) + n\ (\nu) - n(\mu) + \frac{2n(\mu)}{n(\nu)} \\ &= c(\mu) + n(\nu) \left[1 - \frac{n(\mu)}{n(\nu)} \right] + \frac{2n(\mu)}{n(\nu)} . \end{aligned}$$

Thus $c(\nu) \leqq c(\mu)$ if and only if

$$n(\nu) \left(1 - \frac{n(\mu)}{n(\nu)} \right) + \frac{2n(\mu)}{n(\nu)} \leqq 0$$

that is, if

$$\frac{2y}{y-1} \leqq x$$

where $x = n(\nu)$ and $y = n(\mu)/n(\nu)$. Now $n(\mu) > 6$ implies that either

(*i*) $n(\nu) = 2$ and $\dfrac{n(\mu)}{n(\nu)} \geqq 4$

or

(*ii*) $n(\nu) = 3$ and $\dfrac{n(\mu)}{n(\nu)} \geqq 3$

or

(*iii*) $n(\nu) \geqq 3$.

The condition $2y/(y - 1) \leqq x$ is fulfilled in all three cases, and so $c(\nu) \leqq c(\mu)$.

Case (*ii*): $n(\mu) = n(\nu)$ and $O(\mu)$ covers $O(\nu)$. There exist integers j,k such that j divides k, $j \geqq 2$ in $O(\mu)$, and $O(\nu)$ results from $O(\mu)$ by replacing one occurrence of k with one each of j and k/j. Then

$$\begin{aligned} c(\nu) &= n(\nu) + 2 \sum_{x \epsilon 0(\nu)} x \\ &= n(\mu) - 2k + 2j + (2k/j) + 2 \sum_{x \epsilon 0(\mu)} x \\ &= c(\mu) - 2k + 2j + (2k/j). \end{aligned}$$

Since j divides k and $j \geqq 2$, $k \geqq 2j$ and $k \geqq 2k/j$, so

$$k \geqq 2 \max\left(j, \frac{k}{j}\right) > j + \frac{k}{j}$$

and $c(\nu) < c(\mu)$.

Theorem 3: If $\nu \epsilon C_N$ and $O(\nu)$ does not consist entirely of prime numbers (possibly repeated), then there exists a network μ in C_N of $s(\nu) + 2$ stages with $c(\mu) < c(\nu)$, and ν cannot be optimal in C_N.

Proof: There is a value of k in the range $1 \leqq k \leqq n(\nu) - 1$ for which n_k is not a prime, say $n_k = ab$. Define stages $\mathcal{S}_j(\mu)$, $j = 1, \cdots, s(\nu) + 2$ as follows:

$$\mathcal{S}_{j+1}(\mu) = \mathcal{S}_j(\nu), \qquad j = k + 1, \cdots, n(\nu);$$

let Π_a, Π_b be partitions of $X = \{1, \cdots, N\}$ with

$$\begin{aligned} |\,\Pi_a\,| = N/a \quad &\text{and} \quad A \,\epsilon\, \Pi_a \Rightarrow |\,A\,| = a \\ |\,\Pi_b\,| = N/b \quad &\text{and} \quad B \,\epsilon\, \Pi_b \Rightarrow |\,B\,| = b. \end{aligned}$$

Set

$$\begin{aligned} \mathcal{S}_{k+1}(\mu) &= \bigcup_{A \epsilon \Pi_a} A^2 \\ \mathcal{S}_k(\mu) &= \bigcup_{B \epsilon \Pi_b} B^2 \\ \mathcal{S}_j(\mu) &= \mathcal{S}_j(\nu) & j &= 1, \cdots, k - 1 \\ \mathcal{S}_j(\mu) &= \mathcal{S}_{s(\nu)-j+1}(\mu) & \text{all } j &= 1, \cdots, s(\nu) + 2 \end{aligned}$$

By Theorem 2 of Ref. 1 permutations $\varphi_1, \cdots, \varphi_{s(\nu)-1}$ can be found so that the network

$$\mu = \mathcal{S}_1\varphi_1, \cdots, \varphi_{s(\mu)-1}\mathcal{S}_{s(\mu)}$$

is in C_N and is rearrangeable. It is apparent that $n_{m(\mu)} = n_{m(\nu)}$ and that $O(\nu)$ covers $O(\mu)$. Hence the argument for case (*ii*) of Theorem 2 shows that μ has strictly lower cost than ν.

Corollary 1: If $N > 6$ and is not prime, then a network ν consisting of one square switch is not optimal.

VII. PRINCIPAL RESULTS

Definition 8: An element $T(\nu)$ of $T(C_N)$ is *ultimate* if there are no $\mu \in C_N$ such that $T(\nu)$ covers $T(\mu)$.

Remark 2: $T(\nu)$ is ultimate if and only if $n(\nu)$ is prime and $O(\nu)$ consists entirely of prime numbers.

Definition 9: An element $T(\nu)$ of $T(C_N)$ is *penultimate* if it covers an ultimate element.

Definition 10: p_n , $n = 1, 2, \cdots$, is the nth prime.

Definition 11: $\pi(n)$ is the prime decomposition of n, that is, the set of numbers (with repetitions) such that

$$n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_l^{\alpha_l}$$

if and only if $\pi(n)$ contains exactly α_i occurrences of p_i , $i = 1, \cdots ,$ l, and nothing else.

Definition 12: p is the largest prime factor of N.

Lemma 1: If $p = 3$ and $N > 6$ is even, then the following conditions are equivalent:

(*i*) *ν is optimal*
(*ii*) *$T(\nu)$ is penultimate and $n(\nu) = 6$ or 4*
(*iii*) *$T(\nu) = (\pi(N/)6,)6$ or $(\pi(N/4),4)$.*

Proof: By Theorems 2, 3 only ν with $n(\nu) \leqq 6$ and $O(\nu)$ consisting entirely of primes can be optimal. Writing $N = 2^x 3^y$ with $x \geqq 1$ and $y \geqq 1$, it is seen that such ν must have a cost $c(\nu)$ having one of the forms

$$2 + 2[2(x - 1) + 3y] = 4x + 6y - 2,$$

$$3 + 2[2x + 3(y - 1)] = 4x + 6y - 3,$$

$$4 + 2[2(x - 2) + 3y] = 4x + 6y - 4 \qquad \text{(only occurs if } x > 1\text{)},$$

$$6 + 2[2(x - 1) + 3(y - 1)] = 4x + 6y - 4.$$

The least of these is either of the last two, which correspond to $n(\nu) = 6$ if $x = 1$ or to $n(\nu) = 6$ or 4 if $x > 1$. It is apparent that *(ii)* is equivalent to *(iii)*.

Lemma 2: If $p = 2$, and $N > 4$, then the following conditions are equivalent:

(i) ν is optimal
(ii) $T(\nu)$ is penultimate and $n(\nu) = 4$
(iii) $T(\nu) = (\pi(N/4), 4)$.

Proof: With $N = 2^x$ it can be seen as in Lemma 1 that only those ν can be optimal whose cost $c(\nu)$ has one of the forms

$$2 + 2[2(x - 1)],$$

$$4 + 2[2(x - 2)].$$

The second of these is the better, and corresponds to $n(\nu) = 4$.

Theorem 4: Let μ be a network such that a prime number $r > n(\mu)$ occurs in $O(\mu)$. Let M result from $O(\mu)$ by replacing one occurrence of r by $n(\mu)$. Then for any network ν with

$$T(\nu) = (M,r)$$

it is true that

$$c(\nu) < c(\mu)$$

i.e., ν is strictly better than μ. Among such ν, that is best for which r is largest.

Proof: Existence of a rearrangeable ν satisfying $T(\nu) = (M,r)$ is guaranteed by Theorem 1. For the rest of the proof, we observe that $r > n(\mu)$ and

$$\begin{aligned} c(\mu) &= n(\mu) + 2 \sum_{x \in O(\mu)} x \\ &= n(\mu) + 2r - 2n(\mu) + 2 \sum_{x \in M} x \\ &= r - n(\mu) + c(\nu). \end{aligned}$$

Theorem 5: If $n(\mu) \leqq 6$, $n(\mu) = 2^x 3^y 5^z$, some prime number $r > 3$ occurs in $0(\mu)$, and if M results from $0(\mu)$ by replacing one occurrence of r by x occurrences of 2, y occurrences of 3, and z occurrences of 5 then for any network $\nu \in C_N$ with

$$T(\nu) = (M,r)$$

it is true that

$$c(\nu) \leqq c(\mu)$$

i.e., ν is at least as good as μ. Among such ν, that is best for which r is largest.

Proof: Existence of a rearrangeable $\nu \in C_N$ satisfying $T(\nu) = (M,r)$ is given by Theorem 1. For the rest of the proof, we observe that $r \geqq 5$ and

$$\begin{aligned} c(\mu) &= n(\mu) + 2 \sum_{u \in O(\mu)} u \\ &= n(\mu) + 2r - 4x - 6y - 10z + 2 \sum_{u \in M} u \\ &= r + n(\mu) - 4x - 6y - 10z + c(\nu). \end{aligned}$$

Since x, y, and z can only assume the values 0 and 1, with $z = 1$ if and only if $x = y = 0$, we have $c(\mu) \geqq c(\nu)$, the best ν corresponding to the largest r.

Definition 13: $Q = \{(A,r) : r \text{ a prime and } A = \pi(N/r)\}$.

Definition 14: $L = T^{-1}(Q)$.

Remark 2: Q consists of all the absolute minima in the partial ordering $\leqq$ of $T(C_N)$, i.e., $\nu \in L$ implies that there are no $\mu \in C_N$ for which

$$T(\mu) < T(\nu).$$

Theorem 6: If $p > 3$, then all optimal networks belong to L.

Proof: Let $\mu \in C_N - L$ be given. We show that there exists a $\nu \in L$ that is at least as good.

Case 1: There is a sequence $\mu = \mu_1, \mu_2, \cdots, \mu_n, \nu$ with $\mu_n \neq \nu$, $\nu \in L$, $n(\mu_n) > 6$,

$$T(\mu_1) \geqq T(\mu_2) \geqq \cdots \geqq T(\mu_n)$$

and such that $T(\mu_n)$ covers $T(\nu)$. Then the numbers $n(\mu_j)$, $j = 1, \cdots, n$ are all > 6, and the result follows from Theorem 2.

Case 2: All sequences $\mu = \mu_1, \mu_2, \cdots, \mu_n, \nu$ with $\mu_n \neq \nu$, $\nu \in L$, $T(\mu_1) \geqq T(\mu_2) \geqq \cdots \geqq T(\mu_n)$, and such that $T(\mu_n)$ covers $T(\nu)$, are such that $n(\mu_n) \leqq 6$. Consider such a sequence. Let i be the smallest index j for which $n(\mu_j) \leqq 6, j = 1, \cdots, n$. Then Theorem 2 gives $c(\mu) \geqq c(\mu_i)$. Since $n(\mu_i) \leqq 6$ and $T(\mu_n)$ covers $T(\nu)$, it follows that $O(\mu_i)$ contains an occurrence of $p > 3$. Hence by Theorem 5 there exists a network $\eta \in C_N$ with $n(\eta) = p$ and

$$c(\eta) \leqq c(\mu_i) \leqq c(\mu).$$

Let $\xi \in L$ be such that $n(\xi) = p$ and $T(\eta)$ covers $T(\xi)$. Then $c(\xi) \leqq c(\eta)$ by case (ii) of Theorem 2. Hence

$$c(\xi) \leqq c(\mu)$$

$$\xi \in L.$$

Theorem 7: If $N \leqq 6$ *and* ν *is optimal, then* ν *is a square switch and* $c(\nu) = N$.

Proof: For prime N with $2 \leqq N < 6$ the result is obvious. If $N = 6$ and $\nu \in C_6$ then exactly one of the following alternatives obtains:

$$\begin{aligned} T(\nu) &= (\theta, 6) && \text{and} \quad c(\nu) = 6 \\ T(\nu) &= (\{3\}, 2) && \text{and} \quad c(\nu) = 8 \\ T(\nu) &= (\{2\}, 3) && \text{and} \quad c(\nu) = 7. \end{aligned}$$

The first alternative is optimal, and there is exactly one $\nu \in C_6$ such that $T(\nu) = (\theta, 6)$, viz., the 6 × 6 square switch. Similarly, if $N = 4$ and $\nu \in C_4$, then $T(\nu) = (\theta, 4)$ or $(\{2\}, 2)$; the former has cost 4, the latter 6.

Definition 15: For $n \geqq 2$, $D(n)$ is the sum of the prime divisors of n counted according to their multiplicity; thus if

$$n = 2^{\alpha_1} 3^{\alpha_2} \cdots p_k{}^{\alpha_k}$$

then

$$D(n) = \sum_{j=1}^{k} p_j \alpha_j = \sum_{x \in \pi(n)} x.$$

Definition 16: $c(N) = \min \{c(\nu) : \nu \in C_N\}$.

Theorem 8:

$$c(N) = \begin{cases} N \text{ if } N \leqq 6 & \textit{or } N \textit{ is prime} \\ p + 2D(N/p) & \textit{if } N > 6 \textit{ and either } p > 3 \textit{ or } N \textit{ is odd} \\ 2D(N/2) & \textit{if } N > 6 \textit{ in all other cases (i.e., } p = 2, \textit{ or} \\ & \quad p = 3 \textit{ and } N \textit{ is even).} \end{cases}$$

Proof: Putting together Lemmas 1, 2 and Theorems 1, 2, 3, 4, 6, and 7 we obtain the following values for the minimal cost in crosspoints per terminal on a side for networks in C_N:

$$c(N) = \begin{cases} N \text{ if } N \leqq 6 \text{ or } N \text{ is prime} & \\ p + 2 \sum\limits_{x \epsilon \pi (N/p)} & \textit{if } p > 3, N > 6 \\ 6 + 2 \sum\limits_{x \epsilon \pi (N/6)} x = 2 \sum\limits_{x \epsilon \pi (N/2)} x & \textit{if } p = 3, N > 6, N \textit{ even} \\ 3 + 2 \sum\limits_{x \epsilon \pi (N/3)} x = 3 + 6(\log_3 N - 1) & \\ & \textit{if } p = 3, N > 6, N \textit{ odd} \\ 4 + 2 \sum\limits_{x \epsilon \pi (N/4)} x = 4(\log_2 N - 2) = 2 \sum\limits_{x \epsilon \pi (/2)} x & \\ & \textit{if } p = 2, N > 6; \end{cases}$$

simplification gives Theorem 8.

REFERENCES

1. Beneš, V. E., Permutation Groups, Complexes, and Rearrangeable Connecting Networks, B.S.T.J., this issue, p. 1619.
2. Shannon, C. E., Memory Requirements in a Telephone Exchange, B.S.T.J., **29,** July, 1950, pp. 343–349.

Attitude Determination and Prediction of Spin-Stabilized Satellites

By L. C. THOMAS and J. O. CAPPELLARI

(Manuscript received February 17, 1964)

Techniques for both attitude determination and prediction for spin-stabilized satellites are developed. Their use is demonstrated using Telstar I and II satellite data. It is shown that an inclined dipole model of the earth's magnetic field and the method of averaging the gravitational and magnetic torques over each anomalistic period of the satellite permits attitude predictions to within a few tenths of a degree of determined values in most instances. In those few cases where departures are above one degree, explanations are presented to show the reason for such discrepancies.

The usefulness of combining optical flash and solar sensor data for attitude determination and their inherent accuracy are demonstrated. Optical flash data can provide loci with a resolution of 0.1°. Solar sensor loci are resolved to within 1°.

All techniques described have been consolidated into working computer programs which follow closely the mathematical analysis presented. A number of important supporting calculations such as the solar position, sidereal time, orbit updating, etc. are also developed. Because of the complexities of the mean torque and gyroscopic equations, the precessional techniques presented are most useful in computer embodiments.

Table of Contents

I. INTRODUCTION

To maintain a defined attitude in space, the Telstar I and II satellites were spin stabilized. By this method of passive attitude control, a satellite is rotated about an axis of symmetry and consequently exhibits the characteristics of a gyroscope. In the absence of disturbing torques, the satellite's spin axis maintains its spatial orientation fixed with respect to an inertial reference frame throughout its orbit. For the Telstar I and II satellites, this is desirable because of certain required attitude constraints. First of all, the satellite communication antenna is not omnidirectional. More energy is radiated along the equator of the satellite than along its spin poles, as shown by the antenna pattern of Fig. 1. This fact dictates an attitude for which the line of sight from

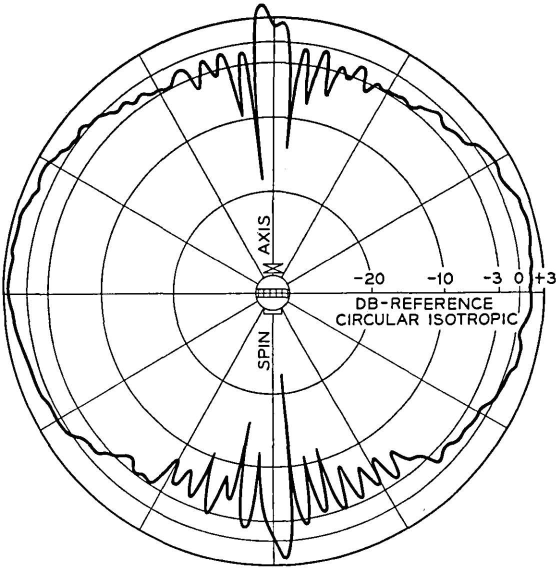


Fig. 1 — Antenna pattern at 4170 mc.

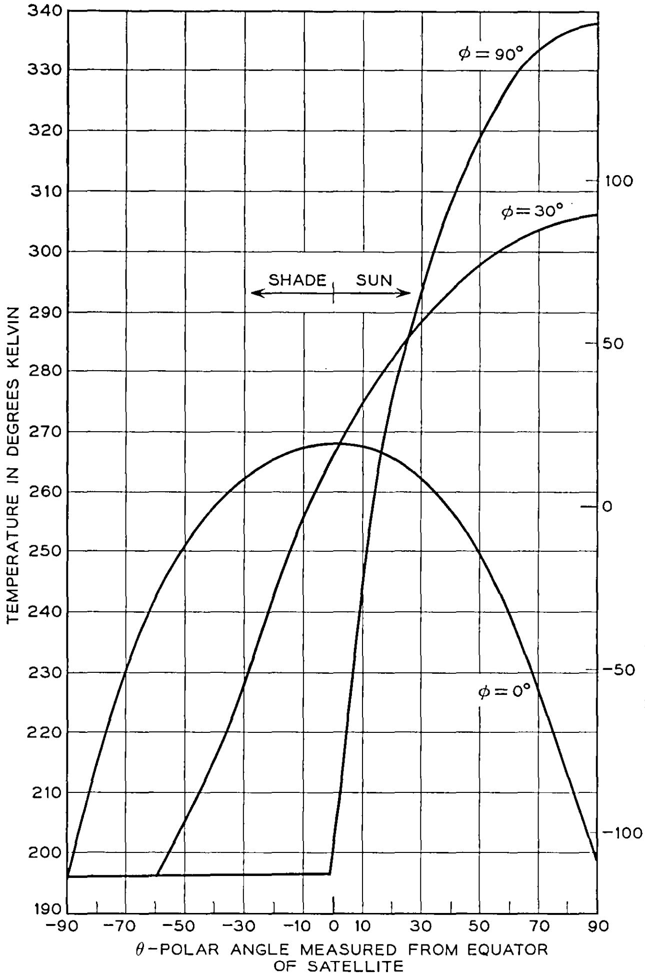


Fig. 2 — Skin temperature distribution vs polar angle.

a ground station to the satellite avoids its poles. The second attitude constraint involves temperature considerations. A good degree of temperature control is obtained by orienting the spin axis so that it is nearly perpendicular to the satellite-sun line. In this manner, temperature balance is maintained by the satellite's spin as indicated in Fig. 2. Here, φ is the solar offset angle, defined as the angular departure of the spin axis from perpendicularity with the satellite-sun line. Solar offsets of about 15° result in temperature deviations of about 150° (see Fig. 2) and are tolerable.

An ideal orientation from a communications standpoint would be to have the spin axis nearly parallel to the earth's surface as it passes over any ground station. A spin axis perpendicular to the orbital plane would accomplish this, but would produce a maximum axis tilt toward the sun equal to the sum of the orbital inclination and the earth's $23\frac{1}{2}$° tilt with respect to the plane of the ecliptic (see Fig. 3). Under these conditions the spin axis of the first Telstar satellite would have a maximum solar tilt of 68°. This exceeds the 15° tilt limit dictated by temperature balance. On the other hand, if the spin axis is made perpendicular to the ecliptic plane (which automatically insures its perpendicularity to the sun), the axis tilt with respect to the line of sight from any ground station over which it may be passing will range from 90° to 90° minus the sum of the orbit inclination and ecliptic obliquity or 22° (see Fig. 4).* From a communications standpoint this is tolerable, since nulls in the antenna pattern are major only within about 15° of the spin

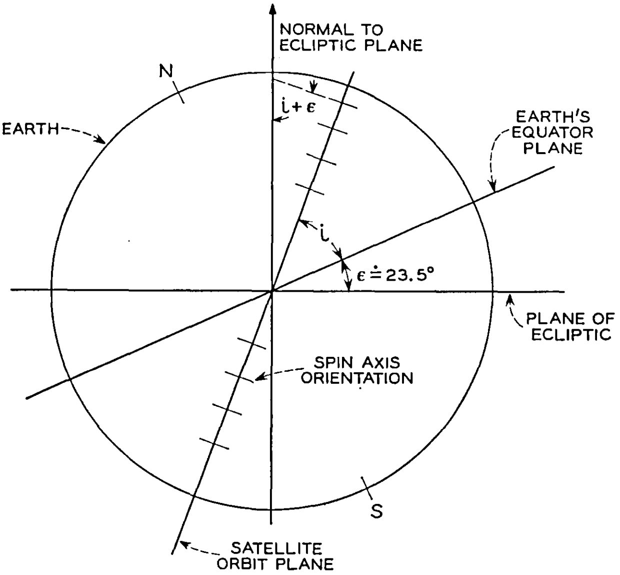


Fig. 3 — Geometry of spin axis perpendicular to orbit plane.

* With this orientation, stations south of the satellite's instantaneous earth latitude would experience angles less than 22°, such as station B of Fig. 4. The major ground stations for Telstar I and II, however, are all at latitudes above 43°. Since the inclinations of the Telstar I and II satellites are 45° and 42.7° respectively, these stations are almost always north of the satellite.

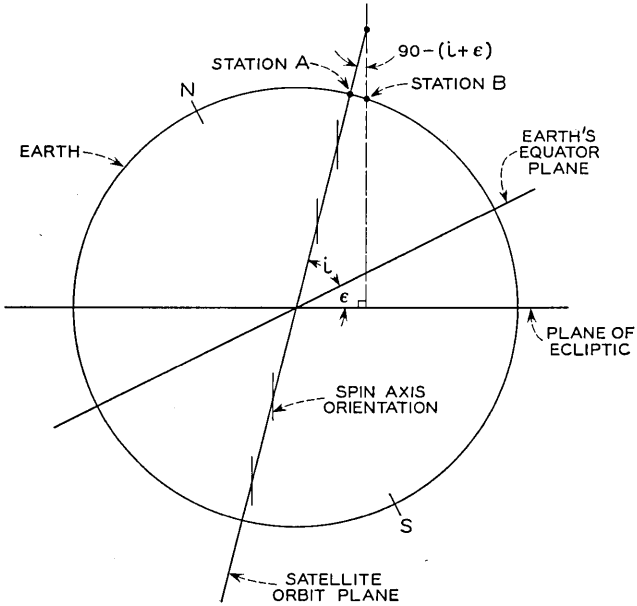


Fig. 4 — Geometry of spin axis perpendicular to ecliptic.

poles (see Fig. 1). Rising and setting satellites may experience a greater tilt to the ground station line of sight, but this is unavoidable.

Because of these considerations, the Telstar I and II satellites were launched with their spin axes as nearly perpendicular to the ecliptic as the powered flight trajectory of the Thor-Delta launch vehicle would permit and still meet certain orbital requirements, such as inclination, apogee height, and perigee height.[1] The predicted attitudes of these satellites at orbit injection are calculated from telemetry data from the first and second rocket stages. Table I lists these attitudes as the right ascension and declination direction of the north pole of the satellite.

Table I — Attitude of the Telstar I and II Satellites at Injection, Determined from Powered Flight Data

Satellite	Initial Attitude	
	Right Ascension	Declination
Telstar I	83.73°	−66.80°
Telstar II	82.23°	−57.31°

The north pole is defined here as the direction of advance of a right-handed screw turning in the direction of the satellite's spin. It also happens to be the spin pole which carries a helical telemetry antenna.

The initial attitudes of the Telstar I and II satellites are changed by gyroscopic precession. This motion of spin-stabilized satellites is chiefly produced by both magnetic and gravitational torques. The former is a result of interaction between residual and eddy-current-produced magnetic fields of the satellite and the earth's magnetic field. The latter is produced by differential gravitational forces acting across the body of the satellite. In the present cases, the magnetic torques are several orders of magnitude in excess of the gravitational torques.

It is necessary to predict precession to allow proper scheduling and planning for satellite use, to resolve certain attitude determination ambiguities, and to sensibly plan alteration of satellite attitude in a prescribed manner when needed. (See Section V for a description of this technique.) It is the purpose of this paper to outline the methods of attitude determination and correction, to develop the precessional theory, and to show the application of these to the Telstar I and II satellites.

II. ATTITUDE DETERMINATION — GENERAL REMARKS

The attitude of the Telstar I satellite has been determined through the analysis of two sets of data: the time of optical flashes of sunlight from three mirrors attached to the surface of Telstar and the current produced by six on-board solar sensors located on the ends of three orthogonal axes.

The first of these sets of data in combination with the spatial position of the satellite, sun, and observer's position determines a locus of possible spin axis positions which would result in the observed flash. This locus describes a cone in space whose axis is the mirror normal. The six solar sensors are designed to determine the direction of the satellite-sun line with respect to a satellite frame of reference and thus to the spin axis. Knowing this angle, again there is defined a conical array of possible spatial spin axis directions. The intersection of the optical cone with this solar sensor cone should, therefore, determine two possible attitudes. A priori knowledge of the approximate attitude as provided by launch data and/or a succession of measurements and predictions over an interval of several days permits the determination of a unique spin axis direction in space.

In practice, the solar sensors present a few difficulties. First of all, the deduction of the sun's position from solar cell current entails inferring a solar angle between each cell's normal and the sun's direction. Such a direction from each of three cells uniquely establishes this solar aspect.* Thus calibration curves relating electrical output to light intensity must be employed for each cell along with temperature corrections. While D. W. Hill[2] has prepared a computer program to take the drudgery out of this work, it remains difficult to calculate the solar angle for a cell if it is illuminated in a direction far from its normal. Moreover, the solar sensor data, which are reported every minute by telemetry, were found to be not always mutually consistent. Often, the solar direction calculated over 30 minutes from a succession of telemetry frames had a spread ranging from 1° to over 8°. A correlation between these deviations and the spatial position of the satellite exists which suggests a biasing of solar cell data, on occasion, by the earth's reflectivity.

To show the geometry of this situation, consider the satellite position shown in Fig. 5. Here one half of the satellite to the right of line AB is illuminated by the sun. At least one half, to the upper left of line CD, is illuminated by the earth in reflecting about 38 per cent† of the sun's total incident light. Thus a solar cell in region BOD would record only direct sunlight, one in region AOC would record only earthshine, and

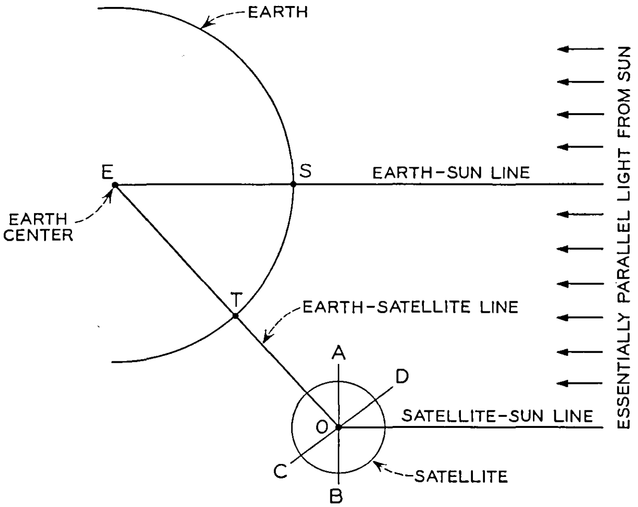


Fig. 5 — Satellite shadow geometry.

* The solar aspect is the angle between the satellite's spin axis and the satellite-sun line.

† This is not the albedo but the earth's mean reflectivity or ratio of mean earth brightness measured at a spot along the earth-sun line relative to the brightness of a perfectly diffusing disk of the same size and at the same distance that the

one in region AOD would record both. There would be no light from either the sun or the earth in region COB. By the difference in solar illumination and earthshine, data from cells in regions BOD and AOC are easily separated and the latter disregarded. However, those cells in region AOD measure both sunlight and earthshine in an inseparable manner and therefore report erroneously the sun's position. If AOD is a large angle, the spin of the satellite will carry solar sensors into that region frequently and result in sizable variations in successive solar aspect determinations. Since AOD equals SET, as the subsatellite position (T) departs from the subsolar point (S), the solar cell data must be carefully interpreted to avoid false conclusions. Under these conditions, it would be wrong, for example, to simply calculate the mean of all solar aspect determinations over a succession of telemetry frames. This procedure, in general, would not yield a good estimate of the true aspect. A better technique would be to calculate the mean for only those solar aspect determinations in which all three cells entering into the determination exhibited current readings above those which could possibly be produced by either the earth's reflectivity or low angles of solar illumination. Operating in this fashion, the true solar aspect can be determined to within about 1°.

Since the maximum attitude resolution obtainable using optical flashes from the Telstar satellites is about 0.1°, it was decided to rely on these for attitude determinations insofar as possible. Two groups of flashes close together in time are needed for an attitude fix, however, and in cases where only a single one existed, the attitude was determined by a combination of mirror flashes and the solar sensor data previously described.

The optical reflections are characterized by a series of intermittent flashes provided by the spin of the Telstar satellites. The time midpoint of these flash series is determined by photoelectric equipment[3] at Bell Telephone Laboratories in Holmdel, N. J. The time of each flash series determines a conical locus of possible spin axis positions about the flashing mirror's normal. The tip of the spin axis vector, therefore, lies on a circle on the celestial sphere. One of the intersections of two such circles defines the attitude, provided the two corresponding flash series occur close enough together in time so that no appreciable precession occurs during the separation interval. Since the mirrors employed have their normals far removed from the spin axis (68° and 95°), it is

earth is from the given spot. Quite naturally a phase law applies which reduces the light reflected to points off the earth-sun line. Off-line brightness varies crudely as the ratio of observable illuminated area of the earth as seen from the point in question to the total observable illuminated area seen from a point on the earth-sun line.

always possible to select the valid locus intersection from past performance and attitude predictions.

III. ATTITUDE DETERMINATION — COMPUTATIONAL TECHNIQUE

A computational technique has been developed for the calculation of spin loci. It consists of four basic parts, which are:

(*i*) the determination of the right ascension and declination of the normal of the flashing mirror;

(*ii*) the determination of the right ascension and declination of the sun in the case of solar cell data;

(*iii*) the construction of a circular locus on the celestial sphere centered on the right ascension and declination of the mirror normal and having a radius of 68° or 95°, depending upon the mirror involved (in the case of solar aspect data from the sensors, the locus is centered on the right ascension and declination of the sun and has a radius equal to the measured solar aspect angle);

(*iv*) the plotting, by computer microfilm techniques, of these circles and others determined from additional flash series and solar aspect data to ascertain intersections and corresponding satellite attitudes.

3.1 *The Right Ascension and Declination of the Mirror Normal*

Consider the generalized mirror orientation shown in Fig. 6. Here, a ray from the sun strikes the mirror and is reflected toward a particular

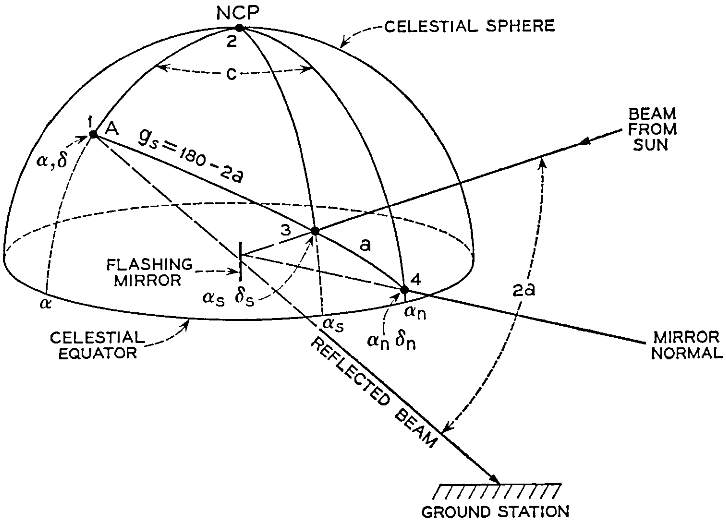


Fig. 6 — Geometry of mirror normal.

ground station. For this to be possible, the mirror normal must bisect the angle $2a$ and lie in the plane determined by the mirror-sun line and the mirror-ground station line.

Sketched around the mirror in Fig. 6 is the celestial sphere. On this sphere the solar direction, as seen from the mirror, is indicated by its right ascension, α_s, and its declination, δ_s. In like fashion, the direction of the satellite mirror as seen from the ground station is specified as α,δ and the direction of the mirror normal is $\alpha_n\delta_n$.

To determine α_n and δ_n, we begin by solving for arc g_s and A in the spherical triangle 1, 2, 3:

$$\cos g_s = \sin \delta \sin \delta_s + \cos \delta \cos \delta_s \cos (\alpha_s - \alpha) \tag{1}$$

$$\cos A = \frac{\sin \delta_s - \sin \delta \cos g_s}{\cos \delta \sin g_s} \tag{2}$$

where angles g_s and A may have values from 0° to 180°.

By triangle 1, 2, 4, one obtains the declination of the normal as

$$\delta_n = \sin^{-1} [\sin \delta \cos g_s + \cos \delta \sin g_s \cos A] \tag{3}$$

where δ_n may have values between ±90°.

Using the same triangle we may write

$$C = \sin^{-1} \left[\frac{\sin (g_s + a) \sin \delta_n}{\sin A} \right] \tag{4}$$

where

$$A = \frac{180 - g_s}{2}. \tag{5}$$

C may have values between 0° and 180°. Since α_n may either exceed α by C or be less by the same amount owing to the two possible orientations of triangle 1, 2, 4, we have

$$\alpha_n = \alpha + \text{SIGNF } (C, \alpha_s - \alpha) \tag{6}$$

where SIGNF, a common computer symbol, indicates that the algebraic sign to be affixed to C shall be determined by the quantity $\alpha_s - \alpha$.

Right ascension is measured eastward from the vernal equinox as a 0 reference through 360°, and if α_s and α should lie on opposite sides in this reference (C remaining less than 180°), (6) becomes

$$\alpha_n = \alpha + \text{SIGNF } (C, \alpha - \alpha_s). \tag{7}$$

Thus (6) is employed if $| \alpha_s - \alpha | - 180°$ is negative and (7) if otherwise.

3.2 *The Right Ascension and Declination of the Sun*

The apparent right ascension and declination of the sun at any specified time is computed from the mean orbital elements of the sun.[4] These may be expressed beginning with the true solar mean anomaly in degrees as

$$M_s = 358.47583 + 0.9856002670d' - 0.000150T^2 - 0.000003T^3$$

where

T = the time in Julian centuries of 36525 ephemeris days from January 0.5, 1900, ephemeris time[5]
d' = ephemeris days from same epoch.

We may express the above equation in a more useful form for the present calculations by changing the epoch to January 1.0, 1960 and writing an equivalent expression as

$$M_s \doteq 357.41283 + 0.985600267d$$

where

d = ephemeris days since the 1960 epoch
= mean solar days since 1960 epoch + 1 second for the years 1963, 1964, 1965.

One obtains the apparent mean solar anomaly, used in the present calculations to determine the apparent position of the sun, by antedating for the solar light transit time. For this reason d is increased by 0.005375 day, which is the light transit time at mean solar distance. Since the earth is about 3×10^6 miles closer to the sun in winter as compared to summer, this can produce an error in apparent solar position of about 2 seconds of arc.

The eccentricity of the earth's orbit is

$$\begin{aligned} e &= 0.01675104 - 0.00004180T - 1.26 \times 10^{-7}T^2 \\ &= 0.01700254 \text{ on January 1.0, 1960.} \end{aligned}$$

Also, the degrees of mean celestial longitude of the perigee of the sun's mean orbit about the earth as a reference is

$$\begin{aligned} L &= 279.69668 + 0.9856473354d' + 0.000303T^2 \\ &= 282.25247 + 0.470684 \times 10^{-4}d. \end{aligned}$$

Finally, the mean obliquity of the ecliptic in degrees is

$$\epsilon = 23.452294 - 0.0130125T - 1.64 \times 10^{-6}T^2 + 5.03 \times 10^{-7}T^3.$$

The apparent mean anomaly and eccentricity along with Kepler's equation permit the calculations of the apparent true anomaly of the sun by standard techniques.[6] This along with the apparent mean longitude of solar perigee and mean obliquity of the ecliptic allows a direct determination of the apparent right ascension and declination of the sun at any date by the simple geometry shown in Fig. 7.

The apparent right ascension and declination of the sun computed from its mean elements in this manner can depart from the tabular values (which include nutation) of *The American Ephemeris and Nautical Almanac* for the years 1962, 1963, and 1964 by about 5 seconds of arc as a maximum. This is well within the approximate 1° error of the measuring techniques for solar aspect as well as the 0.1° for the mirror technique.

3.3 *Construction of the Locus Circle*

As previously stated, the direction of the spin axis may be specified as a circle on the celestial sphere, centered on the right ascension and declination of the mirror normal. This locus may be generated by using the spherical triangle shown in Fig. 8. Here the dotted curve indicates the attitude locus of all possible positions of the spin axis. Let $\alpha_l\delta_l$ be any point on this circle making a fixed arc F with α_n , δ_n . In the case of either the Telstar I or II satellite, F will equal 68° or 95°, the angles

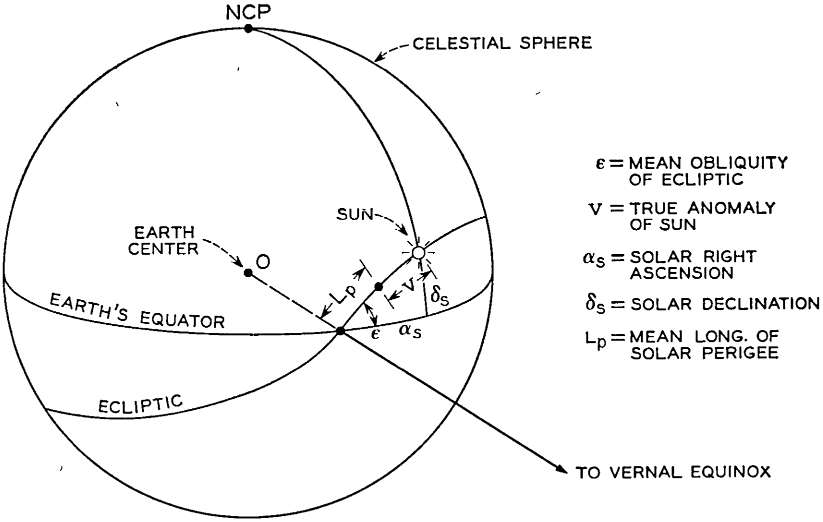

NCP
CELESTIAL SPHERE
ε = MEAN OBLIQUITY OF ECLIPTIC
V = TRUE ANOMALY OF SUN
α_S = SOLAR RIGHT ASCENSION
δ_S = SOLAR DECLINATION
L_P = MEAN LONG. OF SOLAR PERIGEE
SUN
EARTH CENTER
O
L_P
V
δ_S
EARTH'S EQUATOR
ε
α_S
ECLIPTIC
TO VERNAL EQUINOX

Fig. 7 — Geometry of the sun's position.

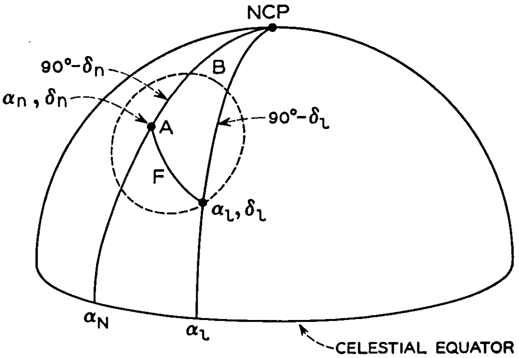


Fig. 8 — The attitude locus.

the mirror normals make with the spin axis. From Fig. 8,

$$\delta_l = \sin^{-1} [\sin \delta_n \cos F + \cos \delta_n \sin F \cos A] \tag{8}$$

where

A = a running variable which takes on values from 0 to 360° to generate the locus

δ_l = declination of a locus point which ranges from $-90°$ to $+90°$, in general.

Also

$$\sin B = \frac{\sin F \sin A}{\cos \delta_l} \tag{9}$$

$$\cos B = \frac{(\cos F - \sin \delta_n) \sin \delta_l}{\cos \delta_n \cos \delta_l} \tag{9a}$$

$$B = \tan^{-1} \frac{\sin B}{\cos B} \tag{10}$$

and

$$\alpha_l = \alpha_n + B. \tag{10a}$$

In general, B may range from 0° to 360°. All quadrant ambiguities presented by (10) are resolved by noting the algebraic signs of (9) and (9a).

If there is a time uncertainty in measuring the midpoint of a flash series, this will result in more than one possible mirror normal, $\alpha_n\delta_n$, and hence a number of attitude loci, since the satellite and sun will occupy successive positions along their paths within the time error.

If during a single sky trajectory a ground station records two separate flash series (or a single series plus solar aspect data), each will generate a circular locus. The intersections of these loci determine attitude with an ambiguity of two. That is, there are two intersections and therefore two possible satellite attitudes which satisfy the two flash series. A priori knowledge of the attitude from previous measurements or initially from launch parameters together with some knowledge of the expected precession will permit selecting the proper intersection. Knowledge of expected precession is also needed to determine whether or not the flash series are close enough together in time to neglect incremental precession during the time between the series. Such precession can alter the position of the loci intersections.

Fig. 9 shows an example of the attitude determination of the Telstar II satellite for passes 135, 136, 199, 272, and 472. Solid lines indicate attitude loci determined from mirror flashes. The line of zeros is typical of those determined by solar sensor data. Two attitude determinations are shown by intersections on passes 135, 136, and on pass 272.

IV. ATTITUDE PREDICTION

Acting upon an orbiting spin-stabilized satellite to produce precession are certain disturbing torques. Those to be considered here are gravitational and magnetic torques. In the case of the Telstar I and II satellites, these are dominant over atmospheric drag torques, solar radiation torques, electrostatic torques, and others.

4.1 *Assumptions for Gravity Torque Calculations*

Differential gravity forces acting across the body of the satellite can produce torques which tend to rotate the body. These forces exist simply because the strength of the earth's gravity field is a function of the distance from earth. A body of finite size must, therefore, experience such torques.

For the purpose of calculating the mean gravity torques, the earth is assumed spherical with its radius equal to the equatorial radius. For the Telstar satellites, gravity torques are at least an order of magnitude less than magnetic torques, and therefore neglecting the earth's oblateness produces at worst only a second-order error.

It is also assumed that the moments of inertia of the satellites about all axes perpendicular to the spin axis and passing through the mass center are equal. The orbit is assumed elliptical and Keplerian, since the earth's figure produces but second-order effects over a single satel-

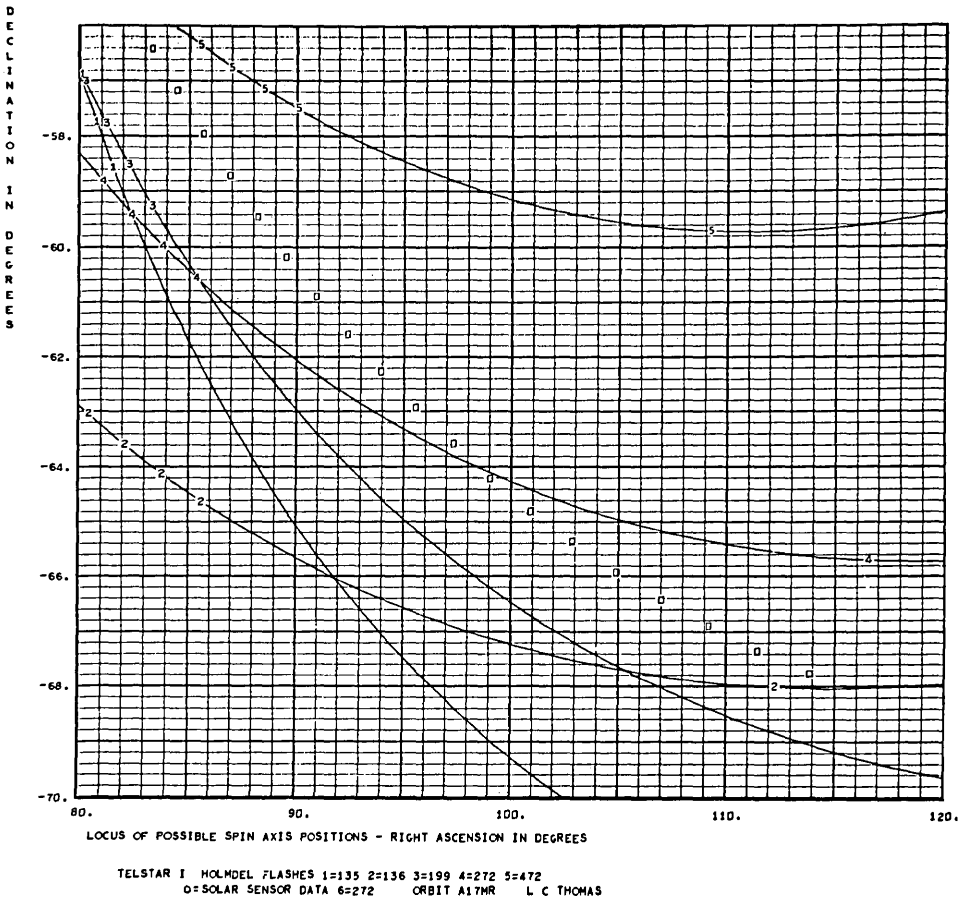


Fig. 9 — Computer plot of mirror and solar loci.

lite period. The orbit, however, for each succession of calculations will be updated using mean orbital elements, which include all secular perturbations, to produce the Keplerian orbit of the best fit at the beginning of each calculation.

4.2 *Assumptions for Magnetic Torque Calculations*

In order to construct a mathematical model which on one hand reasonably well represents the physics of the situation, but on the other hand does not, by its complexity, produce numerical equations costly to compute, the following simplifying assumptions are made:

(*i*) The magnetic field of the earth is represented by a dipole centered at the earth's center and inclined β degrees (equal 11.4°) to the earth's spin axis.[7,8]

(*ii*) Mass symmetry about the satellite's spin axis exists, as in Section 4.1.

(*iii*) The effects of magnetic moments transverse to the spin axis are either negligible or average out due to the satellite's spin.

(*iv*) For the calculation of mean or net magnetic torques over an orbital period, a Keplerian orbit is assumed as in Section 4.1.

4.3 *Assumptions for Gyroscopic Equations of Motion*

It will be useful in simplifying the gyroscopic equations of motion (see Section 4.7) to assume that the satellite angular momentum vector coincides with the spin axis and equals the spin rate times the moment of inertia about the spin axis. This is the same as assuming the entire angular momentum of the satellite is the result of its spin alone, and neglects that small amount provided by the precession of the spin axis itself. Precessional dampers[9] are provided on the Telstar satellites to prevent coning or rapid changes in attitude at rates comparable to the spin rate. This, more than ever, makes the assumption quite reasonable.

4.4 *Coordinate Systems*

It will be convenient to establish certain useful coordinate systems and their interrelationships prior to the torque calculations. These will be defined and related by Euler-type axis rotations expressed by matrices. All coordinate systems are rectangular and right-handed in the conventional sense such that the rotation of an X axis into a Y axis determines the positive direction of a Z axis as the direction of progress of a right-handed screw. Each system will be named, described, and interrelated in that order.

4.4.1 *The Earth-Centered Inertial System* (*IS*)

In this system, the three mutually perpendicular X, Y, and Z axes have their common origin at the earth's center. X contains the vernal equinox and increases positively from the earth's center in its direction. Y is perpendicular to X in the earth's equatorial plane. Z contains the earth's spin axis and increases positively toward the north celestial pole. This is the basic system to which others will be referred.

4.4.2 *Rigid-Body Systems*

Two rigid-body systems will be used. Both have their origin at the satellite mass center. The first of these is to be known as the SANOR or satellite nonrotating system. In this system, the z axis defines the satellite's spin axis, being positive in the direction of advance of a right-hand screw spinning with the satellite. The x axis is in the satellite equatorial plane, ψ degrees rotated from X (see Fig. 10) about Z. This axis defines the xy-XY plane intersection. The y axis is in the satellite equatorial plane orthogonal to x. At this point, we note that the rotational equations of motion for the satellite will be developed in this SANOR system.

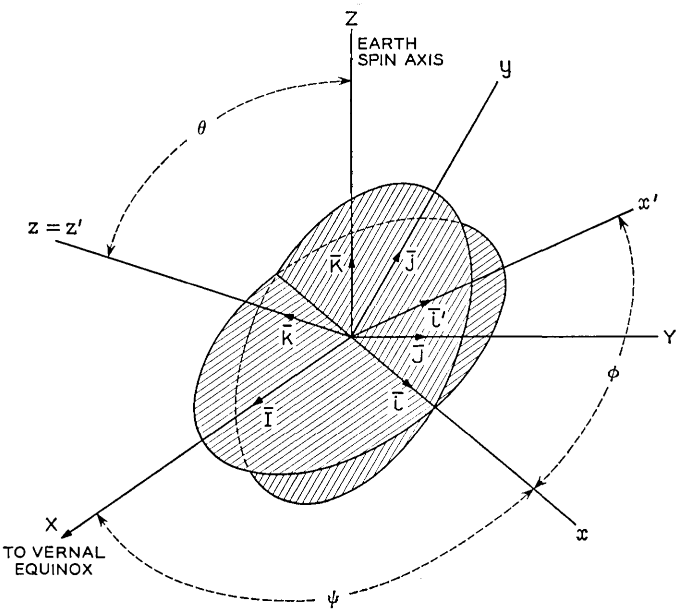


Fig. 10 — The IS, SANOR and SAR systems.

The SAR or satellite rotating system, as its name implies, differs from the SANOR system in that it rotates with the satellite. It has axes x', y', z', where z' coincides with z, and x' and y' are defined as being rotated φ degrees from x and y respectively about z.

To relate the SANOR and SAR systems to IS, we proceed as follows (refer to Fig. 10). The ψ rotation about Z yields the following matrix

$$D = \begin{vmatrix} \cos\psi & \sin\psi & 0 \\ -\sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{vmatrix}. \tag{11}$$

The θ rotation about x yields

$$C = \begin{vmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{vmatrix}. \tag{12}$$

The φ rotation about z gives

$$B = \begin{vmatrix} \cos\varphi & \sin\varphi & 0 \\ -\sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{vmatrix}. \tag{13}$$

To transfer points in the IS to points in the SANOR system, we have

$$\begin{vmatrix} x \\ y \\ z \end{vmatrix} = (CD) \begin{vmatrix} X \\ Y \\ Z \end{vmatrix} \tag{14}$$

or, just to shorten the notation

$$x = (CD)X \tag{15}$$

where

$$(CD) = \text{multiplication of the } C \text{ and } D \text{ matrices.}$$

In like fashion

$$\mathrm{a}' = (BCD)X \tag{16}$$

$$= AX \tag{17}$$

where

$$A \equiv (BCD) = \text{multiplication of the } B,\ C,\ D \text{ matrices.}$$

Carrying out this multiplication, we may express A as the following

$$A = \begin{vmatrix} \cos\psi\cos\varphi - \cos\theta\sin\varphi\sin\psi & \cos\varphi\sin\psi + \cos\theta\cos\psi\sin\varphi & \sin\varphi\sin\theta \\ -\sin\varphi\cos\psi - \cos\theta\sin\psi\cos\varphi & -\sin\varphi\sin\psi + \cos\theta\cos\psi\cos\varphi & \cos\varphi\sin\theta \\ \sin\theta\sin\psi & -\sin\theta\cos\psi & \cos\theta \end{vmatrix}. \tag{18}$$

The matrix CD may be obtained from A by letting $\varphi = 0$,

$$(CD) = \begin{vmatrix} \cos\psi & \sin\psi & 0 \\ -\sin\psi\cos\theta & \cos\psi\cos\theta & \sin\theta \\ \sin\psi\sin\theta & -\cos\psi\sin\theta & \cos\theta \end{vmatrix}. \tag{19}$$

Quite obviously the inverse operations apply by taking the transposes of the matrices. Indicating a transposed matrix by the symbol ~, we have

$$X = (\tilde{D}\tilde{C})x \tag{20}$$

and

$$X = \tilde{A}x'. \tag{21}$$

4.4.3 *Orbital Coordinate Systems*

Let an orbit defining system (ORDEF) be described along the lines shown in Fig. 11. Here x_g is the intersection of the orbit plane and the earth's equator plane with $+x_g$ drawn toward the ascending node of the orbit from the earth's center. Axis z_g is normal to the orbit plane, positive in a direction a right-handed screw along oz_g would advance if turned in the direction of the satellite's orbital motion. Axis y_g completes a right-handed system by being mutually perpendicular to x_g and z_g.

The satellite defining system (SADEF) has its z_s axis collinear with z_g, but x_s passes through the instantaneous satellite position, ω degrees from x_g. Axis x_s therefore will be referred to as the local vertical of the satellite.

The ORDEF and SADEF systems are related to the inertial system (IS) in a manner strictly analogous to the SANOR and the SAR matrices described in the previous section. The only change in the matrices is that

$$\Omega \text{ replaces } \psi$$

$$i \text{ replaces } \theta$$

$$\omega \text{ replaces } \varphi.$$

Thus we may write

$$x_g = (CD)_i X \tag{22}$$

$$X = (\tilde{D}\tilde{C})_i x_g \tag{23}$$

$$x_s = A_i X \tag{24}$$

$$X = \tilde{A}_i x_s \tag{25}$$

where the subscript i indicates the Ω, i, ω substitution.

4.4.4 *Magnetic Coordinate System*

A magnetic coordinate system (MAG) based on an inclined dipole model of the earth's magnetic field is constructed as follows. Let axis x_m define the intersection of the geomagnetic equatorial plane with the earth's geographical equatorial plane, positive toward the ascending node of the geomagnetic equator (η° from X, see Fig. 12). Axis z_m is

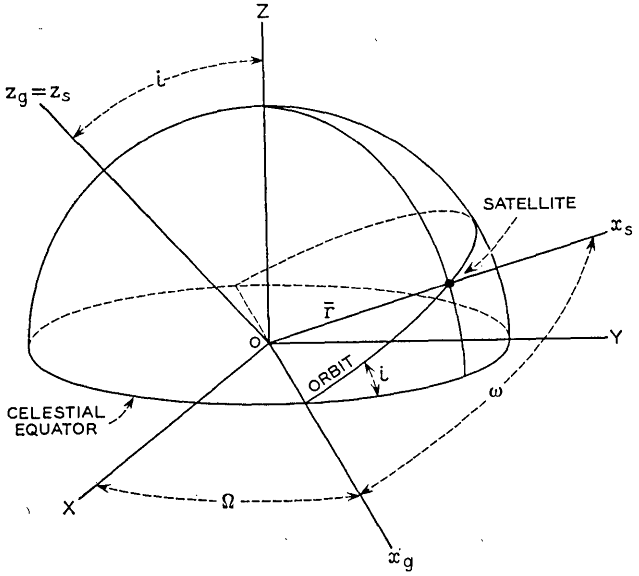


Fig. 11 — The ORDEF and SADEF coordinate systems.

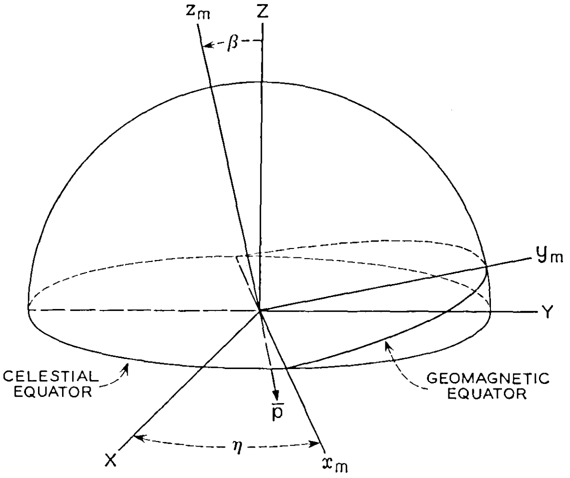


Fig. 12 — The geomagnetic coordinate system.

normal to the geomagnetic equator plane, positive toward the north geographical hemisphere, and $\beta°$ from Z. Here β is simply the magnetic dipole inclination to the earth's spin axis. Axis y_m is in the geomagnetic equatorial plane orthogonal to x_m.

Transformations from the MAG system into the IS system proceed as in (15) thru (20) with

$$\eta \text{ replacing } \psi$$

$$\beta \text{ replacing } \theta$$

in the matrices, so that

$$x_m = (CD)_\eta X \tag{26}$$

and

$$X = (\tilde{D}\tilde{C})_\eta x_m . \tag{27}$$

where the subscript η on the matrices indicates the above substitutions.

Fig. 13 shows the orientation of the earth-centered inclined dipole which produces a field of best fit[7,8] to a field based on all observations of field vectors made anywhere over the earth's surface. The anomalies, or differences between the actual and dipole field, decrease more rapidly with increasing heights above the earth's surface than does the dipole field itself, making the fit better and better as altitude increases.

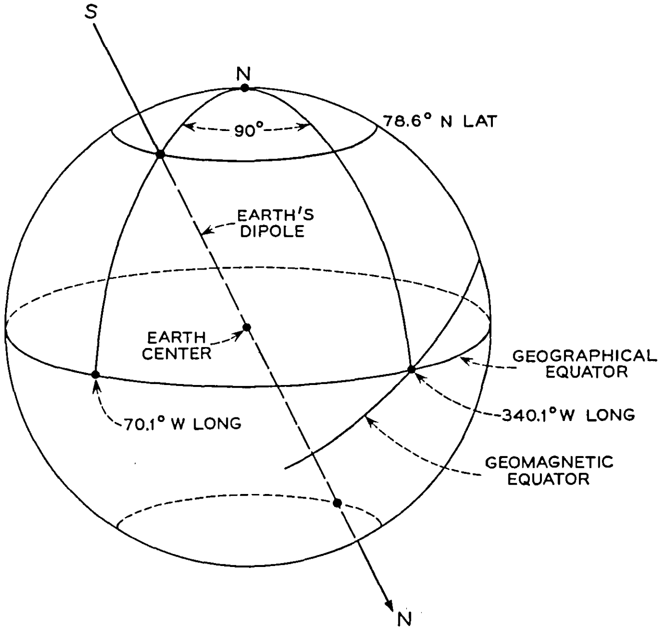


Fig. 13 — The earth-centered inclined dipole.

4.5 *The Mean Gravity Torque*

If the potential energy of an orbiting satellite is expanded about its center of mass in a Taylor series and differentiated with respect to the angles giving its orientation with respect to the IS coordinate system, the instantaneous gravitational torque acting on the satellite may be expressed as[10]

$$\mathbf{T}_g = 3\mu^2(I_3 - I)(\mathbf{i}_s\cdot\mathbf{k})(\mathbf{i}_s \times \mathbf{k}) \tag{28}$$

where

$$\mu^2 = GM/r^3 \tag{29}$$

GM = universal gravitational constant × the mass of the earth
r = geocentric distance of satellite
I_3 = moment of inertia about satellite's spin axis, z
I = moment of inertia transverse to z (assuming all transverse moments to be equal, i.e., $I_x = I_y$)
$\mathbf{k}$ = unit vector along satellite spin axis, z (more generally the

vector along the axis of cylindrical mass symmetry, which is the spin axis by assumptions in Section 4.1).

To determine the mean gravity torque, $\mathbf{T}_g$ will be integrated over a nodal period (T_n) assuming no precession during that period; i.e., the SANOR system stays fixed with respect to IS during the integration. Then

$$(\mathbf{T}_g)_{\text{mean}} = \frac{1}{T_n}\int_0^{T_n} 3\mu^2(I_3 - I_1)(\mathbf{i}_s\cdot\mathbf{k})(\mathbf{i}_s \times \mathbf{k})\,dt. \tag{30}$$

By interchanging the order of the dot product and taking $\mathbf{k}$ constant,

$$(\mathbf{T}_g)_{\text{mean}} = \frac{3GM(\Delta I)}{T_n}\left[\mathbf{k}\cdot\left\{\int_0^{T_n} \frac{\mathbf{i}_s\mathbf{i}_s}{r^3}\,dt\right\} \times \mathbf{k}\right]. \tag{31}$$

If a pure Keplerian orbit[6] is assumed, we have

$$r = \frac{a(1 - e^2)}{1 + e\cos(\omega - P)} \tag{32}$$

$$r^2\dot{\omega} = \frac{2\pi a^2(1 - e^2)^{\frac{1}{2}}}{T_n} \tag{33}$$

$$GM = \frac{4\pi^2 a^3}{T_n^2} \tag{34}$$

where

r = radius vector from the focus of the elliptical orbit to the satellite
a = semimajor axis of ellipse
e = eccentricity of the ellipse
P = argument of perigee
ω = argument of the satellite.

Hence the integral becomes

$$\int_0^{T_n} \frac{\mathbf{i}_s\mathbf{i}_s}{r^3}\,dt = \int_0^{\omega=2\pi} \frac{\mathbf{i}_s\mathbf{i}_s}{r^3\dot{\omega}}\,d\omega \tag{35}$$

and

$$\begin{aligned}(\mathbf{T}_g)_{\text{mean}} &= \frac{6\pi(\Delta I)}{T_n^2(1 - e^2)^{\frac{3}{2}}}\\ &\times\left[\mathbf{k}\cdot\int_0^{2\pi} \mathbf{i}_s\mathbf{i}_s(1 + e\cos\omega\cos P + e\sin\omega\sin P)\,d\omega \times \mathbf{k}\right].\end{aligned} \tag{36}$$

Since in the ORDEF coordinates (see Fig. 11)

$$\mathbf{i}_s = \cos \omega \mathbf{i}_g + \sin \omega \mathbf{j}_g \tag{37}$$

the dyad, $\mathbf{i}_s\mathbf{i}_s$ becomes

$$\mathbf{i}_s\mathbf{i}_s = \cos^2 \omega\, \mathbf{i}_g\mathbf{i}_g + \sin^2 \omega\, \mathbf{j}_g\mathbf{j}_g + \sin \omega \cos \omega(\mathbf{i}_g\mathbf{j}_g + \mathbf{j}_g\mathbf{i}_g). \tag{38}$$

Substituting this into (36) results in three integrals. Those containing $\mathbf{i}_g\mathbf{i}_g$ and $\mathbf{j}_g\mathbf{j}_g$ yield π as a result of the integration. The other integral is 0. Therefore,

$$(T_g)_{\text{mean}} = \frac{6\pi^2(\Delta I)}{T_n^2(1-e^2)^{\frac{3}{2}}}[\mathbf{k}\cdot(\mathbf{i}_g\mathbf{i}_g + \mathbf{j}_g\mathbf{j}_g) \times \mathbf{k}]. \tag{39}$$

By some vector maneuvering (see Appendix A) this reduces to

$$(T_g)_{\text{mean}} = -\frac{6\pi^2(\Delta I)}{T_n^2(1-e^2)^{\frac{3}{2}}}[(\mathbf{k}\cdot\mathbf{k}_g)(\mathbf{k}_g \times \mathbf{k})]. \tag{40}$$

To transform the ORDEF vectors into the SANOR system, we write

$$x_g = (CD)_i X \tag{22}$$

$$X = (\tilde{D}\tilde{C})x \tag{20}$$

$$\therefore\ x_g = (CD)(\tilde{D}\tilde{C})x. \tag{41}$$

Performing the indicated operations, after a bit of labor we find that

$$\begin{aligned}\mathbf{k}_g = {} & (\sin \Omega \sin i \cos \psi - \cos \Omega \sin i \sin \psi)\mathbf{i} \\ & + (-\sin \Omega \sin i \sin \psi \cos \theta - \cos \Omega \sin i \cos \psi \cos \theta + \cos i \sin \theta)\mathbf{j} \\ & + (\sin \Omega \sin i \sin \psi \sin \theta + \cos \Omega \sin i \cos \psi \sin \theta + \cos i \cos \theta)\mathbf{k}\end{aligned} \tag{42}$$

where, referring to Figs. 10 and 11, we see that

$$\begin{aligned}\psi &= \text{an Euler rotational angle} \\ i &= \text{orbital inclination} \\ \Omega &= \text{ascending node of orbit.}\end{aligned}$$

Substituting into (40), the mean gravity torque reduces to

$$\begin{aligned}(\mathbf{T}_g)_{\text{mean}} = {} & \frac{6\pi^2\Delta I}{T_n^2(1-e^2)^{\frac{3}{2}}}\{\cos i \cos \theta + \sin i \sin \theta \cos(\Omega - \psi)\} \\ & \cdot [\![(\sin i \cos \theta \cos(\Omega - \psi) - \cos i \sin \theta)\mathbf{i} + (\sin i \sin(\Omega - \psi))\mathbf{j}]\!].\end{aligned} \tag{43}$$

4.6 *The Mean Magnetic Torque*

It is well known[11] that the scalar potential (Φ_m) of a magnetic dipole may be expressed as

$$\Phi_m = \frac{1}{4\pi\mu_0}\frac{\mathbf{p}\cdot\mathbf{i}_s}{r^2}$$

$$= \frac{1}{4\pi\mu_0}\frac{\mathbf{p}\cdot\mathbf{r}}{r^3} \tag{44}$$

$$= -\frac{p}{4\pi\mu_0}\frac{\mathbf{k}_m\cdot\mathbf{r}}{r^3} \tag{45}$$

where

μ_0 = permeability of free space
= $4\pi \times 10^{-7}$ webers/ampere-meter (Ref. 12)
= $4\pi \times 10^{-7}$ henry/meter (Ref. 12)
$\mathbf{p}$ = magnetic moment of earth's field, direction and magnitude
= 10^{17} weber-meters
= $10^{17}(10^{10}/4\pi)$ emu = 8.06×10^{25} emu
= 8.06×10^{25} erg/gauss (Ref. 12, p. 25)
$\mathbf{i}_s$ = unit vector along the x_s axis, local satellite vertical
r = distance from dipole center (earth's center) to satellite
$\mathbf{k}_m$ = unit vector describing direction of geomagnetic moment
p = magnitude of $\mathbf{p}$
$\mathbf{r}$ = vector form of r along i_s axis.

A satellite magnetic moment (**M**) in the earth's magnetic field (**H**) will produce a torque

$$\mathbf{T} = \mathbf{M} \times \mathbf{H}. \tag{46}$$

If the satellite spins rapidly, any magnetic moment components perpendicular to the spin axis will tend to produce torques which average to 0, while the component along the spin axis will produce a net torque of

$$\mathbf{T}_m = \mathbf{M}_s \times \mathbf{H} \tag{47}$$

$$= M_s\mathbf{k} \times \mathbf{H} \text{ ergs} \tag{48}$$

where

M_s = satellite magnetic moment along its spin axis (weber-meters or ergs/gauss)
$\mathbf{k}$ = unit vector along satellite spin axis (SANOR system)
$\mathbf{H}$ = geomagnetic field (ampere-turns/meters).

It is desired to integrate the instantaneous torque, $\mathbf{T}_m$, over one anomalistic period in order to calculate the mean torque. To do this, we shall express **H** in the MAG system and eventually convert this expression into SANOR terms to be comparable with M_s.

We begin by setting

$$\mathbf{H} = -\nabla\Phi_m \tag{49}$$

from (45) and (49),

$$\mathbf{H} = -\frac{h}{\mu_0}\nabla\left[\frac{-\mathbf{k}_m\cdot(x_m\mathbf{i}_m + y_m\mathbf{j}_m + z_m\mathbf{k}_m)}{r^3}\right] \tag{50}$$

where

$$x_m, y_m, z_m = \text{components of } \mathbf{r} \text{ in MAG system}$$

and

$$h = p/4\pi. \tag{51}$$

Therefore,

$$\mathbf{H} = \frac{h}{\mu_0}\nabla\,(z_m/r^3) \tag{52}$$

and since

$$r^3 = (x_m^2 + y_m^2 + z_m^2)^{\frac{3}{2}}$$

we have

$$\mathbf{H} = \frac{-h}{\mu_0 r^5}[3x_m z_m\mathbf{i}_m + 3y_m z_m\mathbf{j}_m + (3z_m^2 - r^2)\mathbf{k}_m]. \tag{53}$$

Rewriting to spotlight **r** components, using the following relationships normalized to r,

$$x_m/r = r_x, \qquad y_m/r = r_y, \qquad z_m/r = r_z \tag{54}$$

one obtains

$$\mathbf{H} = -\frac{h}{\mu_0 r^3}[3r_x r_z\mathbf{i}_m + 3r_y r_z\mathbf{j}_m + (3r_z^2 - 1)\mathbf{k}_m] \tag{55}$$

or

$$\mathbf{H} = \frac{p}{4\pi\mu_0 r^3}[\mathbf{k}_m - 3(\mathbf{i}_s\cdot\mathbf{k}_m)\mathbf{i}_s] \tag{56}$$

since

$$\mathbf{r} = r\mathbf{i}_s. \tag{57}$$

By the following equations [(24) through (65)], the variables $\mathbf{i}_s$ and $\mathbf{k}_m$ will be expressed in forms useful in the ensuing mean magnetic torque calculations.

Let us first express $\mathbf{i}_s$ in the MAG system. As shown previously,

$$x_s = A_i X \tag{24}$$

and

$$X = (\tilde{D}\tilde{C})_\eta X_m \tag{27}$$

therefore

$$x_s = A_i(\tilde{D}\tilde{C})_\eta X_m \tag{58}$$

or

$$\begin{aligned} \mathbf{i}_s = {} & |\,(\cos\Omega\cos\omega - \cos i \sin\omega \sin\Omega) \\ & \cdot(\cos\omega\sin\Omega + \cos i \cos\Omega \sin\omega)(\sin\omega\sin i)\,| \\ & \times \begin{vmatrix} \cos\eta & -\sin\eta\cos\beta & \sin\eta\sin\beta \\ \sin\eta & \cos\eta\cos\beta & -\cos\eta\sin\beta \\ 0 & \sin\beta & \cos\beta \end{vmatrix} \begin{vmatrix} \mathbf{i}_m \\ \mathbf{j}_m \\ \mathbf{k}_m \end{vmatrix} . \end{aligned} \tag{59}$$

Expanding,

$$\begin{aligned} \mathbf{i}_s = {} & [\cos\eta(\cos\Omega\cos\omega - \cos i \sin\omega\sin\Omega) + \sin\eta(\cos\omega\sin\Omega \\ & + \cos i\cos\Omega\sin\omega)]\mathbf{i}_m + [-\sin\eta\cos\beta(\cos\Omega\cos\omega \\ & - \cos i \sin\omega\sin\Omega) + \cos\eta\cos\beta(\cos\omega\sin\Omega \\ & + \cos i\cos\Omega\sin\omega) + \sin\beta\sin\omega\sin i]\mathbf{j}_m + [\sin\eta\sin\beta \\ & (\cos\Omega\cos\omega - \cos i\sin\omega\sin\Omega) - \cos\eta\sin\beta(\cos\omega\sin\Omega \\ & + \cos i\cos\Omega\sin\omega) + \cos\beta\sin\omega\sin i]\mathbf{k}_m . \end{aligned} \tag{60}$$

This will be used later.

Also, $\mathbf{i}_s$ in the IS system may be written [see (24)] as

$$\mathbf{i}_s = A_i\mathbf{I} \tag{61}$$

$$\begin{aligned} \mathbf{i}_s = {} & [\cos\Omega\cos\omega - \cos i \sin\omega\sin\Omega]\mathbf{I} \\ & + [\cos\omega\sin\Omega + \cos i\cos\Omega\sin\omega]\mathbf{J} \\ & + [\sin\omega\sin i]\mathbf{K}. \end{aligned} \tag{62}$$

In a similar manner, $\mathbf{k}_m$ in the IS system is expanded as follows:

$$x_m = (CD)_\eta X \tag{63}$$

or

$$\begin{aligned} \mathbf{k}_m &= \sin\eta \sin\beta\, \mathbf{I} - \cos\eta \sin\beta\, \mathbf{J} + \cos\beta\, \mathbf{K} && (64) \\ &= \sin\beta(\sin\eta\, \mathbf{I} - \cos\eta\, \mathbf{J}) + \cos\beta\, \mathbf{K}. && (65) \end{aligned}$$

The problem is now to integrate the instantaneous torque, $\mathbf{T}_m$, over one anomalistic period, T_a, to obtain the mean magnetic torque,

$$\begin{aligned} (\mathbf{T}_m)_{\text{mean}} &= \frac{1}{T_a}\int_{t_i}^{t_i+T_a} \mathbf{T}_m dt = \frac{1}{T_a}\int_{v_i}^{v_i+2\pi} \mathbf{T}_m \frac{dt}{dv}\, dv \\ &= \frac{1}{T_a}\int_{v_i}^{v_i+2\pi} \mathbf{T}_m \frac{r^2}{h}\, dv \end{aligned} \tag{66}$$

where

T_a = anomalistic period of the satellite
$\mathbf{T}_m$ = instantaneous magnetic torque
t = time
v = true anomaly of the satellite
r = geocentric satellite distance
h = defined by (68) below.

Substituting into (66) from (48) and (56)

$$(\mathbf{T}_m)_{\text{mean}} = \frac{M_s p}{4\pi\mu_0 T_a}\, \mathbf{k} \times \int_{v_i}^{v_i+2\pi} \frac{\mathbf{k}_m - 3(\mathbf{i}_s\cdot\mathbf{k}_m)\mathbf{i}_s}{r^3}\left(\frac{r^2}{h}\right) dv. \tag{67}$$

Since

$$h = \frac{2\pi a^2(1 - e^2)^{\frac{1}{2}}}{T_a} \qquad (68)^{13}$$

and

$$r = \frac{a(1 - e^2)}{1 + e\cos v} \qquad (69)^{13}$$

where

a = semimajor axis of orbit
e = orbital eccentricity
T_a = anomalistic period of the satellite

(67) becomes

$$(\mathbf{T}_m)_{\text{mean}} = \frac{M_s p}{8\pi^2 \mu_0 a^3 (1 - e^2)^{\frac{3}{2}}} \cdot \mathbf{k} \times \int_{v_i}^{v_i+2\pi} [\mathbf{k}_m - 3(\mathbf{i}_s \cdot \mathbf{k}_m)\mathbf{i}_s](1 + e \cos v)\,dv \tag{70}$$

$$= \mathfrak{M}\mathbf{k} \times \int_{v_i}^{v_i+2\pi} [\mathbf{k}_m - 3(\mathbf{i}_s \cdot \mathbf{k}_m)\mathbf{i}_s](1 + e \cos v)\,dv. \tag{71}$$

First we shall evaluate the integral of (71) beginning by substituting the $\mathbf{k}_m$ value given in (65), to obtain

$$\frac{(\mathbf{T}_m)_{\text{mean}}}{\mathfrak{M}} = (\sin \beta) k \times \int_{v_i}^{v_i+2\pi} [\sin \eta \mathbf{I} - \cos \eta \mathbf{J} - 3\{\mathbf{i}_s \cdot (\sin \eta \mathbf{I} - \cos \eta \mathbf{J})\}\mathbf{i}_s](1 + e \cos v)\,dv \tag{72}$$

$$+ (\cos \beta)\mathbf{k} \times \int_{v_i}^{v_i+2\pi} \{\mathbf{K} - 3(\mathbf{i}_s \cdot \mathbf{K})\mathbf{i}_s\}(1 + e \cos v)\,dv$$

$$= \mathrm{I} + \mathrm{II}. \tag{73}$$

Integral II, the simpler of the two, is evaluated in Appendix B. The result is given below:

$$\mathrm{II} = \pi \cos \beta [\{-2 \sin \theta + 3 \sin^2 i \sin \theta + 3 \sin i \cos i \cos \theta \cos (\Omega - \psi)\}\mathbf{i} + 3 \sin i \cos i \sin (\Omega - \psi)\mathbf{j}]. \tag{88}$$

Now we must evaluate the first integral, I, of (73) but first let us state that by examining (72) it is perfectly obvious that the I integral does not exist for a noninclined dipole. For this case, it follows from (73) with $\cos \beta = 1$ that

$$(\mathbf{T}_m)_{\text{mean}} = \frac{M_s p}{8\pi^2 \mu_0 a^3 (1 - e^2)^{\frac{3}{2}}}\ \mathrm{II}. \tag{89}$$

By suitable variable substitution the first integral I of (73) may be compressed to yield

$$\mathrm{I} = (\sin \beta)\mathbf{k} \times (\mathbf{I}A - \mathbf{J}B - 3\mathbf{I} \cdot \mathbf{C} + 3\mathbf{J} \cdot \mathbf{E}) \tag{90}$$

where

$$A = \int_{v_i}^{v_i+2\pi} \sin \eta (1 + e \cos v)\,dv \tag{91}$$

$$B = \int_{v_i}^{v_i+2\pi} \cos \eta (1 + e \cos v) dv \tag{92}$$

$$\mathbf{C} = \int_{v_i}^{v_i+2\pi} \mathbf{i}_s \mathbf{i}_s \sin \eta (1 + e \cos v) dv \tag{93}$$

$$\mathbf{E} = \int_{v_i}^{v_i+2\pi} \mathbf{i}_s \mathbf{i}_s \cos \eta (1 + e \cos v) dv. \tag{94}$$

Now η must be expressed as a function of v in order to evaluate A, B, $\mathbf{C}$, and $\mathbf{E}$. For simplification of the integrals we shall also let the initial time for the integration be the time when the satellite passes through perigee. Then

$$\eta = \eta_0 + \omega_E t \tag{95}$$

$$= \eta_0 + \omega_E T_a M / 2\pi \tag{96}$$

$$= \eta_0 + bM \tag{97}$$

where

η_0 = initial position of the ascending node of the geomagnetic equator at perigee passage with respect to the IS system (see Fig. 12)
ω_E = angular velocity of the earth in the IS system
t = time measured from passage of satellite through perigee
M = mean anomaly of the satellite (radians)
T_a = anomalistic period of the satellite (time units)
b = $(\omega_E / 2\pi) T_a$ = number of turns of earth in time T_a.

Therefore

$$\sin \eta = \sin \eta_0 \cos (bM) + \cos \eta_0 \sin (bM) \tag{98}$$

$$\cos \eta = \cos \eta_0 \cos (bM) - \sin \eta_0 \sin (bM). \tag{99}$$

Unfortunately, M is related to v through Kepler's equation as

$$M = E - e \sin E \qquad (100)^{14}$$

where

E = the eccentric anomaly
e = eccentricity of the orbit

and

$$E = 2 \tan^{-1} \left[\left(\frac{1 - e}{1 + e} \right)^{\frac{1}{2}} \tan \frac{v}{2} \right] = 2 \tan^{-1} \left(q \tan \frac{v}{2} \right). \qquad (101)^{14}$$

Equations (100) and (101) certainly define M as a function of v, but in a most complicated manner. It can be shown, however, that a plot of v and M versus time normalized to T_a will look like Fig. 14. So as a reasonable approximation to M we might consider

$$M = v - (\lambda/b) \sin v \tag{102}$$

where

λ/b = maximum amplitude of the true anomaly "sine" wave of Fig. 14 = $2e$.[15]

Note that (102) resembles Kepler's equation with v replacing E. With less sophistication we might even let

$$M = v. \tag{103}$$

The only justification here is that we shall be dealing in mean torques averaged over an orbital period, and the v function makes one oscilla-

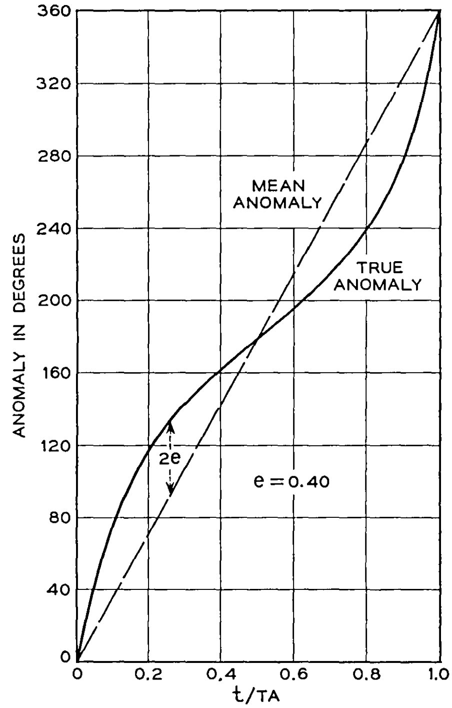


Fig. 14 — Mean and true anomaly comparison.

tion about the M function in that time. If (103) is assumed, (98) and (99) remain unaltered except that v replaces M. If (102) is used, (98) and (99) become

$$\begin{aligned}\sin \eta &= \sin \eta_0[\cos bv \cos (\lambda \sin v) + \sin bv \sin (\lambda \sin v)] + \\ &\qquad \cos \eta_0[\sin bv \cos (\lambda \sin v) - \cos bv \sin (\lambda \sin v)] \qquad (104)\\ &= \sin \eta_0[A'] + \cos \eta_0[B'] \qquad (105).\end{aligned}$$

where A' and B' are defined by comparing (105) to (104). In like fashion,

$$\cos \eta = \cos \eta_0[A'] - \sin \eta_0[B']. \qquad (106)$$

We may express the sin (λ sin v) and cos (λ sin v) portions of A' and B' as

$$\sin (\lambda \sin v) \doteq \lambda \sin v - \frac{\lambda^3 \sin^3 v}{6} \qquad (107)$$

and

$$\cos (\lambda \sin v) \doteq 1 - \frac{\lambda^2 \sin^2 v}{2}. \qquad (108)$$

These approximations are reasonably valid for near earth satellites not having high eccentricities and avoid Bessel function complications. The neglect of higher-order terms in (107) and (108) for $e = 0.25$ and $T_a \doteq 150$ minutes results in errors of less than one part in 10^6.

As case I we shall evaluate the integrals A, B, **C**, and **E** using (98) and (99) with v replacing M. For case II we shall return to evaluate these integrals again using (105), (106), (107), and (108). Case I details are given in Appendix C. Case II is outlined by Appendix D.

Using the expanded II integral of (88) and the evaluated A, B, **C**, and **E** integrals from Appendices C and D, (73) may now be written as

$$\begin{aligned}(\mathbf{T}_m)_{\text{mean}} = \mathfrak{M}\Big[&\cos \beta \mathbf{k} \times \int_0^{2\pi} \{\mathbf{k} - 3(\mathbf{i}_s \cdot \mathbf{K})\mathbf{i}_s\}(1 + e \cos v)dv \\ &+ \sin \beta \mathbf{k} \times \{A\mathbf{I} - B\mathbf{J} - 3\mathbf{I} \cdot (\mathbf{i}_g\mathbf{i}_g C_1 + \mathbf{j}_g\mathbf{j}_g C_2 \\ &+ (\mathbf{i}_g\mathbf{j}_g + \mathbf{j}_g\mathbf{i}_g)C_3) + 3\mathbf{J} \cdot (\mathbf{i}_g\mathbf{i}_g E_1 + \mathbf{j}_g\mathbf{j}_g E_2 \qquad (130)\\ &+ (\mathbf{i}_g\mathbf{j}_g + \mathbf{j}_g\mathbf{i}_g)E_3)\}\Big]\end{aligned}$$

where

A, B, C_1, C_2, C_3, E_1, E_2, E_3 are all defined in Appendix C.

The following, expressed in the SANOR system, will now be substituted into (130) (refer to Fig. 11)

$$\mathbf{k} \times \mathbf{I} = \cos\psi\mathbf{j} + \sin\psi\cos\theta\mathbf{i} \tag{131}$$

$$\mathbf{k} \times (-\mathbf{J}) = \cos\psi\cos\theta\mathbf{i} - \sin\psi\mathbf{j} \tag{132}$$

$$\mathbf{J}\cdot\mathbf{i}_g = \sin\Omega \tag{133}$$

$$\mathbf{J}\cdot\mathbf{j}_g = \cos\Omega\cos i \tag{134}$$

$$\mathbf{I}\cdot\mathbf{i}_g = \cos\Omega \tag{135}$$

$$\mathbf{I}\cdot\mathbf{j}_g = -\sin\Omega\cos i \tag{136}$$

$$\mathbf{k} \times \mathbf{i}_g = -\cos\theta\sin(\Omega - \psi)\mathbf{i} + \cos(\Omega - \psi)\mathbf{j} \tag{137}$$

$$\begin{aligned}\mathbf{k} \times \mathbf{j}_g = (-\sin i\sin\theta - \cos i\cos\theta\cos(\Omega - \psi))\mathbf{i}&\\ - \cos i\sin(\Omega - \psi)\mathbf{j}&\end{aligned} \tag{138}$$

$$\mathbf{k} \times \mathbf{K} = -\sin\theta\mathbf{i}. \tag{139}$$

These yield the mean magnetic torque, which is

$$\begin{aligned}(\mathbf{T}_M)_{\text{mean}} = {}& \frac{M_s p}{8\pi^2\mu_0 a^3(1 - e^2)^{\frac{3}{2}}}\Big\{\pi\cos\beta[\mathbf{i}\{-2\sin\theta + 3\sin^2 i\sin\theta\\ &+ 3\sin i\cos i\cos\theta\cos(\Omega - \psi)\}\\ &+ \mathbf{j}\{3\sin i\cos i\sin(\Omega - \psi\}]\Big\}\\ &+ \frac{M_s p\sin\beta}{8\pi^2\mu_0 a^3(1 - e^2)^{\frac{3}{2}}}\Big\{A[\sin\psi\cos\theta\mathbf{i} + \cos\psi\mathbf{j}]\\ &+ B[\cos\psi\cos\theta\mathbf{i} - \sin\psi\mathbf{j}]\\ &+ 3[\{E_1\sin\Omega + \mathrm{E}_3\cos\Omega\cos i - C_1\cos\Omega\\ &+ C_3\sin\Omega\cos i\}\{-\cos\theta\sin(\Omega - \psi)\mathbf{i}\\ &+ \cos(\Omega - \psi)\mathbf{j}\} + \{E_2\cos\Omega\cos i + E_3\sin\Omega\\ &+ C_2\sin\Omega\cos i - C_3\cos\Omega\}\{-(\sin i\sin\theta\\ &+ \cos i\cos\theta\cos(\Omega - \psi))\mathbf{i} - \cos i\sin(\Omega - \psi)\mathbf{j}\}]\Big\}\end{aligned} \tag{140}$$

or, collecting on $\mathfrak{M}\cos\beta$ and $\mathfrak{M}\sin\beta$,

$$\begin{aligned}(\mathbf{T}_M)_{\text{mean}} = \pi\mathfrak{M}\cos\beta\Big[\mathbf{i}\{&-2\sin\theta + 3\sin^2 i\sin\theta + 3\sin i\cos i\\ &\cdot\cos\theta\cos(\Omega - \psi)\} + \mathbf{j}\{3\sin i\cos i\sin(\Omega - \psi)\}\Big]\end{aligned} \tag{141}$$

$$
\begin{aligned}
+ \mathcal{M} \sin \beta \Big[\mathbf{i} \Big\{ & A \sin \psi \cos \theta + B \cos \psi \cos \theta \\
& - 3[\cos \theta \sin (\Omega - \psi)(E_1 \sin \Omega + E_3 \cos \Omega \cos i \\
& - C_1 \cos \Omega + C_3 \sin \Omega \cos i) + \{\sin i \sin \theta \\
& + \cos i \cos \theta \cos (\Omega - \psi)\}(E_2 \cos \Omega \cos i + E_3 \sin \Omega \\
& + C_2 \sin \Omega \cos i - C_3 \cos \Omega)] \Big\} + \mathbf{j} \Big\{ A \cos \psi \\
& - B \sin \psi + 3[\cos (\Omega - \psi)(E_1 \sin \Omega \\
& + E_3 \cos \Omega \cos i - C_1 \cos \Omega + C_3 \sin \Omega \cos i) \\
& - \cos i \sin (\Omega - \psi)(E_2 \cos \Omega \cos i \\
& + E_3 \sin \Omega + C_2 \sin \Omega \cos i - C_3 \cos \Omega)] \Big\} \Big].
\end{aligned}
$$

Using the magnetic moment program, it has been shown in the case of the Telstar satellites that letting $M = v$ produces precessional results that follow the case II approximation over 1,000 orbits to within 0.01°. To document this result, the residual magnetic moment used was -0.9 microweber-meter, spin rate equaled 20 to 10 radians per second, spin axis and transverse moments of inertia were 4 slug-feet^2, orbit perigee was set to 4,500 miles, and eccentricity ranged from 0 to 0.95. Slight changes in the x-y torques of the order of thousandths of a microfoot pound were observed as the principal differences in the case I and II approximations for these eccentricity ranges. While these differences are negligible for the Telstar I and II satellites, other satellites in sufficiently lower orbits or having greater residual magnetic moments could require the case II approximation (no complete study has been made to date to bound the required ranges of the above mentioned variables for case I to achieve agreement to within 0.01° of case II).

4.7 *Equations of Motion of a Body Symmetrical about Its Spin Axis*[16]

The mean gravity and magnetic torque equations have been derived in the previous section. To analyze the motion of a spin-stabilized satellite responding to these torques, certain gyroscopic equations must now be developed. We begin by relating the vector angular momentum, $\mathcal{H}$, for any rotating body to the external forces acting on the body as

$$\mathcal{H}' = \mathbf{N} \tag{142}$$

where

$\mathbf{N}$ = resultant moment of all external forces acting on the body
$\mathcal{H}'$ = the time derivative of $\mathcal{H}$.

Using the assumptions in 4.3, we write in the SANOR system

$$\mathcal{H} = \dot{\varphi} I_3 \mathbf{k} \tag{143}$$

where

$$\dot{\varphi} = \text{the satellite spin rate.}$$

From this, it follows that (see Fig. 10)

$$\mathcal{H}' = I_3[\ddot{\varphi}\mathbf{k} + \dot{\varphi}(\boldsymbol{\omega} \times \mathbf{k})] \tag{144}$$

where

$\boldsymbol{\omega}$ = angular velocity of the SANOR system referenced to IS coordinates.

We may express $\boldsymbol{\omega}$ in the SANOR system as (see Fig. 10)

$$\boldsymbol{\omega} = \omega_x \mathbf{i} + \omega_y \mathbf{j} + \omega_z \mathbf{k} \tag{145}$$

or

$$\boldsymbol{\omega} = \dot{\theta}\mathbf{i} + \dot{\psi} \sin \theta \mathbf{j} + \dot{\psi} \cos \theta \mathbf{k} \tag{146}$$

where

ω_x, ω_y, ω_z = angular velocity about the a, y, and z axes respectively.

Combining (142), (144), and (146), after performing the indicated operations

$$\mathbf{N} = I_3(\dot{\varphi}\dot{\psi} \sin \dot{\theta}\mathbf{i} - \varphi\dot{\theta}\mathbf{j} + \ddot{\varphi}\mathbf{k}). \tag{146}$$

Expressing (146) as x, y, and z torques in the SANOR system,

$$T_x = I_3\dot{\varphi}\dot{\psi} \sin \theta \tag{147}$$

$$T_y = -I_3\dot{\varphi}\dot{\theta} \tag{148}$$

$$T_z = I_3\ddot{\varphi}. \tag{149}$$

Quite obviously it is the x and y torques which produce precession. These torques, by the assumptions of Section IV, are the sums of the x and y components of the gravity and magnetic torques expressed by (43) and (141). T_z is zero as a result of assuming zero magnetic moment transverse to the spin axis and mass symmetry about that axis. We note that a nonzero T_z implies a change in the spin rate. For the Telstar I and II satellites this takes place principally because of induced eddy currents which produce transverse moments.*

* The general equations of motion referred to the center of mass for a rigid body spinning about the z axis, and symmetrical about this axis, are

4.8 *Alternate Inertial Coordinate System*

Clearly, (143) is valid only if the angular momentum vector, $\mathcal{H}$, coincides with the satellite spin axis **k**. We note in (147), which was derived from (143), that the above assumption will be approximated only if $\dot{\psi}$ sin θ is small as compared to $\dot{\varphi}$. Otherwise the direction of $\mathcal{H}$ shall certainly be influenced by that term as well as $\dot{\varphi}$. We note too that a valid condition within the bounds of the assumption is for $\dot{\psi}$ sin θ to be small even with large $\dot{\psi}$ provided only that θ itself be appropriately small. That is, solutions to the equation of motion exist for large $\dot{\psi}$, and those will occur only for small θ because of (143).* But if θ be near 0° or 180°, a singularity exists in (147), for $\dot{\psi}$, as a result, approaches infinity. An exit from this dilemma may be secured simply by transforming to a new reference set of inertial coordinates in place of the IS system whenever θ becomes small. Naturally, the new set should be chosen so that the equivalent θ then existing will be large. This is accomplished by redefining IS so that

$$X_2 \text{ corresponds to } Y$$

$$Y_2 \text{ corresponds to } Z$$

$$Z_2 \text{ corresponds to } X$$

as shown in Fig. 15. Let us call this new inertial frame the IS2 system.

Transformation equations to relate IS to IS2 are quite simple and are given below:

$$X_2 = QX \tag{150}$$

$$X = \tilde{Q}X_2 \tag{151}$$

$$T_x = I\ddot{\theta} + (I_3 - I)\dot{\psi}^2 \sin\theta\cos\theta + I_3\dot{\varphi}\dot{\psi}\sin\theta$$

$$T_y = I\ddot{\psi}\sin\theta + (2I - I_3)\dot{\psi}\dot{\theta}\cos\theta - I_3\dot{\varphi}\dot{\theta}$$

$$T_z = I_3(\ddot{\psi}\cos\theta - \dot{\varphi}\dot{\theta}\sin\theta + \ddot{\varphi}).$$

For cases where $\dot{\varphi}$ dominates , $\dot{\theta}\dot{\psi}$ sin θ, and $\dot{\psi}$ cos θ, these torques reduce to those given by (147), (148) and (149). We note principally that a change in spin rate reflects a T_x processional torque, from the above equations, but this is generally small and is neglected in this paper. For Telstar I, this torque component is estimated at least two orders of magnitude below the magnetic torques considered herein. (See also Ref. 17.)

* This is the case for precession through or near the north or south celestial pole, where even small changes in attitude, or θ, can produce large changes in $\dot{\psi}$. The situation is quite analogous to an azimuth-elevation antenna tracking a satellite that passes through or near the zenith, where the azimuth rates become extremely high even for small changes in satellite position.

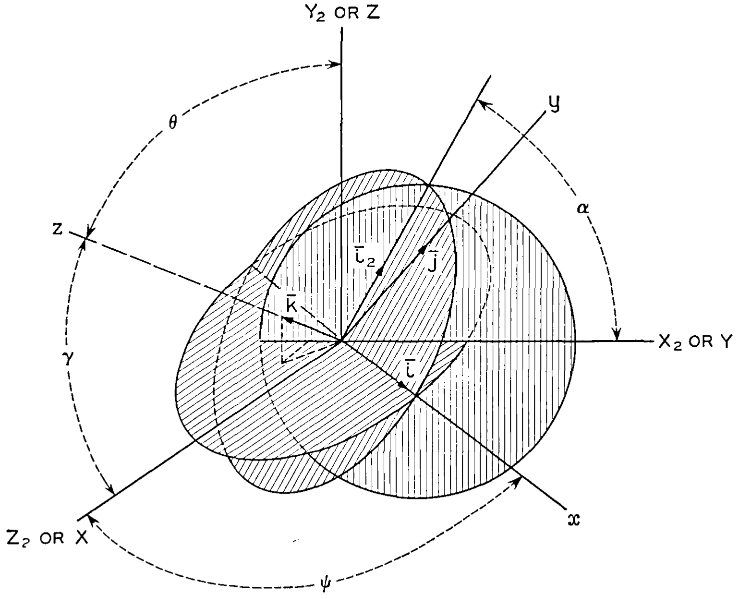


Fig. 15 — Euler rotations in the alpha-gamma system.

where

$$Q = \begin{vmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{vmatrix} \tag{152}$$

and

$$\tilde{Q} = \begin{vmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{vmatrix}. \tag{153}$$

Relationships between the other coordinate systems and the IS2 system may be obtained using equations of Section 4.4 and applying the Q matrix as appropriate, except that in the expanded matrices

α corresponds to ψ

γ corresponds to θ

where

α = The Euler angle measured from X_2 to the intersection of the xy and X_2Y_2 plane, which defines a new x axis called the x_2 axis in a SANOR 2 system. (See Fig. 15.)

γ = The Euler angle measured from Z_2 to $\mathbf{k}$ in a plane perpendicular to the x_2 axis. (Rotation is about the x_2 axis.)

Quite obviously, the steps leading to the mean gravity and mean magnetic torques may be retraced using the Q matrix and the new α,γ Euler angles. This would lead to torque expressions as functions of these desired angles. The same result may be obtained by transforming the final torques expressed in the ψ-θ system to expressions in the α-γ system by using explicit relationships between these two systems. We shall choose this later route and the needed relationships will now be derived.

We note in Fig. 15 that the $\mathbf{k}$ (satellite spin) vector remains in the same spatial position whether expressed in the IS or IS2 system. It seems reasonable then to proceed to relate α-γ to ψ-θ by observing the projections of the $\mathbf{k}$ vector on the XY and X_2Y_2 planes, respectively.

Referring to Fig. 16, we see that

$$\tan \alpha = \frac{-\sin \theta \cos \psi}{-\cos \theta} \tag{157}$$

$$\cos \gamma = \sin \theta \sin \psi \tag{158}$$

$$\sin \gamma = (1 - \sin^2 \theta \sin^2 \psi)^{\frac{1}{2}}. \tag{159}$$

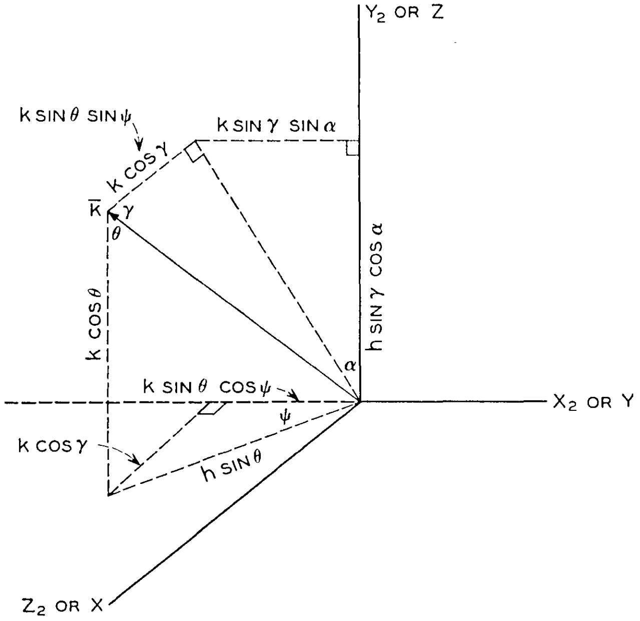


Fig. 16 — Theta-psi and alpha-gamma geometric relationships.

Quite obviously, inverse relationships may be developed. These are

$$\tan \psi = \frac{\cos \gamma}{-\sin \alpha \sin \gamma} \tag{160}$$

$$\cos \theta = -\cos \alpha \sin \gamma \tag{161}$$

$$\sin \theta = (1 - \cos^2 \alpha \sin^2 \gamma)^{\frac{1}{2}}. \tag{162}$$

4.9 *Mean Gravity and Magnetic Torques — Alternate Euler Rotations*

Using (157) through (159) it can be shown that the mean gravity torque expressed in the α-γ system is

$$\begin{aligned}(\mathbf{T}_G)_{\text{mean}} = \frac{6\pi^2 \Delta I}{T_n^2(1 - e^2)^{\frac{3}{2}}} \Big\{ &(-\sin \Omega \sin i \cos \gamma \\ &+ \cos \Omega \sin i \sin \alpha \sin \gamma + \cos i \cos \alpha \sin \gamma) \Big\} \\ &[\mathbf{i}_2(\sin \Omega \sin i \sin \gamma + \cos \Omega \sin i \sin \alpha \cos \gamma \\ &+ \cos i \cos \alpha \cos \gamma) + \mathbf{j}_2(\cos \Omega \sin i \cos \alpha \\ &- \cos i \sin \alpha)].\end{aligned} \tag{163}$$

Likewise, the mean magnetic torque can be expressed as

$$\begin{aligned}(\mathbf{T}_m)_{\text{mean}} = \pi \mathfrak{M} \cos \beta \Big[\mathbf{i}_2 \Big\{ &-2 \cos \alpha \cos \gamma + 3 \sin i(\sin i \cos \alpha \cos \gamma \\ &- \sin \Omega \cos i \sin \gamma - \cos \Omega \cos i \sin \alpha \cos \gamma) \Big\} \\ &+ \mathbf{j}_2 \Big\{ 2 \sin \alpha - 3 \sin i(\sin i \sin \alpha \\ &+ \cos \Omega \cos i \cos \alpha) \Big\} \Big] \\ &+ \mathfrak{M} \sin \beta \Big[\mathbf{i}_2 \Big\{ -A \sin \gamma \\ &- B \sin \alpha \cos \gamma + 3G_1(\sin \Omega \sin \alpha \cos \gamma \\ &- \cos \Omega \sin \gamma) + 3G_2(-\sin i \cos \alpha \cos \gamma \\ &+ \sin \Omega \cos i \sin \gamma + \cos \Omega \cos i \sin \alpha \sin \gamma) \Big\} \\ &+ \mathbf{j}_2 \Big\{ -B \cos \alpha + 3G_1 \sin \Omega \cos \alpha \\ &+ 3G_2(\sin i \sin \alpha + \cos i \cos \Omega \cos \alpha) \Big\} \Big]\end{aligned} \tag{164}$$

where

$$G_1 = E_1 \sin \Omega + E_3 \cos \Omega \cos i - C_1 \cos \Omega + C_3 \sin \Omega \cos i \tag{165}$$

$$G_2 = E_2 \cos \Omega \cos i + E_3 \sin \Omega + C_2 \sin \Omega \cos i - C_3 \cos \Omega. \tag{166}$$

All other terms have meanings described previously.

V. EXPERIMENTAL RESULTS — GENERAL

Using the methods described herein, the attitudes of the Telstar I and II satellites have been predicted, determined, and these two results compared. In the case of the Telstar I satellite, the maximum angular deviation between predicted and determined attitude is 0.90° with the special-case exception of pass 2154, which is covered in detail later.* The average angular deviation is 0.38°. The predicted attitude of the Telstar II satellite deviates from that determined by optical flashes and solar sensors by a maximum of 0.5° and an average of 0.09°. These results are tabulated in Tables II and III. Here the deviation in right ascension and declination is displayed by subtracting the predicted values from the determined values. The angular deviation between predicted and determined attitude is also given as the celestial great circle arc G. When this value is followed by N, this indicates that no attitude fix has been determined by loci intersections (see Section II), and so the shortest arc distance between the corresponding attitude locus and the prediction is quoted. The T preceding an attitude determination indicates the selection of that point as a target toward which a magnetic moment program embodying the above analysis attempts to converge.

5.1 *Experimental Results — The Telstar I Satellite*

The Telstar I satellite entered orbit on July 10, 1962. Its attitude, calculated by combining nominal third-stage burnout parameters and stage 1 and 2 telemetry data, is given in Table I. Between launch and pass 16, there were a number of instances when the orientation coil was inadvertently energized by misinterpreted command signals sent to the satellite at extreme ranges and/or low elevations. This orientation coil (often referred to as a "torque coil") is located in the equatorial plane of the Telstar satellite just under the outer skin. Its purpose is to enable attitude adjustments by the production of a magnetic moment whenever current is sent through the coil. A fixed amount of current may be caused to flow in either direction through the coil so as to produce a magnetic moment, in addition to the residual moment, of ±7.8540 microweber-meters. Fig. 17 diagrams the sense of the magnetic vectors which are called positive, indicates the resulting north and south magnetic poles of the satellite, and shows the corresponding positive direction of the current flow in the orientation coil. The relative magnetic directions and current flow are in accord with established standards.

* There are apparent exceptions on passes 16, 72 and 1657, but these relate to antenna pattern techniques (see Section 5.1) which have an expected accuracy of ±1° to ± 2°.

TABLE II—TELSTAR I SATELLITE—ATTITUDE COMPARISON

Pass Number	Attitude Determination		Type Data for Determ.	Attitude Prediction		Deviation (Det.—Predict.)			MAG MOM (MWM)
	RA	DEC		RA	DEC	RA	DEC	*G*	
Burnout				83.7	−66.8				
16	84.2	−65.8	SM	84.2	−65.18	0	0	0	−0.86328
16	83.0	−66.0	A	84.2	−65.18	−1.2	−0.82	1.4	↓
72				87.78	−66.27			0.5N	↓
72	86.0	−67	A			−1.3	−0.73	1.2	↓
135-136	91.7	−66.0	M	92.19	−66.36	−.49	0.36	0.6	↓
199	single	locus	M	96.49	−65.95			0.4N	↓
271	100	−65	A	100.80	−64.92	−0.8	−0.8	0.5	↓
272	single	locus	M	100.90	−64.90			0.6N	↓
272	↓	↓	S	100.90	−64.90			0.2N	↓
472	↓	↓		107.50	−59.80			0.2N	↓
931	T100.5	−49.84	M	100.74	−49.93	−0.24	0.09	0.18	↓
1051	single	locus	↓	97.25	−50.21			0.9N	−0.86328 +6.9907
1069	↓	↓	↓	98.42	−49.70			0.6N	−0.86328 −8.7173
1114	↓	↓	↓	95.92	−51.01			0.1N	−0.71103
1430	↓	↓	↓	92.62	−59.55			0.1N	↓
1567	T96.2	−64.2	M	96.71	−64.21	−0.51	−0.01	0.23	↓
1657	99.2	−67.0	A	102.81	−66.64	−3.61	0.36	1.7	−0.71103 7.14297 −0.59101
1695				78.58	−57.01				↓
1909				90.59	−59.71			0.7N	↓
2154-55	T104.0	−53.5	M	104.71	−54.81	−0.71	−1.3	1.5	↓
2200	single	locus	↓	106.03	−53.28			0.1N	↓
2264	↓	↓	↓	107.20	−50.74			0.4N	↓
2464	↓	↓	↓	107.09	−43.65			0 N	↓
2482	↓	↓	↓	106.88	−43.04			0 N	↓
2509	↓	↓	↓	106.57	−42.33			0 N	↓
2582-83	105.2	−40.2	M	105.44	−40.53	−0.24	0.38	0.4	↓
3340, 41, 42	106.3	−43.7	M	103.68	−61.54	−2.62	17.8	18.0	↓
3476-7	112.30	−46.32	M	120.21	−66.21	7.91	19.9	22.0	↓
3495				123.16	−66.41				↓
3476-7	T112.30	−46.32	M	112.30	−46.32†	0	0	0	−0.40945

M = Mirror flash data.
A = Antenna pattern data.
S = Solar sensor data.
† Connects 3340 to 3476.

Data relating to the inadvertent uses of the orientation coil are insufficient to reconstruct the detailed precessional motion of Telstar I from launch to pass 16. Cn pass 16, however, an attitude determination was made by reducing optical and solar sensor data. This along with an attitude fix on pass 135–136 made from mirror flash data indicated a magnetic moment of about −0.7 microweber-meter. This was later refined to −0.86328 microweber-meter by causing the magnetic moment program to connect, by the precessional theory herein described, the

Table III—Telstar II Satellite—Attitude Comparison

Pass Number	Attitude Determination		Type Data	Attitude Prediction		Deviation (Det.-Predict.)			MAG MOM (MWM)
	RA	DEC		RA	DEC	ΔRA	ΔDEL	*G*	
Burnout				82.23	−57.31				
62, 63, 63	87.75	−55.08	M	87.75	−55.08	0	0	0	−0.48375
100				88.55	−55.05	0	0	0	−0.48375
132, 133, 133	88.75	−55.34	M	89.23	−54.97	−0.48	−0.37	−0.5	
281				92.17	−54.10			0N	
293, 293	92.3	−54.05	M	92.38	−53.99	−0.08	−0.06	0.1	
324, 25	92.40	−53.71	M	92.90	−53.71	0	0	0	
331				93.01	−53.66			0.08N	
490				94.95	−51.81				
496	T95.00	−51.73	M	95.00	−51.73			0N	
541				95.34	−51.10			0N	
573				95.49	−50.74			0N	
611				95.63	−50.20			0.13N	
643				95.71	−49.80			0.16N	↓

M = Mirror flash data.
S = Solar sensor data.

attitude on pass 16 to a point on the pass 931 attitude locus. The convergence is within 0.18°, as shown in Table II. Using this latter magnetic moment, the precessional motion from passes 16 to 1051 was established. This precessional history is given in Fig. 18, where both the predicted curve and the attitude loci data are displayed.

When an optical flash series is recorded at the Holmdel Laboratories, the midpoint of that series can be determined with a time accuracy of less than ±20 seconds when using visual observation through the on-site telescopes and within ±10 seconds photoelectrically. The early data recorded on the Telstar I satellite did not indicate the expected accuracy at the midpoints, so in Fig. 18 only central loci are plotted. We must therefore think of these loci as having possible tolerances up to ±10 seconds, since photoelectric data were reduced. A tolerance of ±10 seconds can cause the maximum locus limit to be displaced as much as ±0.7° on either side of the plotted central loci. Considering this, the predicted attitude curve passes through every loci and every loci intersection on Fig. 18 with the single exception of pass 1051. The prediction curve would miss the lower limit of this locus by 0.3°.

Extended attitude predictions were made on October 1, 1962 (around pass 800) and covered the period through January 22 (pass 1800). These indicated that attitude adjustments should be initiated during

January, 1963. This was needed to prevent the solar offset from exceeding about 15° as required (see Section I). In order to check the attitude correction procedures, a test maneuver was initiated on pass 1051. It began by turning on the orientation coil, so as to produce a magnetic moment of +7.8540 microweber-meters to oppose the −0.86328 residual moment and yield a resultant of +6.9907 microweber-meters. This was continued through pass 1058, whereupon the coil was turned off until pass 1069. On 1069 the coil was turned on in the opposite sense so as to produce a moment of −7.8540 microweber-meters, giving the satellite a net moment of −8.7173 microweber-meters. On pass 1075, the coil was turned off. All told, the coil was positive for about 18 hours and negative for about 16 hours, a fact which approximately nullified the

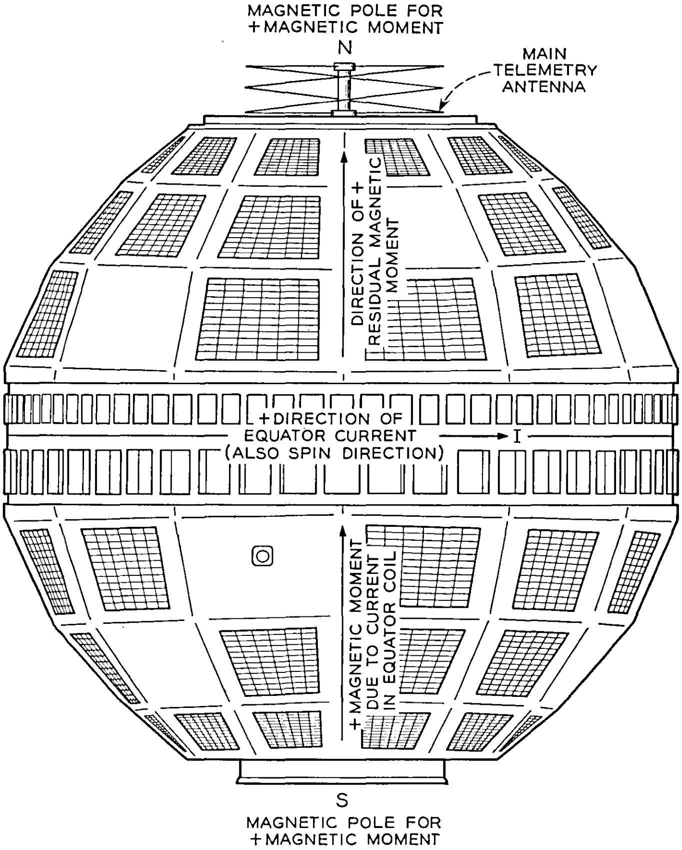


Fig. 17 — Direction of positive magnetic moments for Telstar satellites.

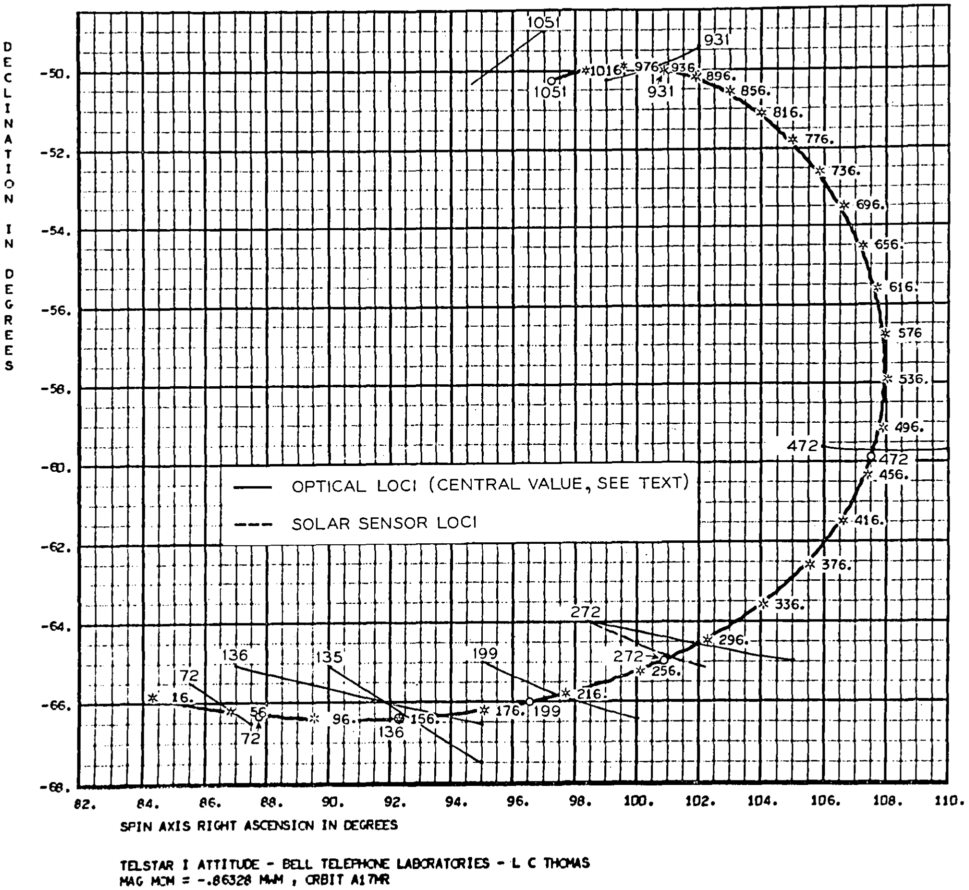


Fig. 18 — Telstar I precessional history — passes 16 through 1051.

precessional effect of the coil as intended for this test. Details of this maneuver are shown in Fig. 19.

Table II summarizes the agreement of loci to prediction during this period. Giving loci of unknown tolerances a ±0.7° spread and using the known tolerance for pass 1114, the residuals shown in Table IV are produced.

The magnetic moment program connected the attitude on pass 1075 to that determined by mirror data on pass 1567 within 0.23°. In so doing, it calculated a residual magnetic moment of −0.71103 microweber-meter for that era. Using this calculated moment, predictions are extended to pass 1657 where, on January 7, 1963, the main orientation

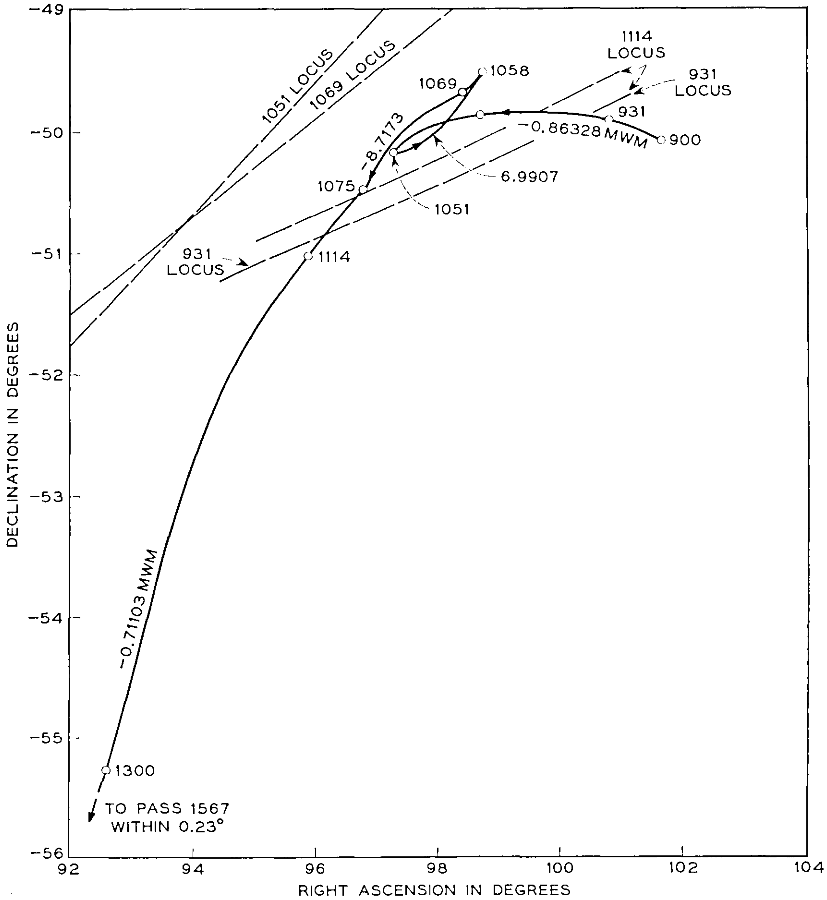


Fig. 19 — Details of the trial torqueing maneuver.

Table IV — Telstar I Satellite — Test Maneuver

Pass Number	Deviation of Prediction from Optical Locus
1051	0.3°
1069	0°
1114	0.1°

maneuver began. The region from pass 1075 to 1657 is shown in Fig. 20 along with the attitude loci. Predictions fit the determinations throughout within about 0.2°.

In addition to using mirror flash and solar sensor data to determine attitude, W. C. Jakes, Jr. and group at the Holmdel Laboratories deduced the attitude for a number of passes in this period by analyzing antenna pattern data.[18] Some of these determinations are included in Table II. Since their expected accuracy is ±1°, these data also serve to corroborate the predicted attitude.

The orientation coil was energized for the major attitude maneuver

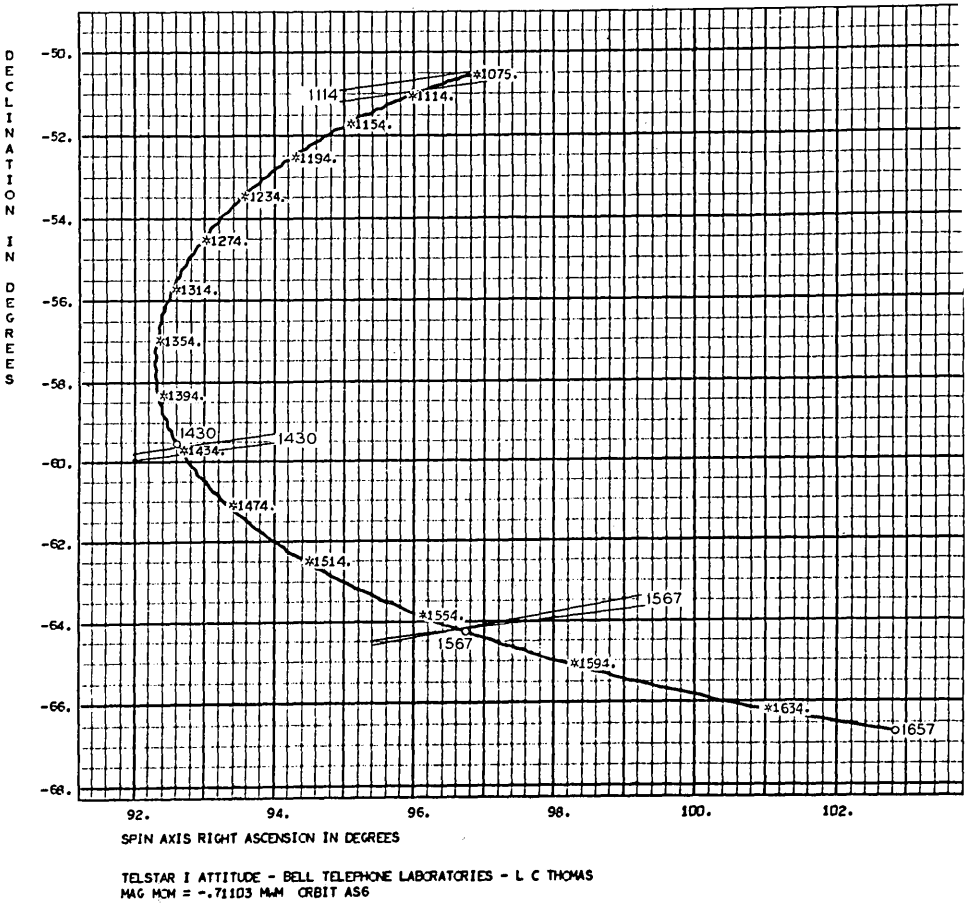


Fig. 20 — Telstar I precessional history — passes 1075 to 1657.

beginning on pass 1657, January 7, 1963, and ending on pass 1695, January 11, 1963. During this time the residual magnetic moment of −0.71103 microweber-meter had superimposed upon it a 7.8540 microweber-meter moment as a result of the coil. The attitude of the Telstar I satellite was changed by over 20° during this maneuver. The attitude and curve for this period are given in Fig. 21. The slight ripples in these curves are due to the inclination of the earth's dipole. Because of the small scale of this plot, these ripples are more noticeable than in the previous graphs.

At this point, we might well insert a brief discussion of the planning

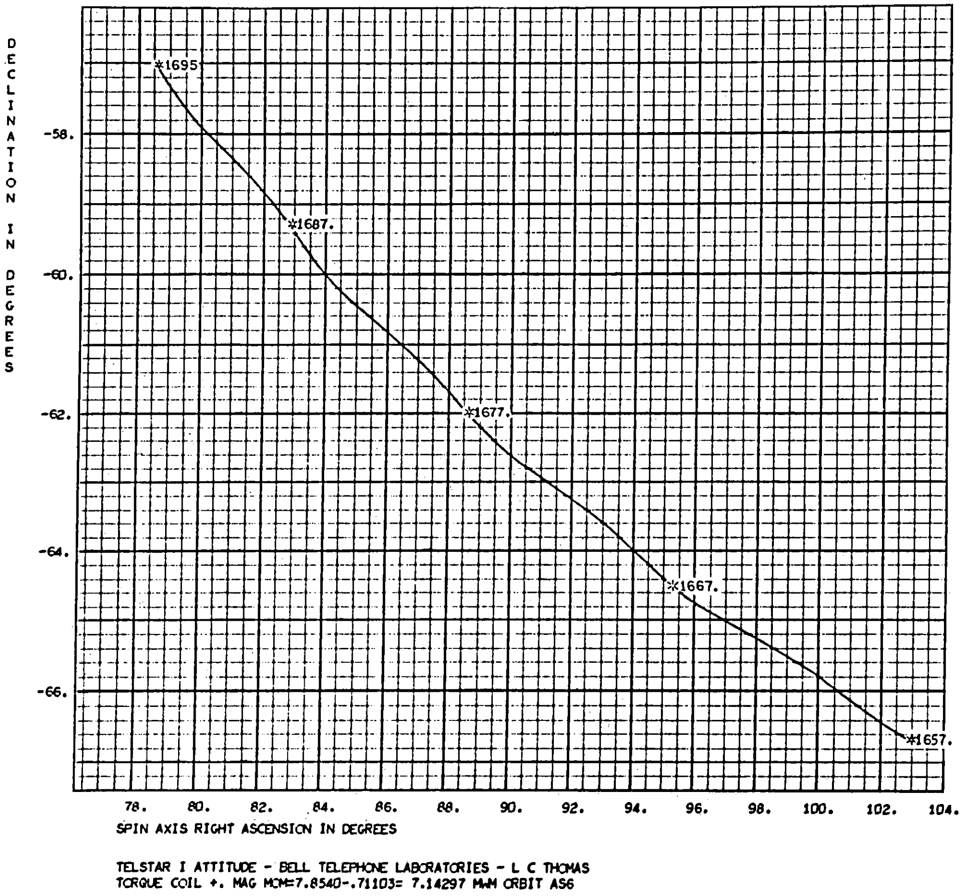


Fig. 21 — Telstar I precessional history — passes 1657 to 1695.

behind this attitude maneuver and why pass 1657 was chosen for its commencement. First of all, it can be shown that for any angular offset in the spin axis from a fixed, desired orientation there exists, for each instant in time, a determinable polarity of voltage for the equatorial coil which will tend to decrease the offset. To utilize this principle fully would require turning on the orientation coil on many occasions when the Telstar I satellite entered the Andover skies. While such a procedure would tend to decrease the angular offset generally in the most direct fashion, there are occasions in the lifetime of the Telstar satellite when leaving the coil on for a number of revolutions will produce faster corrections than pulsing the coil continually at Andover. This is because the instantaneous torque produced by the continuous magnetic moment of the coil in the earth's magnetic field varies during an orbital period in such a way that, in general, the average value exceeds the torque obtainable by energizing the coil only in the Andover skies. The optimum procedure then is for a period of continuous operation and this was undertaken. Fig. 22 shows the results of turning on the coil for steady torqueing at various dates. The heavy spiral shows the predicted position of the spin axis in right ascension and declination if the coil is never energized (the trial maneuver is omitted here for simplicity). The diverging curves indicate the motion due to the torqueing coil. Examination of Fig. 22 shows rapid precession as a result of the torque coil field, so that no mode of continuous coil operation will long permit the spin axis to point close to the south ecliptic pole. The solar offset also must ultimately increase with the enlarging ecliptic angle. It therefore follows that if a simplified mode of coil operation exists, it must begin with a relatively short period of coil operation.

Two factors concerning the precessional motions are noted in Fig. 22. They are:

(*i*) All precessional motions involving the orientation coil lie outside the spiral of residual precession for many orbits, because the net magnetic moment existing when the coil is actuated exceeds the residual magnetic moment.

(*ii*) Torque coil precession for both positive and negative coil polarity begins incrementally at the spiral in opposite directions.

We see that the torque coil should be used sparingly, because of the relatively rapid motion it produces. The problem that remains is to determine the times to turn on and turn off this coil and also to investigate whether or not future coil use is required. Because the spin rate of the satellite is decreasing, equal magnitudes of satellite magnetic moments will cause greater and greater precessional motion as time passes.

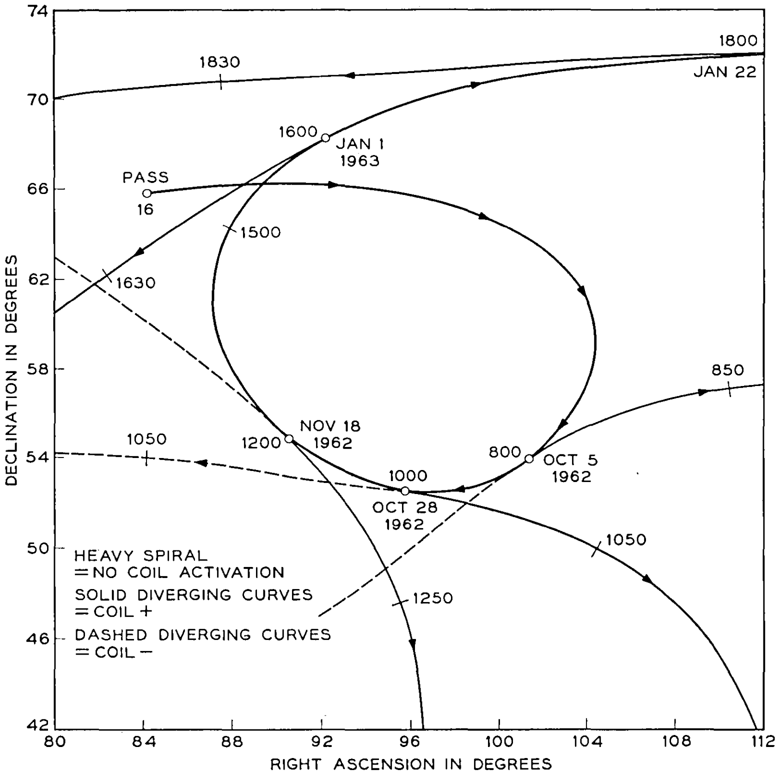


Fig. 22 — Telstar I predicted attitude showing steady torqueing.

It is to be expected that after the coil is actuated and then turned off a spiral of larger excursion than that occurring for passes 16 through 1655 will result strictly because of the lower spin rate existing at the later time. There is no way to avoid this except by repeated coil pulsing.

To avoid, or at least minimize, repeated coil usage, the procedure is to try to keep the maximum ecliptic angle as small as possible. If this value is expected to exceed the allowable solar offset, it is advantageous to time the operation so that the maximum ecliptic angle will occur at the time of minimum solar offset. This criterion governs both the selection of the day for coil turn-on as well as the duration of the attitude maneuver. As may be guessed from Fig. 22 and from the precessional motion factors stated above, only short torqueing operations during the month of January would result in the next precessional spiral (after coil turn-off) returning anywhere near the initial spiral shown in Fig. 22.

To cause the maximum ecliptic angle and the minimum solar offset to occur together, orientation coil use had to begin in the period from January 1 to January 7, 1963. As shown above, the maneuver which did begin on January 7 placed the attitude at right ascension 78.6° and declination at 57.0° by pass 1695.

The magnetic moment program, in connecting pass 1695 attitude to that determined by mirror data on passes 2154 and 2155, determined a residual magnetic moment of -0.59101 microweber-meter. This differs by about 0.1 microweber-meter from the residual moment going into the torqueing maneuver, but appears to be borne out by the closeness of fit to the mirror data from pass 1695 through pass 2583 (see Table II).* Fig. 23 plots this trajectory along with the mirror loci. The fit of attitude predictions to determinations is very good from passes 1695 through 2583 (see Table II). The largest deviation of 1.5° on passes 2154–2155 may be somewhat misleading. Referring to Fig. 23, we see that the predicted attitude on pass 2154 is easily within 0.1° of the 2154 locus. The interaction of the 2154 and 2155 loci however, produces a common area, or attitude box, 1.5° from the pass 2154 prediction. This box is produced by a somewhat grazing intersection of the loci, and its location is therefore affected by the precession which took place between these two passes. Antedating the 2155 locus to account for this would place the attitude box within 0.05° of the 2154 and 2155 prediction, so this fit is quite valid.

We are not in so comfortable a position for passes 3340 and thereafter. Attitude boxes for 3340–1 and 3476–7 are shown as dashed lines on Fig. 23. They are about 20° from the corresponding predicted attitudes, and the reason for this discrepancy is at present unknown. We note that the Telstar I satellite ceased its transmission on pass 2065 (February 20, 1963) because of radiation damage, but that the attitude remained predictable at least through pass 2583. The region from pass 2583 to 3340 is devoid of attitude data because no telemetry could be received to report solar aspect and no mirror flashes were recorded. This was due to the increased activity on Telstar II, a certain proportion of Telstar I passes occurring in daylight hours, and prevailing weather conditions at Holmdel.

Interestingly enough, it is noted that the attitude determinations on passes 3340–1 and 3476–7 can be connected by the magnetic moment program even though they cannot be sensibly joined to the preceding

* It is possible that coil usage can alter the residual moment of the satellite, which was made small in the first place by appropriately balancing much larger magnetic fields.[19]

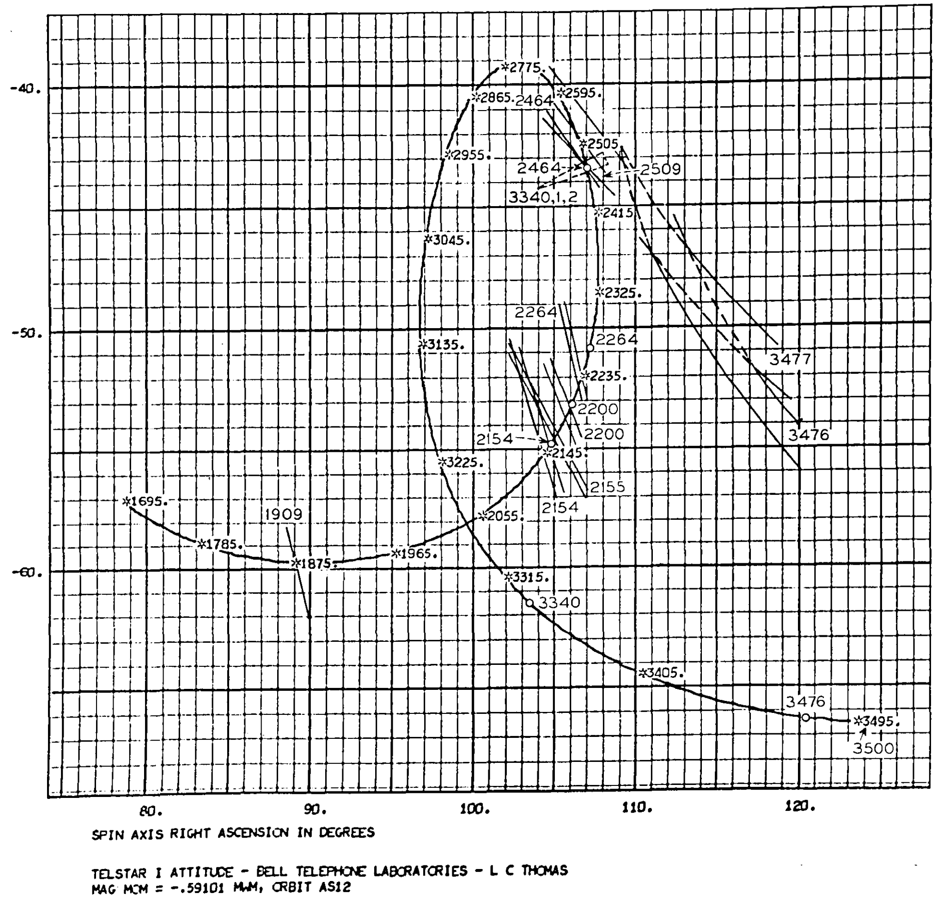


Fig. 23 — Telstar I precessional history — passes 1695 to 3500.

data. Fig. 24 shows this union. The magnetic moment of −0.59101 microweber-meter used in these last three plots cannot be considered too reliable, since it is one of a number which can connect the two attitude boxes shown in Fig. 24. However, it is the same as previously used in the 1695 to 2583 region and may tend to be valid because of this. In any event, more attitude data are required to resolve this issue.

As to what may have caused the anomalous behavior of Telstar I somewhere between pass 2583 and 3440, we offer the following, taken singularly or in combination, as possibilities:

(*i*) the orientation coil has been energized during this period (from

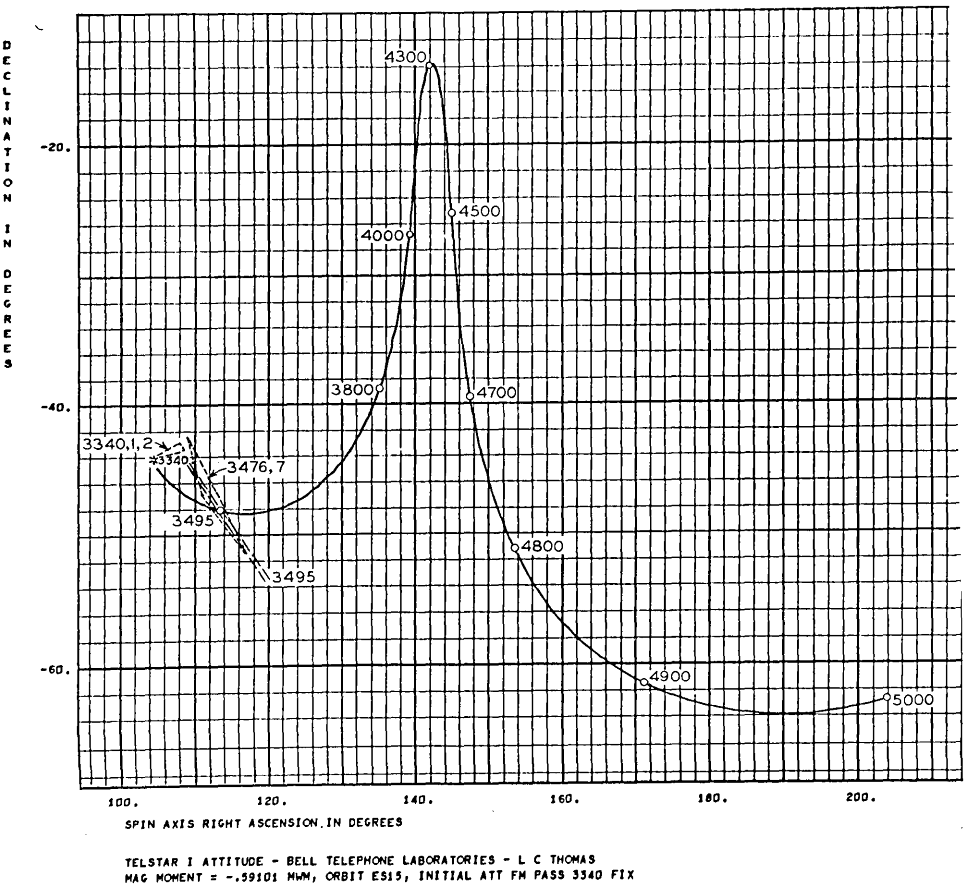


Fig. 24 — Telstar I precessional history and prediction — passes 3340 to 5000.

the present precessional motion, it appears to have been turned off between 3340 and 3477);

(*ii*) the residual magnetic moment has changed a multiplicity of times;

(*iii*) meteoric collision has taken place;

(*iv*) pressure leakage from instrument canister has occurred in a manner to alter the attitude by reaction forces.

5.2 *Experimental Results — The Telstar II Satellite*

The Telstar II satellite entered orbit on May 7, 1963. Its initial attitude calculated from the third-stage burnout parameters is shown in Table III. The predicted burnout attitude proved inaccurate because the third stage of the Thor-Delta vehicle did not perform nominally. This is evidenced by the fact that the orbital period calculated from such assumed nominal performance differed from the actual period by as much as four minutes.

The first attitude fix of the Telstar II satellite occurred on passes 62 and 63, when a total of three flash series were observed. Connecting this determination to that of pass 496 requires a residual magnetic moment of −0.48375 microweber-meter. The fit through pass 643 is given in Table III, and some typical attitude loci are plotted along with the attitude prediction in Fig. 25. Corresponding ecliptic angle, solar aspect, X and Y torques appear in Figs. 26 through 29.

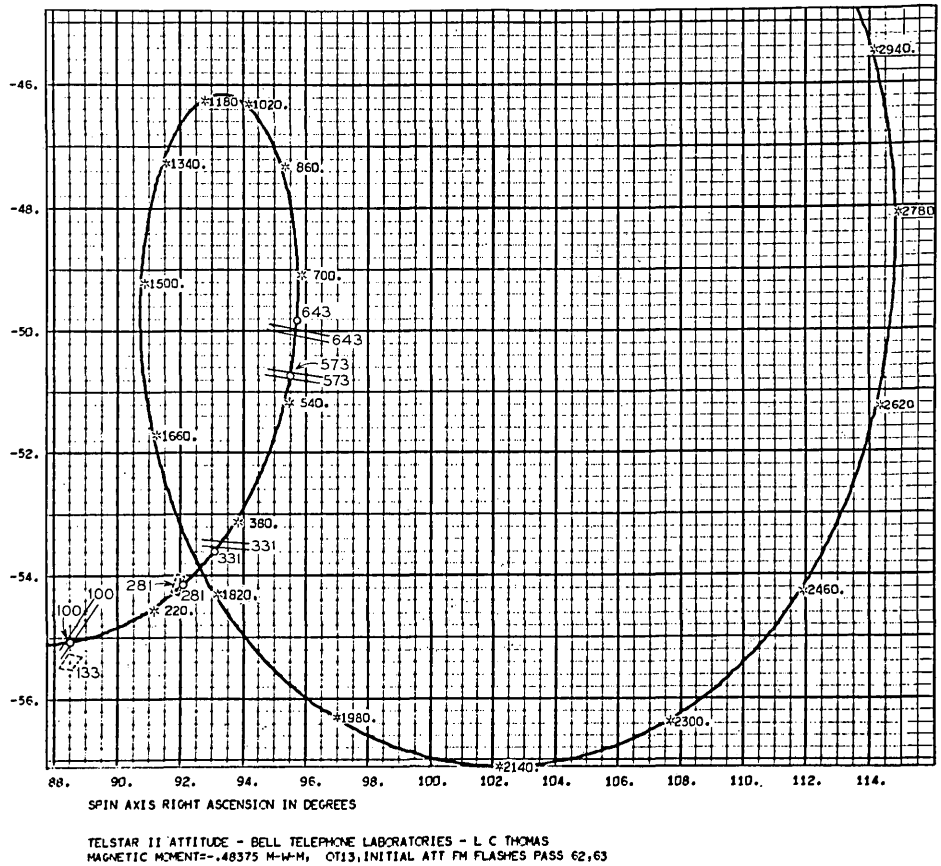
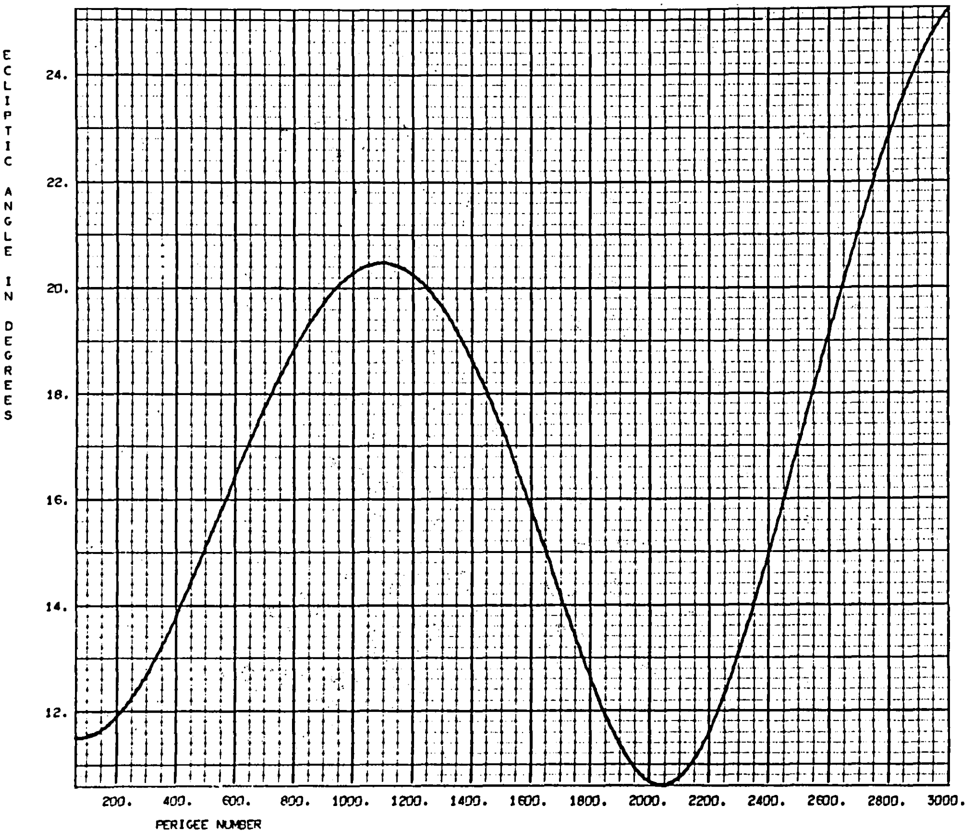


Fig. 25 — Telstar II precessional history and prediction — passes 62 to 3000. (See footnote, p. 1713.)



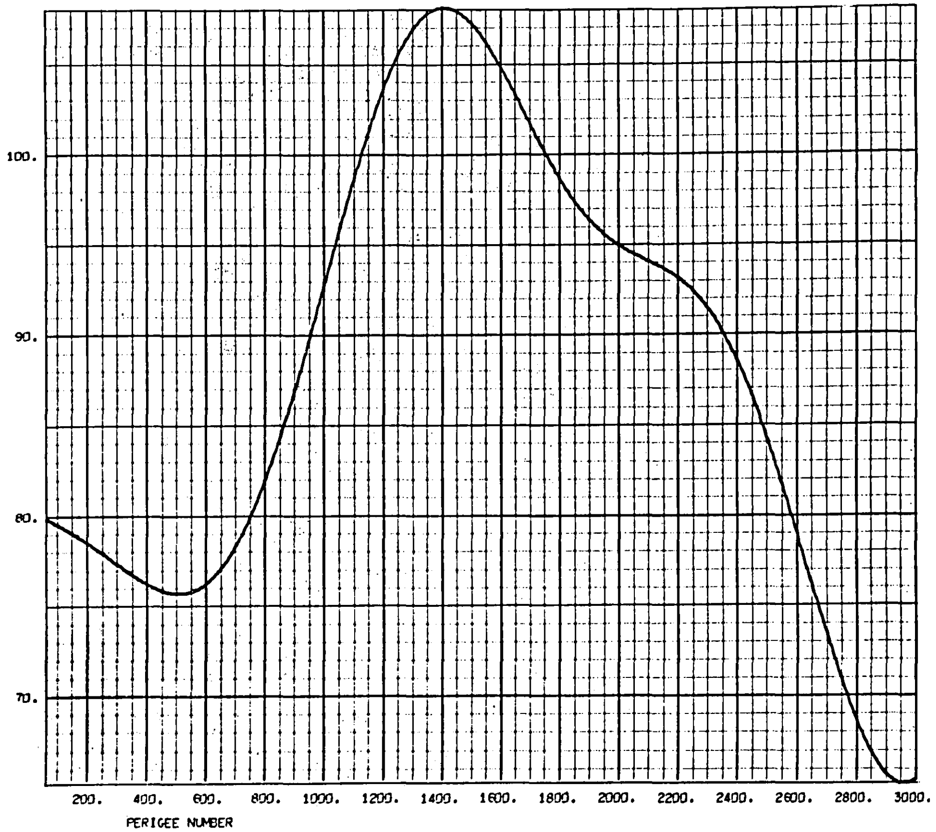
TELSTAR II ATTITUDE - BELL TELEPHONE LABORATORIES - L C THOMAS
MAGNETIC MOMENT=-.48375 M-W-M, OT13,INITIAL ATT FM FLASHES PASS 62,63

Fig. 26 — Telstar II ecliptic angle — passes 62 to 3000.

The solar offset is due to exceed the 15° limit dictated by temperature balance considerations (see Section I) by pass 1250 (see Fig. 27). Since the maximum offset which occurs on pass 1400 is only 18°, no plans are contemplated to reorient the Telstar II satellite strictly to prevent this mild excursion.

There is, however, another excursion beyond the 15° limit, around pass 2670. Besides this excursion being more serious than the former, some interesting differences between these two events exist, as Fig. 25 illustrates.

On pass 1400, the attitude will be returning to a region nearer the south ecliptic pole (located at right ascension 90°, declination 67.5°)



TELSTAR II ATTITUDE - BELL TELEPHONE LABORATORIES - L C THOMAS
MAGNETIC MOMENT=-.48375 M-W-M, OT13,INITIAL ATT FM FLASHES PASS 62,63

Fig. 27 — Telstar II solar aspect — passes 62 to 3000.

thereby limiting the solar offset to values below 15°. On pass 2670, however, the attitude is moving away from the south ecliptic pole, and, in fact, because of its now lower spin rate is entering a spiral of greater excursion than that previously traversed. It therefore behooves us to correct the attitude before that latter spiral occurs.

The question remaining is to determine the most profitable time for such a correction. Ideally, a well-chosen time would meet the following conditions:

(*i*) It would take place just after a good optical attitude fix was established to verify the predictions.

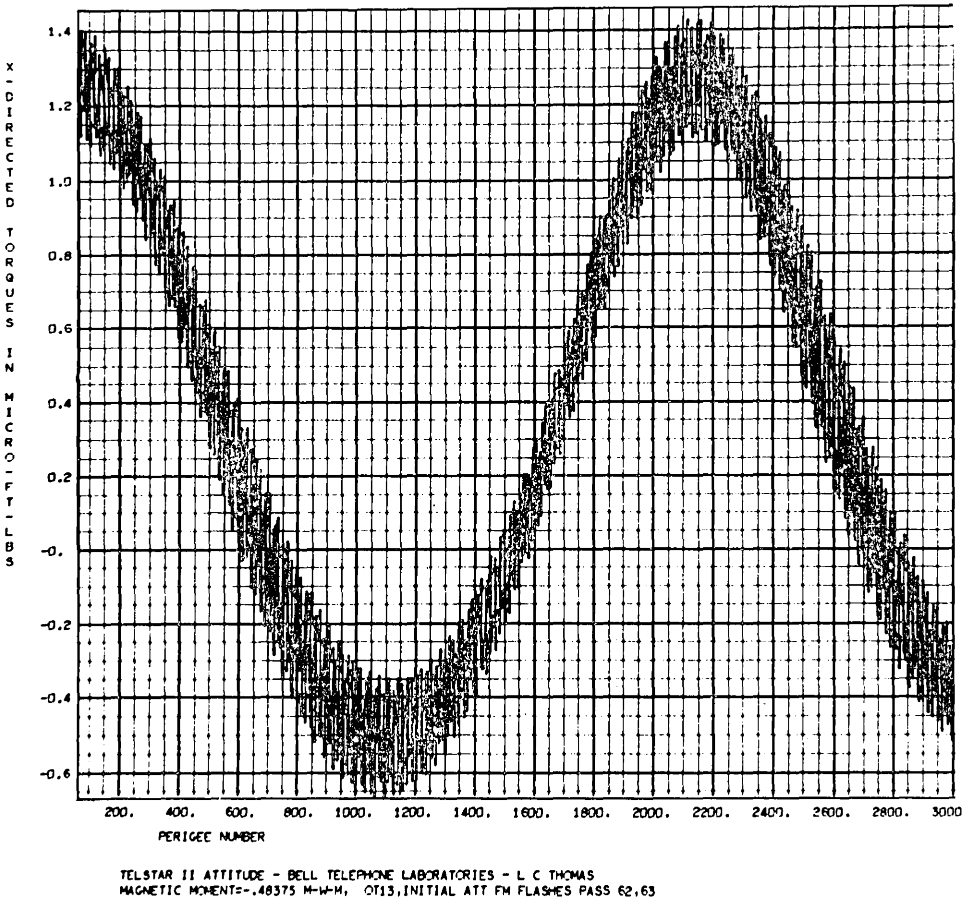


Fig. 28 — Telstar II X torques — passes 62 to 3000.

(*ii*) The attitude maneuver would result in as many mirror flashes as possible during correction to indicate its progress.

(*iii*) The attitude maneuver would end with the satellite in a position to guarantee a suitable attitude for as long a time in the future as is possible.

At the time of this writing, suitable attitude corrections are under study. Most probably the orientation coil will be energized sometime between passes 1300 to 1500 in a negative sense or in a positive sense near pass 2400 so as to drive the attitude downward and to the left in

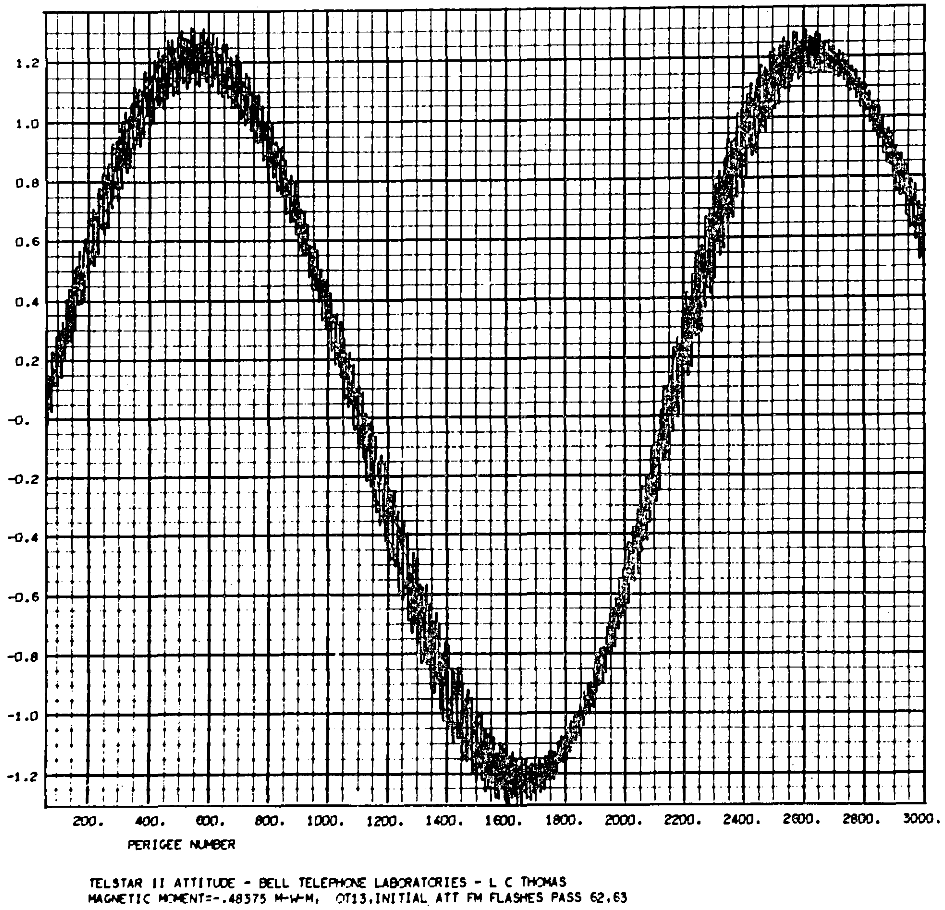


Fig. 29 — Telstar II *Y* torques — passes 62 to 3000.

the sense shown in Fig. 25. This permits the next attitude spiral to occur near the south ecliptic pole.*

VI. CONCLUSIONS

Techniques for both attitude determination and prediction for spin-stabilized satellites have been developed. Their use has been demon-

* The attitude of the Telstar II satellite was successfully reoriented by energizing the coil on pass 2402 (May 17, 1964) and leaving it in that state until pass 2421. Thus attitude predictions given in Fig. 19 beyond pass 2402 are no longer valid.

strated using Telstar I and II satellite data. It has been shown that an inclined dipole model of the earth's magnetic field and the method of averaging the gravitational and magnetic torques over each anomalistic period of the satellite permit attitude predictions to within a few tenths of a degree of determined values in most instances. In those few cases where departures are above one degree, explanations have been presented to show the reason for such discrepancies. The reasons are (1) unknown time errors in determining the midpoint of the optical flash series and (2) grazing intersections of the attitude determining loci. There remains but one anomaly in the precessional motion of the Telstar I satellite which, for the moment, is unexplained. Possible reasons for this anomaly are given in Section 5.1.

It has further been shown that the seemingly crude approximation of letting the mean anomaly of the satellite equal the true anomaly for the purposes of determining the mean torques produces attitudes in close agreement (0.01° over 1000 orbits) with more sophisticated approximations for orbit eccentricities up to 0.9 and perigee radii above about 4500 miles. (See Section 4.6, case II.)

The comparisons made herein of the precession given by the magnetic moment program and the attitude determinations substantiate the simplifying assumptions made in Sections 4.1, 4.2, and 4.3. These assumptions are most instrumental in producing a working technique for both attitude prediction and residual magnetic moment determination which is amenable to analytic solution and conservative of computer time.

Furthermore, the usefulness of combining optical flash and solar sensor data for attitude determination and their inherent accuracy is shown. Optical flash data can provide loci with a resolution of 0.1°. Solar sensor loci are resolved to within 1°. While it is clearly straightforward to determine analytically the boundaries of the attitude boxes from intersecting loci exhibiting estimated time tolerances, this paper indicates the decided advantage of graphing the individual loci to determine the angle of intersection and thereby gain an estimate of the validity of the boundaries in the presence of precession (see, for example, Section 4.1).

Finally, the techniques described have all been consolidated into working computer programs which follow closely the analysis presented. In addition, a number of important supporting calculations such as the solar position, sidereal time, orbit updating, etc. are developed. Because of the complexities of the mean torque and gyroscopic equations, the precessional analysis is most useful when embodied in suitable computer programs.

VII. ACKNOWLEDGMENTS

The author acknowledges his indebtedness to the following contributors as indicated: to J. F. Ossanna for orbital updating techniques specifically, and in general for helpful discussions concerning computer techniques useful for the calculations; to J. S. Courtney-Pratt, J. W. McLaughlin, and J. H. Hett for the development of techniques and equipment to record the optical flashes at the Holmdel Laboratories, and for their splendid cooperation throughout this project; to W. C. Jakes, Jr., L. R. Lowry, W. E. Legg, R. H. Brandt, P. E. Pheffer, J. W. Barto, and E. L. Frantsvog for providing and staffing Holmdel facilities to acquire and track the Telstar satellites and for useful work in ascertaining the attitude from antenna pattern measurements; to D. W. Hill for initially developing the concept of reducing attitude from mirror flashes and solar sensor data; to R. C. Chapman for patiently reading, correcting and rendering advice on this manuscript; to R. H. Shennum for close cooperation, advice and aid during the period of attitude correction; and to I. Welber for helpful integration of this work into the Telstar satellite communications experiment.

APPENDIX A

Derivation of Equation (40)[21]

Beginning with (39) of the main body of the paper, it is certainly evident that

$$\mathbf{k}\cdot(\mathbf{i}_g\mathbf{i}_g + \mathbf{j}_g\mathbf{j}_g) \times \mathbf{k} = (\mathbf{k}\cdot\mathbf{i}_g)(\mathbf{i}_g \times \mathbf{k}) + (\mathbf{k}\cdot\mathbf{j}_g)(\mathbf{j}_g \times \mathbf{k}). \tag{39a}$$

It will be shown that the following identity exists

$$(\mathbf{k}\cdot\mathbf{i}_g)(\mathbf{i}_g \times \mathbf{k}) + (\mathbf{k}\cdot\mathbf{j}_g)(\mathbf{j}_g \times \mathbf{k}) + (\mathbf{k}\cdot\mathbf{k}_g)(\mathbf{k}_g \times \mathbf{k}) \equiv 0. \tag{39b}$$

Rewriting (39b), one obtains

$$[(\mathbf{k}\cdot\mathbf{i}_g)\mathbf{i}_g + (\mathbf{k}\cdot\mathbf{j}_g)\mathbf{j}_g + (\mathbf{k}\cdot\mathbf{k}_g)\mathbf{k}_g] \times \mathbf{k} \equiv 0 \tag{39c}$$

and

$$(\mathbf{k}\cdot\mathbf{i}_g) = (x_g\mathbf{i}_g + y_g\mathbf{j}_g + z_g\mathbf{k}_g)\cdot\mathbf{i}_g \tag{39d}$$

$$= x_g \tag{39e}$$

where

x_g, y_g, z_g = components of **k** along the ORDEF axes.

In like fashion

$$(\mathbf{k}\cdot\mathbf{j}_g) = y_g \tag{39f}$$

$$(\mathbf{k}\cdot\mathbf{k}_g) = z_g\,. \tag{39g}$$

Substituting (39e,f,g) into (39c), we obtain

$$(x_g\mathbf{i}_g + y_g\mathbf{j}_g + z_g\mathbf{k}_g) \times \mathbf{k} \equiv 0 \tag{39h}$$

$$\mathbf{k} \times \mathbf{k} = 0 \tag{39i}$$

$$0 \equiv 0. \tag{39j}$$

Therefore

$$-(\mathbf{k}\cdot\mathbf{k}_g)(\mathbf{k}_g \times \mathbf{k}) \equiv \mathbf{k}\cdot(\mathbf{i}_g\mathbf{i}_g + \mathbf{j}_g\mathbf{j}_g) \times \mathbf{k}, \tag{39k}$$

and (40) follows from (39).

APPENDIX B

Evaluation of the II Integral of Equation (73)[21]

Copying the II integral from (73) of the main body of the paper, we have

$$\mathrm{II} = (\cos\beta)\mathbf{k} \times \int_0^{2\pi} \{\mathbf{K} - 3(\mathbf{i}_s\cdot\mathbf{K})\mathbf{i}_s\}\,(1 + e\cos v)\,dv\,. \tag{74}$$

Split (74) into two parts by letting

$$\begin{aligned}\mathrm{IIA} &= \cos\beta\mathbf{k} \times \mathbf{K} \int_0^{2\pi} (1 + \cos v)\,dv \\ &= \cos\beta\mathbf{K} \times \int_0^{2\pi} (1 + e\cos\omega\cos P + e\sin\omega\sin P)\,d\omega \\ &= (2\pi\cos\beta)\mathbf{k} \times \mathbf{K}\end{aligned} \tag{75}$$

and

$$\begin{aligned}\mathrm{IIB} = (-3\cos\beta)\mathbf{k} \times \mathbf{K}\cdot \int_0^{2\pi} \mathbf{i}_s\mathbf{i}_s(1 &+ e\cos\omega\cos P \\ &+ e\sin\omega\sin P)\,d\omega\,.\end{aligned} \tag{76}$$

But

$$\mathbf{i}_s = \cos\omega\mathbf{i}_g + \sin\omega\mathbf{j}_g \tag{77}$$

therefore

$$\mathbf{i}_s\mathbf{i}_s = \cos^2 \omega \mathbf{i}_g\mathbf{i}_g + \sin^2 \omega \mathbf{j}_g\mathbf{j}_g + \sin \omega \cos \omega(\mathbf{i}_g\mathbf{j}_g + \mathbf{j}_g\mathbf{i}_g). \qquad (78)$$

It follows that

$$\mathrm{IIB} = (-3\pi \cos\beta)\mathbf{k} \times \mathbf{K}\cdot(\mathbf{i}_g\mathbf{i}_g + \mathbf{j}_g\mathbf{j}_g). \qquad (79)$$

So that, combining IIA and IIB

$$\mathrm{II} = (\pi \cos\beta)\mathbf{k} \times [2\mathbf{K} - 3\mathbf{K}\cdot(\mathbf{i}_g\mathbf{i}_g + \mathbf{j}_g\mathbf{j}_g)] \qquad (80)$$

$$= (\pi \cos\beta)\mathbf{k} \times [2\mathbf{K} - 3(\mathbf{K}\cdot\mathbf{j}_g)\mathbf{j}_g] \qquad (81)$$

$$= \pi \cos\beta[2\mathbf{k} \times \mathbf{K} - 3(\mathbf{K}\cdot\mathbf{j}_g)(\mathbf{k}X\mathbf{j}_g)]. \qquad (82)$$

We shall now proceed to express (82) in the SANOR system. From Fig. 10,

$$\mathbf{K} = \sin \theta\mathbf{j} + \cos \theta\mathbf{k}. \qquad (83)$$

From (22) and (20) we write

$$x_g = (CD)_i(\tilde{D}\tilde{C})x \qquad (84)$$

or

$$\begin{aligned}\mathbf{j}_g = {} & (\sin \psi \cos \Omega \cos i - \sin \Omega \cos \psi \cos i)\mathbf{i} \\ & + (\sin i \sin \theta + \sin \Omega \cos i \sin \psi \cos \theta \\ & + \cos \Omega \cos i \cos \psi \cos \theta)\mathbf{j} \qquad (85) \\ & + (\sin i \cos \theta - \sin \Omega \cos i \sin \psi \sin \theta \\ & - \cos \Omega \cos i \cos \psi \sin \theta)\mathbf{k}\end{aligned}$$

or, simplifying

$$\begin{aligned}\mathbf{j}_g = {} & [-\cos i \sin (\Omega - \psi)]\mathbf{i} \\ & + [\sin i \sin \theta + \cos i \cos \theta \cos (\Omega - \psi)]\mathbf{j} \qquad (86) \\ & + [\sin i \cos \theta - \cos i \sin \theta \cos (\Omega - \psi)]\mathbf{k}.\end{aligned}$$

Substitute (83) and (86) into (82) to obtain

$$\begin{aligned}\mathrm{II} = {} & \pi \cos \beta\Big[(-2 \sin \theta)\mathbf{i} - 3[\{\sin i \sin^2 \theta \\ & + \cos i \sin \theta \cos \theta \cos (\Omega - \psi) + \sin i \cos^2 \theta \\ & - \cos i \sin \theta \cos \theta \cos (\Omega - \psi)\}\{(-\sin i \sin \theta \qquad (87) \\ & - \cos i \cos \theta \cos (\Omega - \psi))\mathbf{i} \\ & + (-\cos i \sin (\Omega - \psi))\mathbf{j}\}]\Big].\end{aligned}$$

Finally,

$$\begin{aligned}\mathrm{II} = \pi \cos \beta[\{ -2 \sin \theta + 3 \sin^2 i \sin \theta \\ + 3 \sin i \cos i \cos \theta \cos (\Omega - \psi)\}\mathbf{i} \\ + 3 \sin i \cos i \sin (\Omega - \psi)\mathbf{j}].\end{aligned} \tag{88}$$

APPENDIX C

Case I Expansion

Refer to integrals of (91), (92), (93), and (94) of the main body of the paper:

$$A = \int_0^{2\pi} \sin \eta_0(1 + e \cos v)\ dv = \int_0^{2\pi} \sin \eta_0\, dv + e \int_0^{2\pi} \sin \eta_0 \cos v\, dv \tag{109}$$

$$\begin{aligned}&= \sin \eta_0 \int_0^{2\pi} \cos (bv)\ dv + \cos \eta_0 \int_0^{2\pi} \sin (bv)\ dv \\ &\quad + e\left[\sin \eta_0 \int_0^{2\pi} \cos (bv) \cos v\ dv \right. \\ &\quad \left. + \cos \eta_0 \int_0^{2\pi} \sin (bv) \cos v\ dv\right]\end{aligned} \tag{110}$$

$$= \sin \eta_0(D_1 + eD_3) + \cos \eta_0(D_2 + eD_4) \tag{111}$$

$$= \left(\frac{1}{b} - \frac{eb}{1 - b^2}\right) [\sin \eta_0 \sin 2\pi b + \cos \eta_0(1 - \cos 2\pi b)] \tag{112}$$

where the D factors are listed in Appendix E. In similar manner,

$$B = \cos \eta_0(D_1 + eD_3) - \sin \eta_0(D_2 + eD_4) \tag{113}$$

$$= \left[\frac{1}{b} - \frac{eb}{1 - b^2}\right] [\cos \eta_0 \sin 2\pi b - \sin \eta_0(1 - \cos 2\pi b)]. \tag{114}$$

The **C** integral is expanded as follows

$$\mathbf{C} = \int_0^{2\pi} \mathbf{i}_s\mathbf{i}_s \sin \eta(1 + e \cos v)\ dv \tag{115}$$

but

$$\mathbf{i}_s = \cos \omega\mathbf{i}_g + \sin \omega\mathbf{j}_g\,. \tag{116}$$

Therefore

$$\mathbf{i}_s\mathbf{i}_s = \cos^2 \omega\mathbf{i}_g\mathbf{i}_g + \sin^2 \omega\mathbf{j}_g\mathbf{j}_g + \sin \omega \cos \omega(\mathbf{i}_g\mathbf{j}_g + \mathbf{j}_g\mathbf{i}_g). \tag{117}$$

Let

$$\mathbf{C} = \mathbf{i}_g\mathbf{i}_g C_1 + \mathbf{j}_g\mathbf{j}_g C_2 + (\mathbf{i}_g\mathbf{j}_g + \mathbf{j}_g\mathbf{i}_g)C_3 \tag{118}$$

and evaluate C_1, C_2, C_3 separately.

$$C_1 = \int_0^{2\pi} \cos^2 \omega \sin \eta(1 + e \cos v)\, dv. \tag{119}$$

Letting $\omega = v + P$, expanding $\cos^2 \omega$ and collecting terms, we have

$$\begin{aligned} C_1 = \sin \eta_0[&\cos^2 P(D_5 + eD_7) + \sin^2 P(D_9 + eD_{11}) \\ &- 2 \sin P \cos P(D_{13} + eD_{15})] \\ &+ \cos \eta_0[\cos^2 P(D_6 + eD_8) + \sin^2 P(D_{10} + eD_{12}) \\ &- 2 \sin P \cos P(D_{14} + eD_{16})]. \end{aligned} \tag{120}$$

in like fashion

$$\begin{aligned} C_2 = \sin \eta_0[&\sin^2 P(D_5 + eD_7) + \cos^2 P(D_9 + eD_{11}) \\ &+ 2 \sin P \cos P(D_{13} + eD_{15})] \\ &+ \cos \eta_0[\sin^2 P(D_6 + eD_8) + \cos^2 P(D_{10} + eD_{12}) \\ &+ 2 \sin P \cos P(D_{14} + eD_{16})] \end{aligned} \tag{121}$$

and

$$\begin{aligned} C_3 &= \int_C^{2\pi} \sin \omega \cos \omega \sin \eta(1 + e \cos v)\, dv \\ &= \sin \eta_0[(\cos^2 P - \sin^2 P)(D_{13} + eD_{15}) \\ &\quad + \sin P \cos P\ (D_5 - D_9 + eD_7 - eD_{11})] \\ &\quad + \cos \eta_0[(\cos^2 P - \sin^2 P)(D_{14} + eD_{16}) \\ &\quad + \sin P \cos P\ (D_6 - D_{10} + eD_8 - eD_{12})]. \end{aligned} \tag{122}$$

Since

$$\mathbf{E} = \int_0^{2\pi} \mathbf{i}_s\mathbf{i}_s \cos \eta(1 + e \cos v)\, dv \tag{94}$$

we expand as with $\mathbf{C}$, to yield

$$\mathbf{E} = \mathbf{i}_g\mathbf{i}_g E_1 + \mathbf{j}_g\mathbf{j}_g E_2 + (\mathbf{i}_g\mathbf{j}_g + \mathbf{j}_g\mathbf{i}_g)E_3 \tag{123}$$

so that

$$E_1 = \int_0^{2\pi} \cos^2 \omega \cos \eta (1 + e \cos v)\, dv \tag{124}$$

$$\begin{aligned} &= \cos \eta_0[\cos^2 P(D_5 + eD_7) + \sin^2 P(D_9 + eD_{11}) \\ &\qquad - 2 \sin P \cos P(D_{13} + eD_{15})] \\ &\qquad - \sin \eta_0[\cos^2 P(D_6 + eD_8) + \sin^2 P(D_{10} + eD_{12}) \\ &\qquad - 2 \sin P \cos P(D_{14} + eD_{16})] \end{aligned} \tag{125}$$

and

$$E_2 = \int_0^{2\pi} \sin^2 \omega \cos \eta (1 + e \cos v)\, dv$$

$$\begin{aligned} &= \cos \eta_0[\sin^2 P(D_5 + eD_7) + \cos^2 P(D_9 + eD_{11}) \\ &\qquad + 2 \sin P \cos P(D_{13} + eD_{15})] \\ &\qquad - \sin \eta_0[\sin^2 P(D_6 + eD_8) + \cos^2 P(D_{10} + eD_{12}) \\ &\qquad + 2 \sin P \cos P(D_{14} + eD_{16})] \end{aligned} \tag{127}$$

and

$$E_3 = \int_0^{2\pi} \sin \omega \cos \omega \cos \eta (1 + e \cos v)\, dv \tag{128}$$

$$\begin{aligned} &= \cos \eta_0[(\cos^2 P - \sin^2 P)(D_{13} + eD_{15}) \\ &\qquad + \sin P \cos P(D_5 - D_9 + eD_7 - eD_{11})] \\ &\qquad - \sin \eta_0[(\cos^2 P - \sin^2 P)(D_{14} + eD_{16}) \\ &\qquad + \sin P \cos P(D_6 - D_{10} + eD_8 - eD_{12})]. \end{aligned} \tag{129}$$

APPENDIX D

Case II Expansion

We will now evaluate (91), (92), (93) and (94) once again in a manner similar to Appendix C, but this time using the approximation

$$M = v - (\lambda/b) \sin v. \tag{102}$$

The same notation will be used for the integrals

$$\begin{aligned} A &= \int_0^{2\pi} \sin \eta(1 + e \cos v)\, dv \\ &= \sin \eta_0 \int_0^{2\pi} \{\cos bv \cos (\lambda \sin v) + \sin bv \sin (\lambda \sin v)\} \\ &\qquad \cdot (1 + e \cos v)\, dv \\ &\qquad + \cos \eta_0 \int_0^{2\pi} \{\sin bv \cos (\lambda \sin v) - \cos bv \sin (\lambda \sin v)\} \\ &\qquad \cdot (1 + e \cos v)\, dv. \end{aligned} \tag{143}$$

Using the approximations of (107) and (108), we have

$$\begin{aligned} A = {} & \sin \eta_0[D_1 + eD_3 + \lambda(D_{17} + eD_{14}) - (\lambda^2/2)(D_9 + eD_{11}) \\ & - (\lambda^3/6)(D_{18} + eD_{19})] \\ & + \cos \eta_0[D_2 + eD_4 - \lambda(D_{20} + eD_{13}) - (\lambda^2/2) \\ & (D_{10} + eD_{12}) + (\lambda^3/6)(D_{21} + eD_{22})]. \end{aligned} \tag{144}$$

In like manner,

$$B = \int_0^{2\pi} \cos \eta(1 + e \cos v)\, dv \tag{145}$$

$$\begin{aligned} = {} & \cos \eta_0[D_1 + eD_3 + \lambda(D_{17} + eD_{14}) - (\lambda^2/2)(D_9 + eD_{11}) \\ & - (\lambda^3/6)(D_{18} + eD_{19})] \\ & - \sin \eta_0[D_2 + eD_4 - \lambda(D_{20} + eD_{13}) - (\lambda^2/2)(D_{10} \\ & + eD_{12}) + (\lambda^3/6)(D_{21} + eD_{22})]. \end{aligned} \tag{146}$$

Similarly,

$$C_1 = \int_0^{2\pi} \cos^2 \omega \sin \eta(1 + e \cos v)\, dv \tag{147}$$

$$\begin{aligned} C_1 = {} & \sin \eta_0\Big[\cos^2 P\{(D_5 + eD_7) + \lambda(D_{16} + eD_{26}) - (\lambda^2/2)(D_{23} \\ & + eD_{25}) - (\lambda^3/6)(D_{31} + eD_{33})\} + \sin^2 P\{(D_9 + eD_{11}) \\ & + \lambda(D_{18} + eD_{19}) - (\lambda^2/2)(D_{30} + eD_{38}) - (\lambda^3/6) \\ & \cdot (D_{35} + eD_{41})\} - 2 \sin P \cos P\{(D_{13} + eD_{15}) + \lambda(D_{12} \end{aligned}$$

$$\begin{aligned}&+ eD_{24}) - (\lambda^2/2)(D_{22} + eD_{32}) - (\lambda^3/6)(D_{37} + eD_{39})\}\Big]\\ &+ \cos \eta_0\Big[\cos^2 P\{(D_6 + eD_8) - \lambda(D_{15} + eD_{28}) - (\lambda^2/2)\\ &\cdot(D_{24} + eD_{27}) + (\lambda^3/6)(D_{32} + eD_{34})\} + \sin^2 P\{(D_{10}\\ &+ eD_{12}) - \lambda(D_{21} + eD_{22}) - (\lambda^2/2)(D_{29} + eD_{37})\\ &+ (\lambda^3/6)(D_{36} + eD_{42})\} - 2 \sin P \cos P\{(D_{14} + eD_{16})\\ &- \lambda(D_{11} + eD_{23}) - (\lambda^2/2)(D_{19} + eD_{31}) + (\lambda^3/6)(D_{38}\\ &+ eD_{40})\}\Big].\end{aligned} \tag{148}$$

To simplify the writing of the mean magnetic torque equation, let

$$\begin{aligned}F_1 = (D_5 + eD_7) + \lambda(D_{16} + eD_{26}) - (\lambda^2/2)(D_{23} + eD_{25})\\ - (\lambda^3/6)(D_{31} + eD_{33})\end{aligned} \tag{122}$$

$$\begin{aligned}F_2 = (D_9 + eD_{11}) + \lambda(D_{18} + eD_{19}) - (\lambda^2/2)(D_{30} + eD_{38})\\ - (\lambda^3/6)(D_{35} + eD_{41})\end{aligned} \tag{123}$$

$$\begin{aligned}F_3 = (D_{13} + eD_{15}) + \lambda(D_{12} + eD_{24}) - (\lambda^2/2)(D_{22} + eD_{32})\\ - (\lambda^3/6)(D_{36} + eD_{39})\end{aligned} \tag{124}$$

$$\begin{aligned}F_4 = (D_6 + eD_8) - \lambda(D_{15} + eD_{28}) - (\lambda^2/2)(D_{24} + eD_{27})\\ + (\lambda^3/6)(D_{32} + eD_{34})\end{aligned} \tag{125}$$

$$\begin{aligned}F_5 = (D_{10} + eD_{12}) - \lambda(D_{21} + eD_{22}) - (\lambda^2/2)(D_{29} + eD_{37})\\ + (\lambda^3/6)(D_{36} + eD_{42})\end{aligned} \tag{126}$$

$$\begin{aligned}F_6 = (D_{14} + eD_{16}) - \lambda(D_{11} + eD_{23}) - (\lambda^2/2)(D_{19} + eD_{31})\\ + (\lambda^3/6)(D_{38} + eD_{40}).\end{aligned} \tag{127}$$

By the similarity of C_2 and C_3 integrals to C_1 and of E_1, E_2, E_3 to C_1, C_2, C_3 we immediately write and summarize the following:

$$\begin{aligned}C_1 = {}&\sin \eta_0[F_1 \cos^2 P + F_2 \sin^2 P - 2F_3 \sin P \cos P]\\ &+ \cos \eta_0[F_4 \cos^2 P + F_5 \sin^2 P - 2F_6 \sin P \cos P]\end{aligned} \tag{128}$$

$$\begin{aligned}C_2 = {}&\sin \eta_0[F_1 \sin^2 P + F_2 \cos^2 P + 2F_3 \sin P \cos P]\\ &+ \cos \eta_0[F_4 \sin^2 P + F_5 \cos^2 P + 2F_6 \sin P \cos P]\end{aligned} \tag{129}$$

$$\begin{aligned}C_3 = {}&\sin \eta_0[(F_1 - F_2) \sin P \cos P + F_3(\cos^2 P - \sin^2 P)]\\ &+ \cos \eta_0[(F_4 - F_5) \sin P \cos P + F_6(\cos^2 P - \sin^2 P)]\end{aligned} \tag{130}$$

$$\begin{aligned} E_1 = {} & \cos \eta_0[F_1 \cos^2 P + F_2 \sin^2 P - 2F_3 \sin P \cos P] \\ & - \sin \eta_0[F_4 \cos^2 P + F_5 \sin^2 P - 2F_6 \sin P \cos P] \end{aligned} \tag{131}$$

$$\begin{aligned} E_2 = {} & \cos \eta_0[F_1 \sin^2 P + F_2 \cos^2 P + 2F_3 \sin P \cos P] \\ & - \sin \eta_0[F_4 \sin^2 P + F_5 \cos^2 P + 2F_6 \sin P \cos P] \end{aligned} \tag{132}$$

$$\begin{aligned} E_3 = {} & \cos \eta_0[(F_1 - F_2) \sin P \cos P + F_3(\cos^2 P - \sin^2 P)] \\ & - \sin \eta_0[(F_4 - F_5) \sin P \cos P + F_6(\cos^2 P - \sin^2 P)]. \end{aligned} \tag{133}$$

APPENDIX E

The D Integrals[22]

$$D_1 = \int_0^{2\pi} \cos (bv)\, dv = \frac{1}{b} \sin (2\pi b)$$

$$D_2 = \int_0^{2\pi} \sin (bv)\, dv = \frac{1}{b} [1 - \cos (2\pi b)]$$

$$D_3 = \int_0^{2\pi} \cos (bv) \cos v\, dv = -\frac{b \sin (2\pi b)}{1 - b^2}$$

$$D_4 = \int_0^{2\pi} \sin (bv) \cos v\, dv = -\frac{b}{1 - b^2} [1 - \cos (2\pi b)]$$

$$D_5 = \int_0^{2\pi} \cos (bv) \cos^2 v\, dv = \frac{2 - b^2}{b(4 - b^2)} \sin (2\pi b)$$

$$D_6 = \int_0^{2\pi} \sin (bv) \cos^2 v\, dv = \frac{2 - b^2}{b(4 - b^2)} [1 - \cos (2\pi b)]$$

$$D_7 = \int_0^{2\pi} \cos (bv) \cos^3 v\, dv = -\frac{b(7 - b)^2}{(1 - b^2)(9 - b^2)} \sin (2\pi b)$$

$$D_8 = \int_0^{2\pi} \sin (bv) \cos^3 v\, dv = -\frac{b(7 - b^2)}{(1 - b^2)(9 - b^2)} [1 - \cos (2\pi b)]$$

$$D_9 = \int_0^{2\pi} \cos (bv) \sin^2 v\, dv = \frac{2 \sin (2\pi b)}{b(4 - b^2)}$$

$$D_{10} = \int_0^{2\pi} \sin (bv) \sin^2 v\, dv = \frac{2}{b(4 - b^2)} [1 - \cos (2\pi b)]$$

$$D_{11} = \int_0^{2\pi} \sin^2 v \cos v \cos (bv)\, dv = -\frac{2b \sin (2\pi b)}{(1 - b^2)(9 - b^2)}$$

$$D_{12} = \int_0^{2\pi} \sin^2 v \cos v \sin (bv)\, dv = \frac{-2b}{(1 - b^2)(9 - b^2)} [1 - \cos (2\pi b)]$$

$$D_{13} = \int_0^{2\pi} \sin v \cos v \cos (bv)\, dv = \frac{1 - \cos (2\pi b)}{4 - b^2}$$

$$D_{14} = \int_0^{2\pi} \sin v \cos v \sin (bv)\, dv = -\frac{\sin (2\pi b)}{4 - b^2}$$

$$D_{15} = \int_0^{2\pi} \sin v \cos^2 v \cos (bv)\, dv = \frac{(3 - b^2)}{(9 - b^2)(1 - b^2)} [1 - \cos (2\pi b)]$$

$$D_{16} = \int_0^{2\pi} \sin v \cos^2 v \sin (bv)\, dv = -\frac{(3 - b^2) \sin (2\pi b)}{(9 - b^2)(1 - b^2)}$$

$$D_{17} = \int_0^{2\pi} \sin v \sin (bv)\, dv = -\frac{\sin (2\pi b)}{1 - b^2}$$

$$D_{18} = \int_0^{2\pi} \sin^3 v \sin (bv)\, dv = -\frac{6 \sin (2\pi b)}{(9 - b^2)(1 - b^2)}$$

$$D_{19} = \int_0^{2\pi} \sin^3 v \cos v \sin (bv)\, dv = -\frac{6 \sin (2\pi b)}{(16 - b^2)(4 - b^2)}$$

$$D_{20} = \int_0^{2\pi} \sin v \cos (bv)\, dv = \frac{1 - \cos (2\pi b)}{1 - b^2}$$

$$D_{21} = \int_0^{2\pi} \sin^3 v \cos (bv)\, dv = \frac{6[1 - \cos (2\pi b)]}{(9 - b^2)(1 - b^2)}$$

$$D_{22} = \int_0^{2\pi} \sin^3 v \cos v \cos (bv)\, dv = \frac{6[1 - \cos (2\pi b)]}{(16 - b^2)(4 - b^2)}$$

$$D_{23} = \int_0^{2\pi} \sin^2 v \cos^2 v \cos (bv)\, dv = \frac{2 \sin (2\pi b)}{b(16 - b^2)}$$

$$D_{24} = \int_0^{2\pi} \sin^2 v \cos^2 v \sin (bv)\, dv = \frac{2[1 - \cos (2\pi b)]}{b(16 - b^2)}$$

$$D_{25} = \int_0^{2\pi} \sin^2 v \cos^3 v \cos (bv)\, dv = -\frac{2b(13 - b^2) \sin (2\pi b)}{(25 - b^2)(9 - b^2)(1 - b^2)}$$

$$D_{26} = \int_0^{2\pi} \sin v \cos^3 v \sin (bv)\, dv = -\frac{(10 - b^2) \sin (2\pi b)}{(16 - b^2)(4 - b^2)}$$

$$D_{27} = \int_0^{2\pi} \sin^2 v \cos^3 v \sin (bv)\, dv = -\frac{2b(13 - b^2)[1 - \cos (2\pi b)]}{(25 - b^2)(9 - b^2)(1 - b^2)}$$

$$D_{28} = \int_0^{2\pi} \sin v \cos^3 v \cos (bv)\, dv = \frac{(10 - b^2)[1 - \cos (2\pi b)]}{(16 - b^2)(4 - b^2)}$$

$$D_{29} = \int_0^{2\pi} \sin^4 v \sin (bv)\, dv = \frac{24[1 - \cos (2\pi b)]}{b(16 - b^2)(4 - b^2)}$$

$$D_{30} = \int_0^{2\pi} \sin^4 v \cos (bv)\, dv = \frac{24 \sin (2\pi b)}{b(16 - b^2)(4 - b^2)}$$

$$D_{31} = \int_0^{2\pi} \sin^3 v \cos^2 v \sin (bv)\, dv = -\frac{6(5 - b^2) \sin (2\pi b)}{(25 - b^2)(9 - b^2)(1 - b^2)}$$

$$D_{32} = \int_0^{2\pi} \sin^3 v \cos^2 v \cos (bv)\, dv = \frac{6(5 - b^2)[1 - \cos (2\pi b)]}{(25 - b^2)(9 - b^2)(1 - b^2)}$$

$$D_{33} = \int_0^{2\pi} \sin^3 v \cos^3 v \sin (bv)\, dv = -\frac{6 \sin (2\pi b)}{(36 - b^2)(4 - b^2)}$$

$$D_{34} = \int_0^{2\pi} \sin^3 v \cos^3 v \cos (bv)\, dv = \frac{6[1 - \cos (2\pi b)]}{(36 - b^2)(4 - b^2)}$$

$$D_{35} = \int_0^{2\pi} \sin^5 v \sin (bv)\, dv = -\frac{120 \sin (2\pi b)}{(25 - b^2)(9 - b^2)(1 - b^2)}$$

$$D_{36} = \int_0^{2\pi} \sin^5 v \cos (bv)\, dv = \frac{120[1 - \cos (2\pi b)]}{(25 - b^2)(9 - b^2)(1 - b^2)}$$

$$D_{37} = \int_0^{2\pi} \sin^4 v \cos v \sin (bv)\, dv = -\frac{24b[1 - \cos (2\pi b)]}{(25 - b^2)(9 - b^2)(1 - b^2)}$$

$$D_{38} = \int_0^{2\pi} \sin^4 v \cos v \cos (bv)\, dv = -\frac{24b \sin (2\pi b)}{(25 - b^2)(9 - b^2)(1 - b^2)}$$

$$D_{39} = \int_0^{2\pi} \sin^4 v \cos^2 v \sin (bv)\, dv = \frac{24(6 - b^2)[1 - \cos (2\pi b)]}{b(36 - b^2)(16 - b^2)(4 - b^2)}$$

$$D_{40} = \int_0^{2\pi} \sin^4 v \cos^2 v \cos (bv)\, dv = \frac{24(6 - b^2) \sin (2\pi b)}{b(36 - b^2)(16 - b^2)(4 - b^2)}$$

$$D_{41} = \int_0^{2\pi} \sin^5 v \cos v \sin (bv)\, dv = -\frac{120 \sin (2\pi b)}{(36 - b^2)(16 - b^2)(4 - b^2)}$$

$$D_{42} = \int_0^{2\pi} \sin^5 v \cos v \cos (bv)\, dv = \frac{120[1 - \cos (2\pi b)]}{(36 - b^2)(16 - b^2)(4 - b^2)}.$$

REFERENCES

1. Bennett, S. B., and Thomas, L. C., The *Telstar* Communications Satellite Experiment Plan, AIEE Trans., No. 63-952, May 6, 1963, pp. 9–10. See also B.S.T.J., **42,** July, 1963, p. 765.
2. Hill, D. W., Calculation of the Spin-Axis Orientation of the *Telstar* Satellites from Optical Data, B.S.T.J., **42,** Nov., 1963, p. 2943.
3. Courtney-Pratt, J. S., Hett, J. H., and McLaughlin, J. W., Optical Meas-

urements on *Telstar* to Determine the Orientation of the Spin Axis and the Spin Rate, Jour. SMPTE, **72,** June, 1963, pp. 462–484.

4. *The American Ephemeris and Nautical Almanac for the Year 1962*, U. S. Government Printing Office, Washington, D. C., 1960, p. 482.
5. Ref. 4, pp. 466–471.
6. Smart, W. M., *Textbook on Spherical Astronomy*, Cambridge Press, 1956, pp. 116–117.
7. Vestline, E. H., et al., Description of the Earth's Main Magnetic Field and Its Secular Change, 1940–1945, Carnegie Institution of Washington, Publication 578, 1947.
8. Vestline, E. H., et al., The Geomagnetic Field, Its Description and Analysis, Carnegie Institution of Washington, Publication 580, 1959.
9. Hrycak, P., et al., The Spacecraft Structure and Thermal Design Considerations, B.S.T.J., **42,** July, 1963, p. 974.
10. Roberson, R. E., and Tatistcheff, D., The Potential Energy of a Small Rigid Body in the Gravitational Field of an Oblate Spheroid, J. Franklin Inst., **262,** 1956, pp. 209-214.
11. Harnwell, G. P., *Principles of Electricity and Electromagnetism*, 1st ed., McGraw-Hill, 1938, p. 377. For precessional theory similar to that of Section 4.6, but for circular orbits only, see: Bandeen, W. R., and Manger, W. P., Angular Motion of the Spin Axis of the *Tiros I* Meteorological Satellite Due to Magnetic and Gravitational Torques, NASA TN D-571, April, 1961.
12. Allen, C. W., *Astrophysical Quantities*, Athlone Press, 1955.
13. See Ref. 6, pp. 107–108.
14. See Ref. 6, p. 111–113.
15. Danby, J. M. A., *Fundamentals of Celestial Mechanics*, Macmillan, New York, 1962, pp. 128–129.
16. Goldstein, H., *Classical Mechanics*, Addison-Wesley, 1959.
17. Yu, E. Y., Spin Decay, Spin-Precession Damping, and Spin-Axis Drift of the *Telstar* Satellite, B.S.T.J., **42,** Sept., 1963, p. 2169.
18. Jakes, W. C., Participation of the Holmdel Station in the *Telstar* Project, B.S.T.J., **42,** July, 1963, p. 1421.
19. Harris, D. L., *Planets and Satellites*, ed. Kuiper and Middlehurst, Chicago, 1961.
20. Delchamps, T. B., et al., The Spacecraft Test and Evaluation Program, B.S.T.J., **42,** July, 1963, p. 1020 and p. 1024.
21. Phillips, H. B., *Vector Analysis*, John Wiley, 1933.
22. Pierce, B. O., *A Short Table of Integrals*, 4th ed., Ginn, 1957.

Directional Control in Light-Wave Guidance

By S. E. MILLER

(Manuscript received May 11, 1964)

The transmission of light waves for communication in a medium sheltered from atmospheric effects requires wave guidance providing frequent changes in direction of propagation. This paper shows that, in any electromagnetic waveguide having transverse planes in which the field is essentially equiphase, the transverse width of the field distribution $2a$ *and wavelength* λ *determine the order of magnitude of the direction-determining parameters,* R_{min}, *the minimum bending radius, and* δ_{max}, *the maximum abrupt angular changes, according to the relations*

$$R_{min} = 2(a^3/\lambda^2)$$

$$\delta_{max} = \tfrac{1}{2}(\lambda/a)$$

which are valid in the region $\lambda < a$. *The significance of* R_{min} *is apparent, with the note that in a system containing a multiplicity of bends, an appropriate way of summing the effects of the individual bends should be used to establish an over-all equivalent bend radius for the complete transmission path, which must be larger than* R_{min}. *The quantity* δ_{max} *may be regarded as the maximum value of the accumulated angular errors (rms sum, for example) in a transmission line including reflecting or refracting elements for directional control. For a light beam at* $\lambda = 0.6328$ *microns having a diameter of 1.0 mm,* $\delta_{max} = 0.036°$ *and* $R_{min} = 600$ *meters.*

Small-diameter beams ease the problem of directional control. There is no fundamental reason why small beams should not be achievable with low loss in the straight condition, but many guiding structures do have an inverse relation between beam diameter and straight-condition attenuation coefficient. To explore the direction-controlling properties of specific media and the interaction of R_{min} *and* δ_{max} *with straight attenuation coefficient, the following waveguides and associated criteria for establishing* R_{min} *and* δ_{max} *were studied:*

(1) sequence of lenses: criterion, beam deflection from nominal axis by one beam radius,

(2) hollow dielectric waveguide: criterion, added bend loss equal to straight condition loss,

(3) round metallic circular-electric waveguides: for helix guide, criterion is bend loss equal to straight loss; for simple metallic tube, criterion is a transmission ripple (due to mode conversion) of about 1.7 db.

In all cases the functional dependence on a and λ for R_{min} or δ_{max} was the same (given above) as derived for the generalized electromagnetic waveguide, and the associated constants were in most cases of similar magnitude.

I. INTRODUCTION

In research on techniques for transmitting light waves over appreciable distances for communication it has become evident that control of direction of propagation is an important and difficult problem. Electromagnetic waves in free space travel in a straight line. In a medium that is sheltered from atmospheric effects, frequent changes in direction are necessary to follow vertical terrain contours and to conform to a horizontal path avoiding physical obstacles and regions of high-cost installation. The wave guiding medium must provide these direction changes.

In this paper some simple relations are derived to give the order of magnitude of the direction-determining factors, bending radius and abrupt tilt angle, for any wave guiding structure as a function of wavelength and the transverse dimension of the guided electromagnetic wave beam. These simple relations are then compared to the corresponding more precisely defined quantities for specific waveguides: (1) a sequence of lenses,[1] (2) the hollow-dielectric waveguide,[5] and (3) round waveguides for circular electric waves.

II. DERIVATION OF GENERAL WAVEGUIDE DIRECTIONAL SENSITIVITY

In Fig. 1 we show a generalized waveguide for electromagnetic waves, with an abrupt open end radiating into free space. We assume the field at the aperture is essentially equiphase, which implies ending the guide

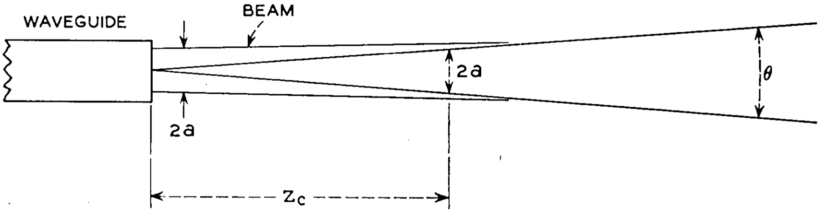


Fig. 1 — Waveguide with abrupt open end.

only at certain longitudinal locations if a periodic form of guidance (such as a sequence of lenses) is employed. Let the field strength variation across the aperture be approximately sinusoidal. Then, approximately, the far-field beam angle θ is

$$\theta = \lambda/a \text{ radians} \tag{1}$$

in which we require $a > \lambda$. Other aperture distributions would give the same order of magnitude for θ. In the near-field region the radiated beam remains collimated in a width approximately $2a$ out to a distance z_c from the aperture, where

$$z_c\theta = 2a \tag{2}$$

$$z_c = 2a^2/\lambda. \tag{2a}$$

The key inference on directional sensitivity is introduced here. Since in the absence of the guide the beam remains confined to essentially the same region as in the presence of the guide, it is concluded that the guide has little influence on the beam over the interval z_c. Thus any appreciable change in direction of wave propagation must not be made in a distance less than z_c.

With reference to Fig. 2, the departure of a circular arc from the tangent is

$$\Delta = \tfrac{1}{2}(l^2/R). \tag{3}$$

We now require that $\Delta = a$ when $l = z_c$. Using (3), (2) and (1), we obtain the minimum bend radius $R_{\min}$

$$R_{\min} = 2a^3/\lambda^2. \tag{4}$$

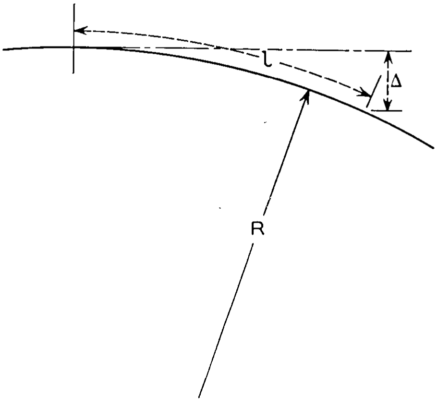


Fig. 2 — Departure of a circular arc from the tangent.

Alternatively, this same relation may be arrived at by specifying that the change in direction shall be one beamwidth θ after traveling a distance z_c in the minimum bending radius $R_{\min}$; i.e., $R_{\min}\theta = z_c$.

Equation (4) gives the order of magnitude of bend radius at which the wave propagation will change character. At longer bend radii the wave propagation will be essentially as in the straight guide, and at shorter bend radii something drastic will happen. Just what changes occur in the latter case depend on the nature of the medium in detail. If the medium is enclosed in a perfect conductor the change will be large mode conversion. If it consists of a sequence of infinitely wide lenses we will see that the change is a wide oscillation of the beam about the nominal axis of propagation. Note that in neither of these cases is energy lost due to the bend. Nonetheless, we regard either change as undesirable.

Consider an abrupt angular change in the guide direction, δ. Following a line of reasoning analogous to that given above, we can say that the character of wave propagation will change rapidly in the region where

$$\delta_{\max} = \theta/2 = \lambda/2a. \tag{5}$$

Smaller values of δ will cause progressively less change in wave character, whereas larger values of δ will cause violent changes.

If we consider the relation of these quantities to a wave guiding medium, it is apparent that $R_{\min}$ is intended as the smallest radius at which the otherwise uniform medium can be bent. When a multiplicity of bends is included in a single transmission link, some way of summing their effects is needed to form an equivalent bending radius which must be greater than $R_{\min}$.

The angle δ is somewhat different. In many media where $a \gg \lambda$ it is possible to insert a large plane reflector and introduce a change of direction of arbitrary size. As long as the guides at both approaches to the reflector are perfectly aligned according to geometric optics, the disturbance on wave propagation may be negligible. However, there will be an error in such angular alignment and $\delta_{\max}$ tells us how large that error may be. When many random angular errors are made, $\delta_{\max}$ is approximately the rms accumulation of such errors.

The numerical values of $R_{\min}$ and $\delta_{\max}$ have been plotted for λ from 0.6328 to 10 microns and beam radius a from 0.1 to 100 millimeters in Figs. 3 and 4, respectively. For example, at $\lambda = 0.6328$ microns and $2a = 1.0$ cm, $R_{\min} = 600{,}000$ meters and $\delta_{\max} = 3.6 \times 10^{-3}$ degrees. Dropping to $2a = 1.0$ mm, $R_{\min} = 600$ meters and $\delta_{\max} = 3.6 \times 10^{-2}$ degrees.

We consider next certain specific wave guiding structures to compare

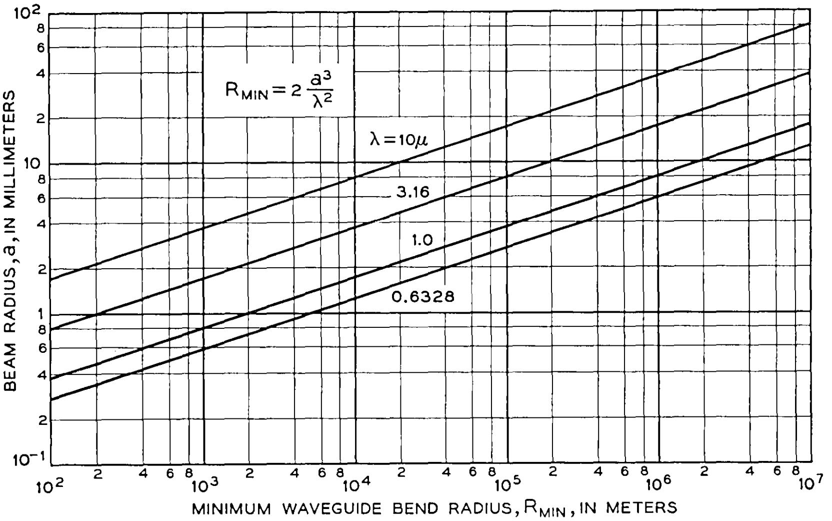


Fig. 3 — Beam radius vs minimum waveguide bend radius.

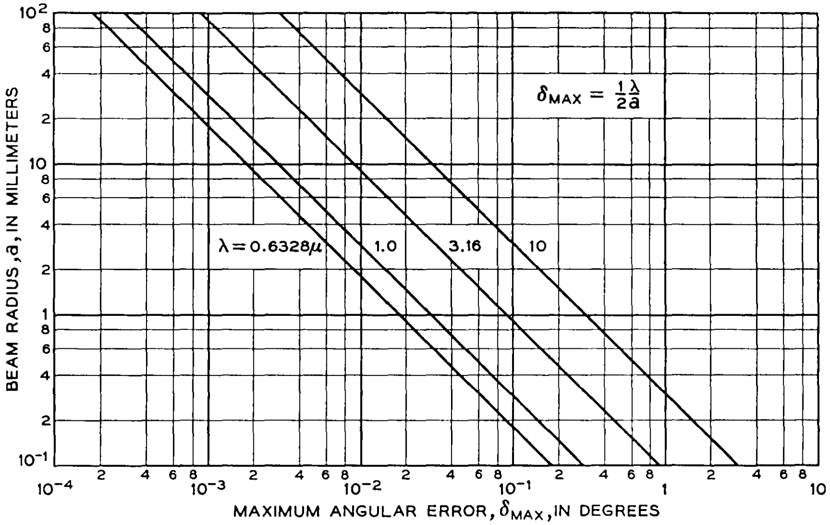


Fig. 4 — Beam radius vs maximum angular error.

the results of directional changes in those structures to the generalized conclusions drawn above.

III. SEQUENCE-OF-LENS WAVEGUIDE

G. Goubau has proposed[1] a waveguide for electromagnetic waves consisting of a series of lenses, and D. Marcuse has used geometric optics to determine the effects of bends in such a waveguide.[2]

If the input to a lens waveguide is a ray which is inclined at an angle δ to the longitudinal waveguide axis (Fig. 5) the departure of the ray from the longitudinal axis has a magnitude at successive lenses which is contained within an envelope which is a sinusoidal function of distance along the longitudinal axis. Starting with the work of Marcuse, one can show that there is an optimum strength of lens which minimizes the departure of a ray from the axis; the optimum focal length f is related to the lens spacing L by

$$2f \doteq L \tag{6}$$

and under that condition the maximum deviation of the ray from the longitudinal axis is

$$r_{\text{max}-1} = \delta L. \tag{7}$$

Consider a region of bend radius R following a straight region of lens waveguide. For a ray incident on the curved region from the axis of the straight region Marcuse has also calculated the ray's departure from the axis in the curved region; for the case $f = L/2$ the maximum departure is

$$r_{\text{max}-2} = L^2/R. \tag{8}$$

We now relate these departures from the guide axis to the transverse dimension of the beam. It is convenient to consider the beam radius to be that value of radius beyond which a completely negligible amount of

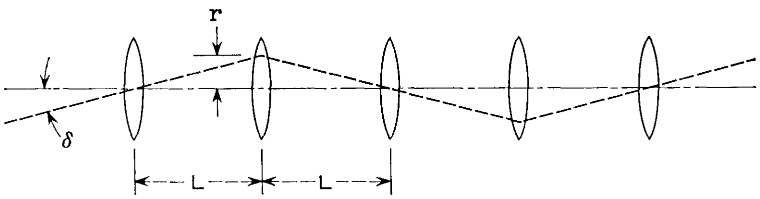


Fig. 5 — Sequence of lenses.

field exists. We define this beam radius as

$$r_b = (N_0 L\,\lambda)^{\frac{1}{2}}. \tag{9}$$

Here N_0 is the Fresnel number which previous work[3,9] shows is on the order of unity for negligibly small diffraction loss when all energy outside the radius r_b is absorbed at the lenses.*

As a criterion of the maximum permissible abrupt angular change, we somewhat arbitrarily set it to be that angle at which $r_{\text{max}-1}$ is equal to the transverse beam radius r_b :

$$\delta_{\text{max}} = (r_b/L) = (N_0\lambda/L)^{\frac{1}{2}} = N_0\lambda/r_b\,. \tag{10}$$

Since $N_0 \cong 1$, this specific guide and criterion gives a permissible angular change of twice that prescribed by (5). This may be considered an excellent agreement.

As a criterion of the minimum permissible bend radius, we set the resulting beam deflection $r_{\text{max}-2}$ equal to the transverse beam radius r_b :

$$R_{\text{min}} = L^2/r_b = r_b^3/(N_0\lambda)^2. \tag{11}$$

Since $N_0 \cong 1$, this specific guide and criterion gives a permissible bend radius of one-half that prescribed by (4), which again may be considered excellent agreement.

The most important aspect of the comparison between (4) and (5), (10), and (11) is that the corresponding equations have the identical dependence on λ and a, which determines in a broad way the magnitude of the direction determining parameters.

In this form of guide we can readily relate the beam radius to the associated lens spacing and the losses. Fig. 6 shows the lens spacing L versus beam radius in the 0.5- to 4-micron wavelength region. As before, N_0 will be about unity, but where extremely low losses per lens are required may have to be slightly greater than unity.[3] For the 1.0-mm beam diameter referred to above, the lens spacing is about 0.4 meter for $\lambda = 0.63$ microns.

In principle, vanishingly small transmission loss could be obtained by appropriate choice of lens diameter (i.e., choice of N_0), if the reflection, absorption, and scattering losses were negligible at the lenses. In practice, such losses may be very real. Fig. 7 shows the total losses per lens required as a function of lens spacing with net transmission loss as a parameter. For 3 db/mile net loss and the 0.4-meter lens spacing, a power loss per lens of about one part in 10^4 is required.

* Further discussion of N_0 and r_b is given in the Appendix.

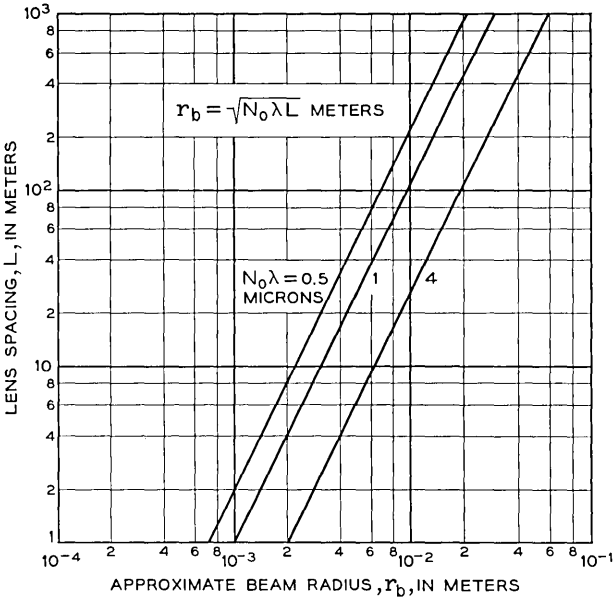


Fig 6. — Approximate beam radius vs lens spacing in a sequence of identical lenses.

In general, of course, the smaller beam diameters which permit rapid direction changes require more tight guidance (closer lens spacing) and tend to increase the losses. Note, however, that the only inherent losses associated with tight guidance for the lens guidance system are due to scattering or reflection at lens surfaces or bulk lens absorption loss, both of which may conceivably be made very small.

IV. HOLLOW DIELECTRIC WAVEGUIDE

E. A. J. Marcatili and R. A. Schmeltzer have proposed a waveguide for light waves consisting of a hollow dielectric tube in which the useful energy is entirely confined to the central hole.[5] When the guide is straight, loss takes place through very slow radiation into the dielectric, which is completely absorbing for the light energy.

The bending radius for such a guide which makes the extra loss due to bending (also a radiation loss) exactly equal to the straight-guide attenuation coefficient has also been determined.[5] They find, for the lowest-order mode (EH_{11}), in which the field varies roughly cosinusoidally from the axis to the inner wall of the tube,

$$R_{\min-D} = 9.5\, a^3/\lambda^2 \tag{12}$$

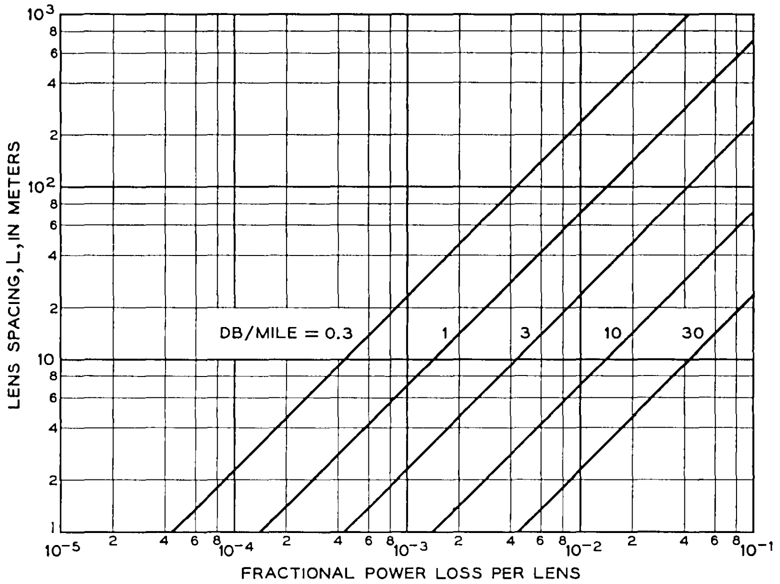


Fig. 7 — Lens loss vs spacing for prescribed total loss.

where a is the inner radius of the tube. Once again, the functional dependence of $R_{\min-D}$ on a and λ is identical to that in (4).

The straight-line attenuation coefficient for the lowest-order mode can be reduced to[5]

$$\alpha_s = 0.214\,\lambda^2/a^3 \tag{13}$$

for an index of refraction of the dielectric tube equal to 1.5. As Marcatili and Schmeltzer have pointed out, the dependence of α_s on a and λ is the exact inverse of that for $R_{\min-D}$; hence for any prescribed straight-guide loss there is a minimum permissible bending radius, which for the lowest-order mode is

$$R_{\min-D} = 2.03/\alpha_s\,. \tag{14}$$

For fixed α_s this is independent of λ. For $R_{\min-D} = 1000$ meters, $\alpha_s = 0.002$ nepers/meter or 27.9 db/mile, and at $\lambda = 0.6328$ microns, the inner radius of the tube $a = 0.35$ mm.

V. CIRCULAR ELECTRIC WAVEGUIDES

We consider now the directional control relations for round waveguides designed for the TE_{01} circular electric wave. Helix waveguide and smoothwalled metallic waveguide involve quite different criteria for tolerable bending radius and will be discussed separately.

Consider first metallic guides in which the losses are negligible compared to the mode coupling coefficients. We assume the degeneracy between TE_{01} and TM_{11} is broken with a dielectric lining or other guide modification. Then the limit on bending radius or abrupt tilt is the interfering effect between the unperturbed TE_{01} energy and the energy which is converted to an undesired mode and reconverted back to the TE_{01} wave.

For an abrupt tilt, the amplitude conversion coefficient from TE_{01} to TE_{12} was found by S. P. Morgan[4] to be, approximately

$$p = 1.935\ (a/\lambda)\delta \tag{15}$$

where δ is the tilt angle in radians. When converted energy from one tilt strikes another tilt (presumed for simplicity here to be the same angle), energy is reconverted back to the TE_{01} wave with the same conversion coefficient given by (15). The amplitude of the reconverted wave compared to the unperturbed wave is then p^2. Depending on the relative phase of the reconverted vector, it may add at any phase angle to the unperturbed wave. Hence the amplitude transmission coefficient varies with wavelength between $(1 + p^2)$ and $(1 - p^2)$. Letting a transmission fluctuation of 1.7 db, corresponding to $p^2 = 0.1$, be the criterion of limiting tilt angle, we find

$$\delta_{\max - M} = p\lambda/1.935a = 0.164\ \lambda/a. \tag{16}$$

Comparing this to the generalized relation for $\delta_{\max}$ in (5), we note the identical dependence on λ and a and a somewhat smaller constant multiplier. In practice, the existence of coupling to other modes would tend to make somewhat smaller values of $\delta_{\max - M}$ needed, but the dependence on λ and a would not be affected.

Still considering guides with negligible losses, we examine the effect of a constant-radius bend. It may be shown that the reconverted vector has the magnitude

$$p_1^2 = k_t^2/(\Gamma_1 - \Gamma_2)^2 \tag{17}$$

where k_t is the distributed coupling coefficient between TE_{01} and an undesired mode and Γ_1 and Γ_2 are the propagation constants for TE_{01} and the undesired modes, respectively. For TE_{01} to TE_{12}

$$k_t \cong j(2a/\lambda_0 R) \tag{18}$$

where R is the bend radius and a is the radius of the guide. Also,

$$(\Gamma_1 - \Gamma_2) \cong j(\beta_1\beta_2) = j\ \lambda_0/a^2. \tag{19}$$

Hence

$$p_1^2 = 4a^6/R^2\lambda^4. \tag{20}$$

As a criterion for minimum bend radius we set p_1^2 equal to 0.1, giving the same ripple in transmission loss as noted above, with the resulting bending radius from (20)

$$R_{\mathrm{min}-M} = (4/p^2)^{\frac{1}{2}}(a^3/\lambda^2) = 6.3\, a^3/\lambda^2. \tag{21}$$

Comparing (21) to (4), we again find the identical dependence on a and λ with a slightly different constant.

Turning now to the case of helix waveguide in which very strong loss is introduced for the undesired mode, we find we do not have explicit forms for the coupling coefficient. We take advantage of some numerical evaluations carried out in the 30- to 100-kmc region on guide varying in diameter from 0.25 inch to 3 inches. It was found that the bend loss coefficient is given by the expression*

$$\alpha_B = 0.0726\ (a^3/R^2\lambda^{2.7}). \tag{22}$$

The TE_{01} loss of the guide when straight is very nearly that of a copper cylinder, given by S. A. Schelkunoff as[6]

$$\alpha_s = 4.46 \times 10^{-6}\, \lambda^{\frac{3}{2}}/a^3. \tag{23}$$

In (23) we have assumed $a^2 \gg (\lambda/2)^2$, so that the cutoff effect is negligible. As our criterion for minimum bending radius we equate the bend loss α_B and the straight-line loss α_s, yielding

$$R_{\mathrm{min}-H} = 128\, a^3/\lambda^{2.1}. \tag{24}$$

The functional dependence of $R_{\mathrm{min}-H}$ on λ and a is very nearly the same as in (4), but the constant multiplier is much greater. This is a consequence of the criterion $\alpha_s = \alpha_B$, which is thereby proven much more stringent than the rather lax transmission ripple criterion used above for the metallic tube guide in which dissipation was negligible. Since α_s of (23) and $R_{\mathrm{min}-H}$ of (24) have different dependence on λ, a change of wavelength will influence $R_{\mathrm{min}-H}$ even though α_s is held constant. We can express this by substituting (23) into (24), giving

$$R_{\mathrm{min}-H} = \frac{128}{\lambda^{0.6}} \times \frac{a^3}{\lambda^{\frac{3}{2}}} = \frac{5.71 \times 10^{-4}}{\lambda^{0.6}\alpha_s}. \tag{25}$$

At longer wavelengths, *smaller* bending radii are tolerable even though

* This is the result of unpublished calculations by the author, based on coupling coefficients derived by methods due to Unger[9] and using coupled-wave theory.[10]

α_s is held constant by increasing a. At a wavelength of 5 mm and a straight attenuation coefficient of 1 db/mile, $R_{\min-H} = 191$ meters. This result and the numerical constant in (22) are dependent to some extent on the wall impedance to the undesired modes used in the numerical evaluations referred to above, which was on the order of one-half the free-space intrinsic impedance.

VI. CONCLUSION

For any guided electromagnetic wave, the order of magnitude of the direction-determining parameters $R_{\min}$ (the minimum bending radius) and $\delta_{\max}$ (the maximum abrupt angular change) are uniquely determined by the wavelength and transverse beam dimension. Equations (4) and (5) were derived, determining $R_{\min}$ and $\delta_{\max}$ for a general guided electromagnetic wave by inferring the tightness of guidance from the behavior of a wave radiated from the open end of the waveguide. Investigation of specific forms of waveguide with precise criteria for setting limits on R and δ (as outlined in the abstract) lead to identical functional forms for $R_{\min}$ and $\delta_{\max}$, with similar constant multipliers.

APPENDIX

Previous workers[3,7] have calculated the diffraction loss at a reflector in a maser interferometer, and the same loss per lens would be expected in a sequence-of-lens waveguide if the entire plane outside the edge of the lens (of radius equal to that of the maser reflector) were absorbing. These losses are plotted versus $N = a^2/L\lambda$ (where a is the reflector radius) in Fig. 3 of Ref. 7 and in Fig. 15 of Ref. 3 for focal length $f = L/2$. We chose N_0 to be that value of N which gives satisfactorily low loss per lens; for example, for $N_0 = 1$, Ref. 7 gives a power loss per lens of one part in 10^4 for the lowest-order wave, and for $N_0 \cong 1.4$ the power loss is one part in 10^6. Fig. 3 of Ref. 1 shows that 99.8 per cent of the energy of the normal mode for infinite lenses lies within the radius $r = (L\lambda)^{\frac{1}{2}}$ at the lens. In practical cases, therefore, N_0 will differ little from unity.

Another item of interest is the relation between r_b of (9) and the field amplitude given by previous workers.[7,8] The field varies as a function of radius r from the axis of the guide according to

$$\exp\left(-r^2/w^2\right) \tag{26}$$

where w is the radius at which the field drops to e^{-1} of its maximum (on axis) value. The value of w varies with longitudinal position between lenses; at midway between lenses $w = w_0$, where

$$w_0 = (L\lambda/2\pi)^{\frac{1}{2}}. \tag{27}$$

At the lenses, $w = w_s$, and for our cases of $f = L/2$

$$w_s = (L\lambda/\pi)^{\frac{1}{2}}. \tag{28}$$

It is apparent from (9) and (28) that

$$r_b = w_s \, (N_0\pi)^{\frac{1}{2}}. \tag{29}$$

In terms of w_s (10) becomes

$$\delta_{\max} = (N_0/\pi)^{\frac{1}{2}}(\lambda/w_s) \tag{30}$$

and (11) becomes

$$R_{\min} = (\pi^{\frac{3}{2}} N_0^{-\frac{1}{2}}) \, (w_s^3/\lambda^2). \tag{31}$$

REFERENCES

1. Goubau, G., and Schwering, F., On the Guided Propagation of Electromagnetic Wave Beams, IRE Trans., **AP-9,** May, 1961, pp. 248–256.
2. Marcuse, D., Propagation of Light Rays through a Lens Waveguide with Curved Axis, B.S.T.J., **43,** March, 1964, pp. 741–753.
3. Fox, A. G., and Li, T., Resonant Modes in a Maser Interferometer, B.S.T.J., **40,** March, 1961, pp. 453–488.
4. Rowe, H. E., and Warters, W. D., Transmission in Multimode Waveguide with Random Imperfections, B.S.T.J., **41,** May, 1962, pp. 1031–1170.
5. Marcatili, E. A. J., and Schmeltzer, R. A., Hollow Dielectric Waveguides for Long Distance Optical Transmission and Lasers, B.S.T.J., this issue, p. 1783.
6. Schelkunoff, S. A., *Electromagnetic Waves*, D. Van Nostrand Co., Princeton, 1945, p. 390.
7. 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.
8. Boyd, G. D., and Kogelnik, H., Generalized Confocal Resonator Theory, B.S.T.J., **41,** July, 1962, pp. 1347–1369.
9. Unger, H. G., Normal Modes and Mode Conversion in Helix Waveguide, B.S.T.J., **40,** January, 1961, pp. 255–280.
10. Miller, S. E., Coupled Wave Theory and Waveguide Applications, B.S.T.J., **33,** May, 1954, pp. 661–719.

Alternating-Gradient Focusing and Related Properties of Conventional Convergent Lens Focusing

By S. E. MILLER

(Manuscript received June 8, 1964)

A series of lenses whose focusing properties are alternately convergent and divergent (alternating-gradient focusing) is of potential interest in the guidance of light waves, and has previously been used to focus electron beams and high-energy particle streams. New information is provided herein on such focusing for the case of equal focal length f (but alternating-gradient) lenses equally spaced a distance L.

The alternating-gradient system formed by adding diverging lenses between the lenses of an all-converging sequence of lenses is found to have the same stability condition as the original system for $0 < L/f < 2$. A physical argument leads to the conclusion that weaker divergent lenses would also leave the stability criterion unchanged.

The focusing effect of the alternating-gradient system is surprisingly close to that of an all-converging lens system. After the focal length of each has been adjusted to an optimum value, the ray departure from the system axis is only 1.67 times as great for the alternating-gradient system as for an all-convergent lens system with the same spacing of convergent lenses.

For weak lenses (i.e. $2f/L \gg 1$) the output ray departure due to input ray displacement is independent of both the focal length and spacing of the lenses, and is independent of lens spacing but proportional to focal length for input ray slope.

Both the alternating-gradient system and all-convergent lens focusing arrangements exhibit discontinuities in the maximum ray displacement versus focal strength relation.

Viewed over-all, alternating-gradient focusing for light guidance does a surprisingly efficient job and may be advantageous over all-convergent lens systems if the alternating-gradient arrangement has structural or economic advantages.

I. INTRODUCTION

In research on guidance of light waves for communication we are considering use of a sequence of lenses of alternately convergent and divergent types. For example, a guidance system using tubular thermal gas lenses might employ continuous flow of gas through a tube whose walls are alternately warmer and cooler than the gas within. Thus, the mechanism used to cause the focusing may have the alternating character, and the question comes to the fore — how well can one focus with such a structure as compared to the use of a sequence of all-convergent lenses?

Alternating-gradient focusing has previously been used on electron beams[1] and on particle accelerators.[2] The present study discovered an error in the previous determination of stability conditions and revealed some little known but interesting properties of alternating-gradient focusing. A comparison is made with conventional focusing using all-convergent lenses.

II. ANALYSIS OF ALTERNATING-GRADIENT SYSTEMS

One might wonder whether a series of equal-focal-length and alternately converging and diverging lenses would give any net focusing at all, since the average dielectric constant along all paths parallel to the axis would be the same. It is well known, however, that a divergent lens followed by a convergent lens of equal focal length, spaced a finite distance less than the focal length, gives a net *converging* lens, and the same is true if the order of the lenses is reversed. Thus, a net focusing is to be expected for an infinite series of such lens pairs.[1]

We consider a sequence of alternating convergent and divergent lenses equally spaced a distance L and of equal focal lengths, f. We follow the method of analysis used by Pierce.[1] There are two cases to cover, one in which the first lens of the array is a divergent lens (obtained by starting at $n = 0$ in Fig. 1) and the other in which the first lens of the array is a convergent lens (obtained by starting at $N = 0$ in Fig. 4, below).

With reference to Fig. 2, and taking the input ray at plane a to have a slope r_a' and a displacement r_a from the longitudinal axis, the output ray from the lens at plane b will be

$$r_b = r_a + Lr_a' \tag{1}$$

$$r_b' = r_a' + (1/f)r_b\,. \tag{2}$$

At plane c this ray will be described by

$$r_c = r_b + Lr_b' \tag{3}$$

$$r_c' = r_b'. \tag{4}$$

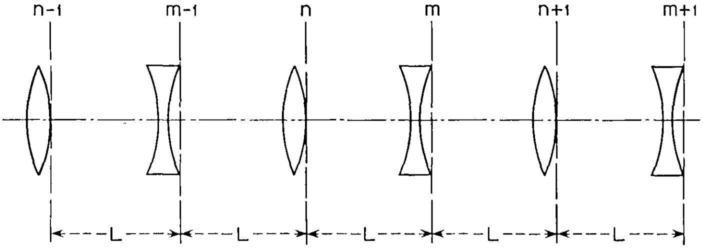


Fig. 1 — Sequence of lenses — first case.

Eliminating r_b and r_b' from (3) and (4) gives

$$r_c = \left(1 + \frac{L}{f}\right) r_a + L\left(2 + \frac{L}{f}\right) r_a' \tag{5}$$

$$r_c' = \left(1 + \frac{L}{f}\right) r_a' + \frac{r_a}{f} . \tag{6}$$

Hence, with reference to Fig. 1, we can write

$$r_{n+1} = \left(1 + \frac{L}{f}\right) r_n + L\left(2 + \frac{L}{f}\right) r_n' \tag{7}$$

$$r_{n+1}' = \left(1 + \frac{L}{f}\right) r_n' + \frac{r_n}{f} - \frac{r_{n+1}}{f} . \tag{8}$$

These two equations lead to

$$r_{n+2} - \left[2 - \left(\frac{L}{f}\right)^2\right] r_{n+1} + r_n = 0. \tag{9}$$

The solution to (9) is

$$r_n = A \cos n\theta + B \sin n\theta \tag{10}$$

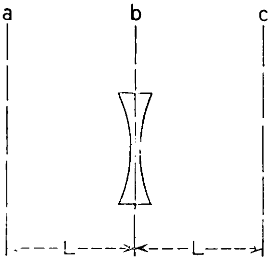


Fig. 2 — Lens subsection for Fig. 1.

where

$$\theta = \cos^{-1}\left[1 - \frac{1}{2}\left(\frac{L}{f}\right)^2\right] \tag{11}$$

and where A and B are constants to be determined.

Equation (11) differs from the corresponding equation on page 200 of Ref. (1), which is believed to be in error. The correct condition for stability, from (11), is

$$0 < \tfrac{1}{2}(L/f)^2 < 2 \tag{12}$$

or

$$0 < L/f < 2.$$

We put in the boundary conditions, at $n = 0$

$$r_n' = r_0' \tag{13}$$

$$r_n = r_0 \,. \tag{14}$$

We make use of a general theorem* stating the orthogonality of the effects of r_0 and r_0' and seek a solution with those quantities as factors. This leads to the following form for r_n, using (10), (7), (13) and (14):

$$r_n = r_0 k_1 \cos(n\theta - \varphi_1) + r_0' L\, k_2 \sin n\theta \tag{15}$$

where

$$k_1 = \left[\frac{2}{1 - (L/2f)}\right]^{\frac{1}{2}} \tag{16}$$

$$\varphi_1 = |\cos^{-1} k_1^{-1}| \tag{17}$$

$$k_2 = \frac{[(2f/L) + 1]}{[1 - \frac{1}{4}(L/f)^2]^{\frac{1}{2}}} = \frac{[2 + (L/f)]}{\sin\theta}\,. \tag{18}$$

The general form of k_1 and k_2 versus L/f is shown in Fig. 3. Further discussion will be postponed to a later point in this paper.

We are also interested in the displacement r_m at the output of the mth diverging lens (Fig. 1). Using the relation

$$r_m = r_n + L r_n' \tag{19}$$

and using (7) for r_n', (15) for r_n with appropriate trigonometric relations, it can be shown that

$$r_m = r_0 k_3 \cos(m\theta - \varphi_3) + r_0' L\, k_4 \cos(m\theta - \varphi_4) \tag{20}$$

* See Appendix B.

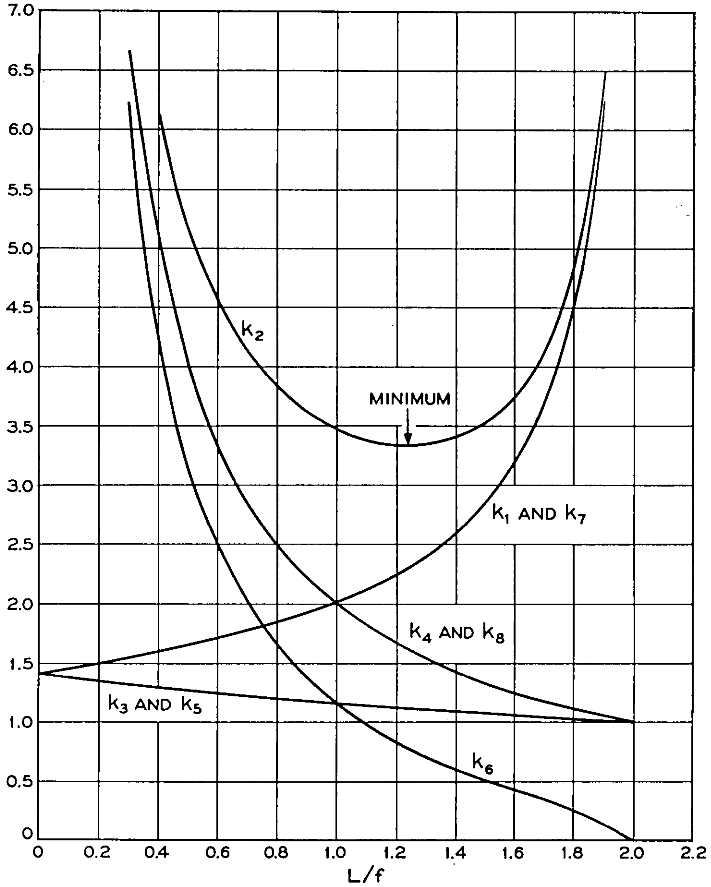


Fig. 3 — Coefficients relating input ray slope and displacement to ray displacement at the *n*th, *m*th, *N*th, and *M*th lenses of the alternating-gradient lens systems of Figs. 1 and 4.

where

$$k_3 = \left[\frac{2}{1 + (L/2f)}\right]^{\frac{1}{2}} \tag{21}$$

$$\varphi_3 = |\cos^{-1} k_3^{-1}| \tag{22}$$

$$k_4 = 2f/L \tag{23}$$

$$\varphi_4 = |\cos^{-1} k_4^{-1}| \tag{24}$$

and θ is again defined by (11). Plots of k_3 and k_4 are given in Fig. 3.

Equations (15) and (20) give the ray displacements at any lens in the system when the input is at a plane adjacent to a converging lens (i.e. at $n = 0$, Fig. 1). Before discussing interesting features of such ray propagation we will give the corresponding solutions for the case where the input is adjacent to a diverging lens (i.e. at $N = 0$, Fig. 4).

With reference to Figs. 4 and 5, and following a derivation similar to that carried out in connection with equations (1) to (9), it is found that

$$r_{N+2} - [2 - (L/f)^2]\, r_{N+1} + r_N = 0. \tag{25}$$

Note that (25) is identical to (9) and with the change of n into N the solution for (25) is again (10) and (11). When the initial conditions are put in, at $N = 0$, $r_N = r_0$ and $r_N' = r_0'$, we get

$$r_N = r_0 k_r \cos (N\theta + \varphi_5) + r_0' L k_6 \sin N\theta \tag{26}$$

where

$$k_5 = \frac{(L/f)[2 - (L/f)]^{\frac{1}{2}}}{\sin \theta} = \left[\frac{2}{1 + (L/2f)}\right]^{\frac{1}{2}} \tag{27}$$

$$\varphi_5 = |\cos^{-1} k_5^{-1}| \tag{28}$$

$$k_6 = \frac{[(2f/L) - 1]}{[1 - \frac{1}{4}(L/f)^2]^{\frac{1}{2}}} = \frac{[2 - (L/f)]}{\sin \theta}. \tag{29}$$

Note that k_5 is identical to k_3 in (21).

For r_M we find

$$r_M = r_0 k_7 \cos (M\theta + \varphi_7) + r_0' L k_8 \sin (M\theta - \varphi_8) \tag{30}$$

where

$$k_7 = \left[\frac{2}{1 - (L/2f)}\right]^{\frac{1}{2}} \tag{31}$$

$$\varphi_7 = |\cos^{-1} k_7^{-1}| \tag{32}$$

$$k_8 = 2f/L \tag{33}$$

$$\varphi_8 = |\cos^{-1} k_8^{-1}|. \tag{34}$$

Note that k_8 and φ_8 are identical to k_4 and φ_4, and that k_7 is identical to k_1. Plots of k_5, k_6, k_7, k_8 are given in Fig. 3.

III. RELATIONS FOR A SEQUENCE OF CONVERGING LENSES

For comparison purposes we will want to refer to the case of a sequence of identical convergent lenses equally spaced. The analysis is similar to that above for the alternating gradient lenses. The results are

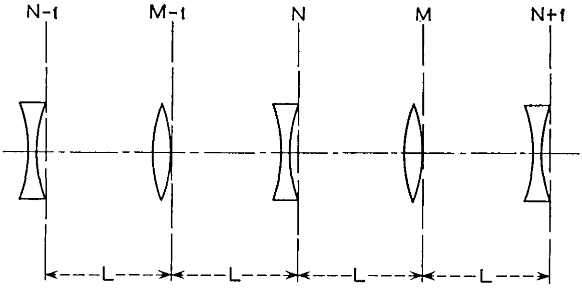


Fig. 4 — Sequence of lenses — second case.

as follows. Let the lens spacing be s, and the focal length be f. Then the displacement r_p at the pth lens is

$$r_p = r_0 k_9 \cos (p\delta - \varphi_9) + r_0' s \, k_{10} \sin p\delta \tag{35}$$

where

$$k_9 = \left[\frac{4f/s}{(4f/s) - 1}\right]^{\frac{1}{2}} \tag{36}$$

$$\varphi_9 = | \cos^{-1} k_9^{-1} | \tag{37}$$

$$k_{10} = \frac{(f/s)^{\frac{1}{2}}}{[1 - (s/4f)]^{\frac{1}{2}}} = \frac{1}{\sin \delta} \tag{38}$$

$$\delta = \cos^{-1} [1 - (s/2f)] \tag{39}$$

$$s/f = 2(1 - \cos \delta). \tag{40}$$

The system is stable in the sense that an input displacement r_0 or slope r_0' will remain bounded as p is increased if

$$0 < s/f < 4. \tag{41}$$

Fig. 6 shows the values of k_9 and k_{10} for comparison to Fig. 3.

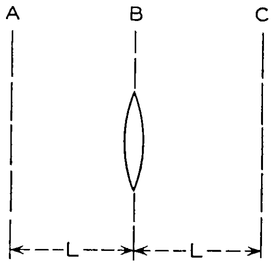


Fig. 5 — Lens subsection for Fig. 4.

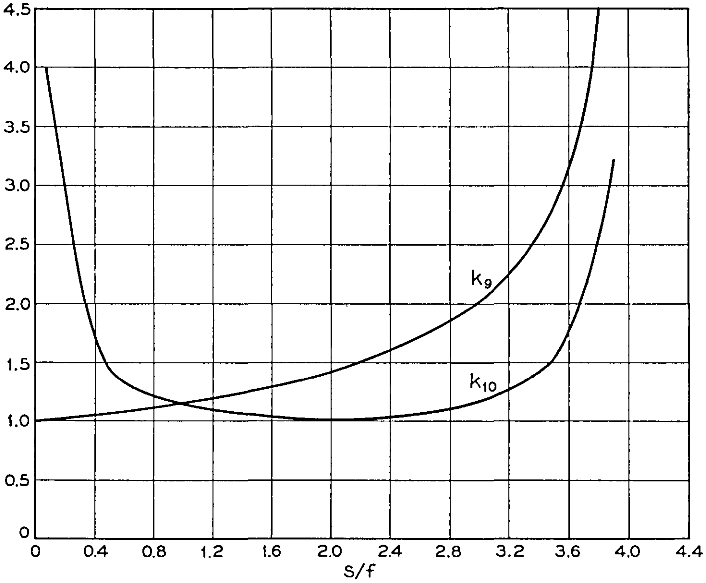


Fig. 6 — Coefficients relating ray slope and displacement to ray displacement at the pth lens of an all-convergent lens system; f = focal length, s = lens spacing.

IV. STABILITY COMPARISON

Since the converging lenses in Fig. 1 or Fig. 4 are spaced a distance $2L$, a comparison of (41) and (12) shows that the alternating-gradient system formed by adding a divergent lens halfway between the convergent lenses of an all-convergent lens system has the same stability condition as the original system. At a later point it will be shown that this is reasonable physically.

V. THE WEAK LENS CASE

When the lenses are weak, i.e., when $2f/L \gg 1$, the general expressions may be simplified to show some remarkable properties of alternating-gradient focusing. When the first lens is a diverging one, (15) and (20) yield

$$r_m = r_n = r_0 \sqrt{2} \cos [n\theta - (\pi/4)] + r_0'2f \sin n\theta \qquad (42)$$

and when the first lens is a converging one, (26) and (30) yield

$$r_N = r_M = r_0 \sqrt{2} \cos [N\theta + (\pi/4)] + r_0'2f \sin N\theta. \qquad (43)$$

These expressions show that for an input ray displacement without slope the maximum displacement of the transmitted ray (as n or N varies) is $\sqrt{2}$ times the input ray displacement, independent of both focal length and lens spacing! Also, for an input ray of zero displacement but finite slope r_0', the maximum displacement of the transmitted ray is $2fr_0'$, independent of lens spacing L. The angle θ is dependent on f and L and goes to zero as $f \to \infty$.

In an all-convergent lens system the similar condition $4f/s \gg 1$ leads to [from (35)]

$$r_p = r_0 \cos p\delta + r_0' \sqrt{fs} \sin p\delta. \tag{44}$$

In comparing the alternating-gradient system to the all-convergent lens system for weak lenses, we see that for an input ray with zero slope r_0' but finite displacement, r_0, the maximum output displacement for the alternating-gradient system is $\sqrt{2}$ times that of the all-convergent lens system. For input ray displacement $r_0 = 0$ but finite r_0', we see that the maximum output displacement for the alternating-gradient system is larger than for the all-converging lens system by the factor [see (42) and (44)]:

$$\frac{2fr_0'}{(fs)^{\frac{1}{2}}r_0'} = (4f/s)^{\frac{1}{2}}. \tag{45}$$

Our assumption of weak lenses made $4f/s \gg 1$, so (45) is a factor of two or more.

In this weak lens case both θ and δ are small angles, and from (11) and (39)

$$\theta \cong L/f \tag{46}$$

$$\delta \cong (s/f)^{\frac{1}{2}}. \tag{47}$$

Using the case of $s = 2L$, which is the alternating-gradient system formed by adding a diverging lens between the lenses of an all-convergent lens system

$$\theta/\delta = (s/4f)^{\frac{1}{2}}. \tag{48}$$

Since $4f/s \gg 1$ by our weak lens definition, θ/δ is less than unity and the period of the alternating-gradient system encompasses a great many more convergent lenses than does the all-convergent lens system with the same spacing of convergent lenses. This is as would be expected.

VI. OPTIMUM FOCAL LENGTHS

We now inquire as to whether there is a best value for the lens strength in order to minimize output ray displacement. On the assumption that the sine and cosine terms of (15), (20), (26), (30) and (35) go through unity for some number of lenses, the question is whether or not the coefficients k_1, $k_2 \cdots k_{10}$ have any minima.

For the all-convergent lens system Fig. 6 illustrates that k_9 has no useful minimum but that k_{10}, relating input ray slope to output ray displacement as in (35), does have a minimum. By setting

$$\frac{d}{df}(k_{10}) = 0 \tag{49}$$

we find

$$\left.\frac{s}{f}\right|_{\text{opt}} = 2 \tag{50}$$

at which condition $k_{10} = 1.0$. We note that the displacement r_p due to r_0' is $r_0'sk_{10} \sin p\delta$, so we have a minimum in this displacement when k_{10} is a minimum provided $\sin p\delta$ goes through unity for some number of lenses p. This is the most typical case, but there are notable exceptions. Suppose, for example, that δ of (39) is $\pi/3$, corresponding to $s/f = 1$; then $p\delta = \pi/3, 2\pi/3, \pi, 4\pi/3$, etc., as illustrated in Fig. 7, and $|\sin p\delta|$ never exceeds $\sin \delta$. Hence the maximum value of $r_0's\, k_{10} \sin p\delta$ is $r_0's$ for $s/f = 1$. It is shown in Appendix A that there is an infinite series of such discrete values, but the largest departure of the maximum value of $r_0'sk_{10} \sin p\delta$ from k_{10} is 15 per cent, occurring at s/f values of 1 and 3, as illustrated in Fig. 14 (see Appendix A).

Turning now to the alternating-gradient system, the only coefficient having a useful minimum is k_2 of (15), relating input ray slope to ray displacement at the converging lenses of Fig. 1. We find the minimum in k_2 by setting

$$\frac{d}{df}(k_2) = 0 \tag{51}$$

which leads to the equation

$$(L/f)^3 + 4(L/f)^2 - 8 = 0. \tag{52}$$

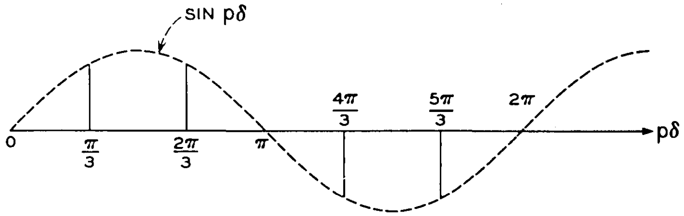


Fig. 7 — Diagram of $\sin p\delta$, $p = 1,2,3 \cdots$, when $\delta = \pi/3$.

The appropriate root of this equation is

$$\left.\frac{L}{f}\right|_{\text{opt}} = 1.237 \tag{53}$$

at which we calculate $\theta = 76.4°$, and $k_2 = 3.33$. Fig. 3 shows this minimum is sharper than the corresponding one for k_{10} of the all-converging lens system, Fig. 6. We note that $k_2 \sin n\theta$ of (15) contains the $\sin n\theta/\sin \theta$ factor, so once again (as described in Appendix A) for $\theta = \pi/3$ and other values, the maximum value of $k_2 \sin n\theta$ will be somewhat less than the value of k_2.

It is important to compare the optimized focusing effect of the alternating-gradient system to that for the all-convergent lens system. We make the comparison on the alternating-gradient system formed by adding a divergent lens of equal focal length in between the lenses of an all-convergent lens system; then we have $s = 2L$. The optimized maximum displacement due to input ray slope is

$$r_0's = 2r_0'L$$

for the convergent lens system, and is

$$3.33\, r_0'L$$

for the alternating-gradient system. It is remarkable that the focusing effect of the alternating-gradient system is so nearly the same as that of the all-convergent lens system. In practice it may be advantageous to get the focusing action in a manner that inherently reverses itself periodically. This analysis shows that such structures are nearly as effective as those wherein the focusing effect is always convergent.

Fig. 3 shows that the focal length which is optimum with respect to the input ray slope (k_2) is also an acceptable region with respect to input ray displacement (k_1 and k_3).

VII. RAY PATHS

One can get a useful physical feel for the wave propagation by tracing the rays in a few of the important cases.

For the all-convergent lens system optimized according to (50), Fig. 8 shows the ray paths for a zero-slope finite-displacement input ray and for a zero-displacement finite-slope input ray.* Here $\delta = 90°$ [see (35)] and a period is completed in 4 lenses. As proved in Appendix B, the

* Note that $k_9 = \sqrt{2}$, with $s = 2f$ in (36), but $k_9 \cos (p\delta - \varphi_9)$ is always ± 1 for any p. (The angle $\varphi_9 = 45°$.) This is an example of the caution that must be exercised in regarding the k's as maximum values of the various terms in r_n, r_m, r_p, etc.

response to an arbitrary input ray can be obtained by a linear superposition of the responses shown in Fig. 8.

For the alternating-gradient system the optimum according to (53) corresponds to an angle θ in (15), (20), (26) and (30) of 76.4°, which makes the ray path periodic only at a very large number of lenses. However, a very useful feel can be obtained from the ray paths for $\theta = 90°$, corresponding to $L = \sqrt{2}f$ and giving a value of k_2 only slightly larger than the minimum value (3.414 compared to 3.33). These ray plots are shown in Figs. 9 and 10 for the two types of input rays at the two possible points in the alternating-gradient system. We note that input ray displacement causes the same maximum displacement in the response regardless of where it occurs. Input ray slope is much more serious when it occurs in front of a diverging lens than when it occurs in front of a converging lens. Again, the response to arbitrary input rays can be obtained by adding the plotted responses.

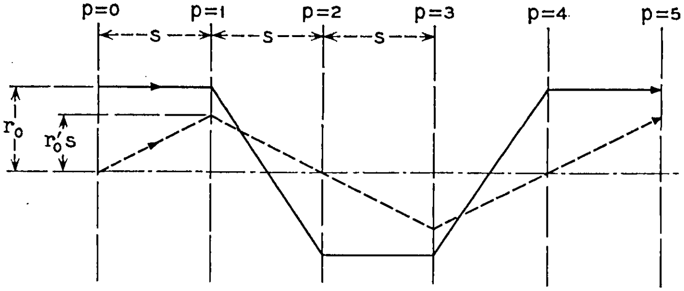


Fig. 8 — Ray paths in the confocal all-convergent lens system, $s = 2f$.

One can gain a little feel for the stability comparison made previously by looking at Figs. 11 and 12. Even though the r_0 term and the r_0' term of (35) for the all-convergent lens system go to infinity individually when $s = 4f$, a suitable combination of input ray slope and displacement remains bounded and this is illustrated in Fig. 11.* Any reduction in focal length f causes instability, and any increase in f leaves the system completely stable. It is clear that adding a lens of any kind at the midpoint between lenses in Fig. 11 will not alter the propagation of that ray. In Fig. 12 we see that adding a divergent lens in between the converging

* One can obtain these values of r_p from (35) by a suitable limiting process. It is helpful to start with the alternative form of (35):

$$r_p = r_0 \left\{ \cos p\delta + \frac{1}{[(4f/s) - 1]^{\frac{1}{2}}} \sin p\delta \right\} + r_0's\, k_{10} \sin p\delta.$$

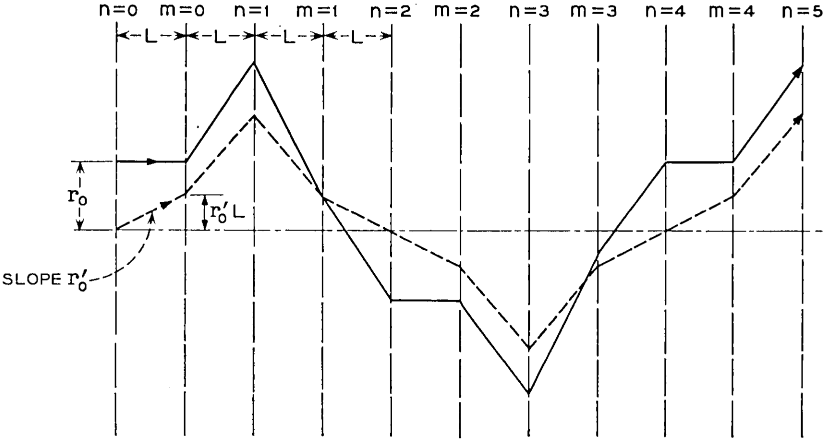


Fig. 9 — Ray paths for system of Fig. 1, $L = \sqrt{2f}$.

lenses will cause reductions in the focal length of the $p = 1$ converging lens to make the ray sent on to the $p = 2$ lens diverge even more; for increases in the focal length of the $p = 1$ lens, the divergent lens reduces the angle of the ray sent on to the $p = 2$ lens. Hence, it is plausible that the addition of the divergent lens between the convergent lenses does not alter the stability requirement on the focal lengths.

Given the mathematically-derived condition that divergent lenses of

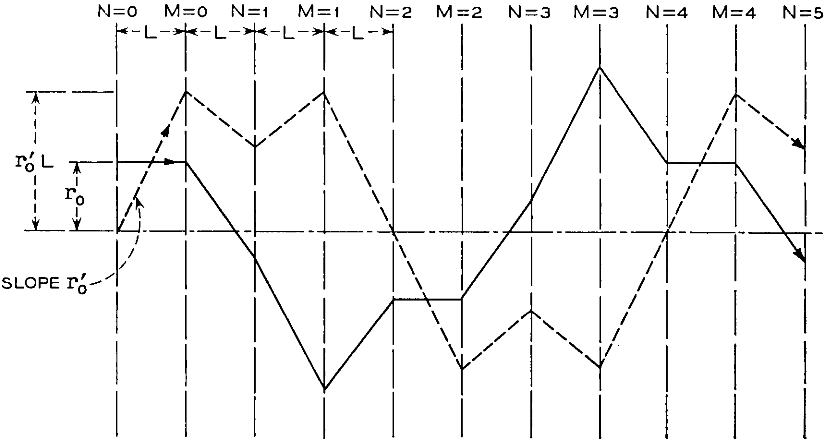


Fig. 10 — Ray paths for system of Fig. 3, $L = \sqrt{2f}$.

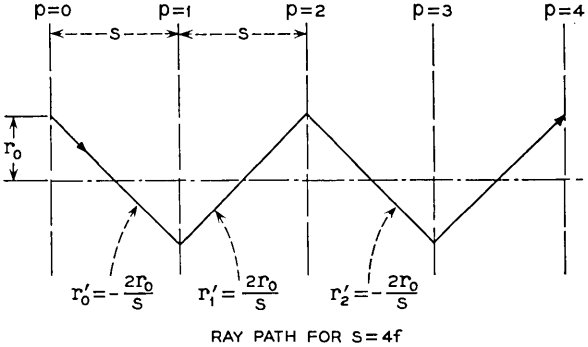


Fig. 11 — Ray path for $s = 4f$ in all-convergent lens system.

focal length f added to a chain of convergent lenses of focal length f do not change the stability criterion as described above, the physical argument just outlined leads to the conclusion that weaker divergent lenses would also leave the stability criterion unchanged.*

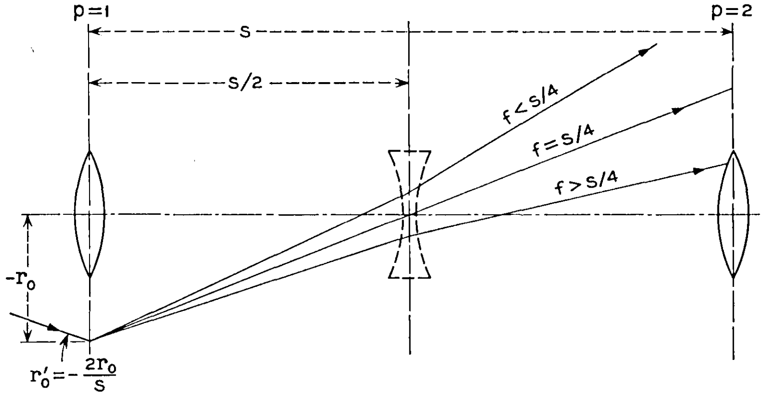


Fig. 12 — Ray path for f near $s/4$ to illustrate effect of divergent lens.

VIII. CONCLUSION

Alternating-gradient focusing is surprisingly close to an all-convergent lens system in focusing ability, and may be preferred if practical matters such as structural features or cost favor the alternating-gradient system.

* When reading this manuscript, Mr. J. P. Gordon commented that this conclusion is in agreement with the work of Boyd and Kogelnik[3] which can be shown to yield the relation $f_2 > f_1 - L/2$ for the required focal length f_2 of the diverging lens in terms of the focal length f_1 of the converging lens and the spacing L.

APPENDIX A

We examine here the maximum value that the term $r_0's\ k_{10} \sin p\delta$ of (35) can take as a function of lens number p when our objective is to minimize the term through appropriate choice of focal length. In the body of the article it has been shown that k_{10} has a minimum at $s = 2f$. This corresponds to a value of $\delta = \pi/2$ from (39) and it is evident that $\sin p\delta = \sin p\ \pi/2$ is either zero or unity for all integral values of p.

In the more general case we want to know the value of

$$k_{10} \sin p\delta = \sin p\delta/\sin \delta. \tag{54}$$

When it is recognized that p may take on all integral values greater than zero it follows that the maximum value of $[(\sin p\delta)/\sin \delta]$ as p varies can never be less than unity for any fixed δ.

It is possible for $[(\sin p\delta)/\sin \delta]$ to have a maximum value which is smaller than $k_{10} = 1/\sin \delta$. That is to say, $\sin p\delta$ does not necessarily go through unity even though p ranges from 0 to ∞ in integral steps.

Referring to Fig. 13, the maximum value of $\sin p\delta$ will be less than unity if

$$\delta(q + \tfrac{1}{2}) = \pi/2 \tag{55}$$

or

$$\delta = \pi/(2q + 1) \tag{56}$$

where $q = 1, 2, 3, \cdots$. It also is true that $\sin p\delta$ will have a maximum value less than unity for

$$\delta = r[\pi/(2q + 1)] \tag{57}$$

where $r = 1, 2, 3, 4, \cdots$.

The values of s/f corresponding to these values of δ and the resultant values of maximum $k_{10} \sin p\delta$ are given in Table I. Column 5 shows the ratio of k_{10} to the maximum of $k_{10} \sin p\delta$, and is a measure of the error

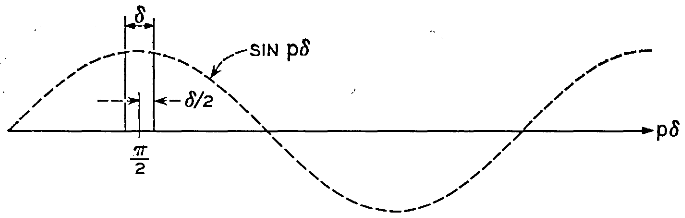


Fig. 13 — Diagram of $\sin p\delta$ yielding $|\sin p\delta| < 1$ for all p.

TABLE I

1 δ	2 s/f	3 k_{10}	4 Max. Value of k_{10} sin $p\delta$	5 Column 3 ÷ Column 4
$\pi/2$	2	1.0	1.0	1.0
$\pi/3$	1	1.15	1.0	1.15
$2\pi/3$	3	1.15	1.0	1.15
$\pi/5$	0.38	1.70	1.62	1.05
$2\pi/5$	1.38	1.05	1.0	1.05
$3\pi/5$	2.62	1.05	1.0	1.05
$4\pi/5$	3.62	1.70	1.62	1.05
$\pi/7$	0.194	2.31	2.255	1.023
$2\pi/7$	0.750	1.28	1.25	1.023
$3\pi/7$	1.554	1.023	1.0	1.023
$4\pi/7$	2.444	1.023	1.0	1.023
$5\pi/7$	3.226	1.28	1.25	1.023
$6\pi/7$	3.806	2.31	2.255	1.023

made in assuming sin $p\delta$ goes through unity. That ratio is $1/\cos(\delta/2)$ where δ is given by (56). All values for a given q in (57) result in the same error, but the various values of r indicate the values of δ and s/f at which that error will appear. Fig. 14 summarizes the data of Table I; for p ranging up to infinity it is only at the discrete values of δ given by (57) that the maximum of k_{10} sin $p\delta$ differs from k_{10}.

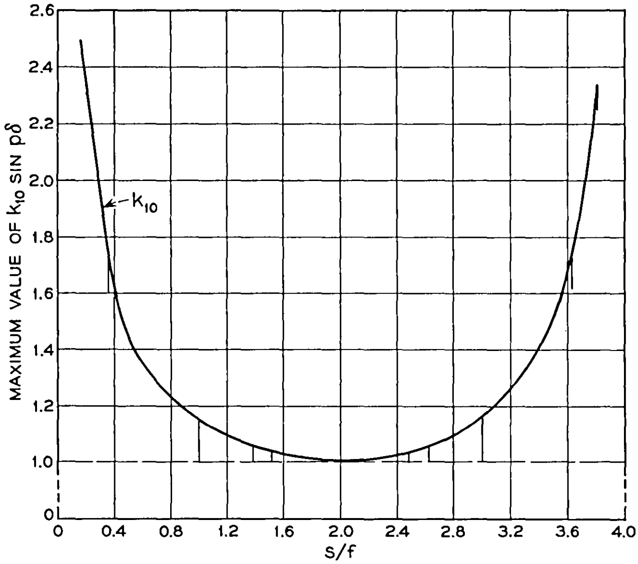


Fig. 14 — Maximum value of k_{10} sin $p\delta$ vs s/f.

If p were finite and the ratio s/f was varied, the plot of Fig. 14 would presumably show finite-width dips of the same over-all depth as those plotted.

APPENDIX B

It is the purpose of this appendix to point out that the "thin lens" description of light ray propagation leads to the following conclusion (see Fig. 15):

The slope and displacement of the output ray of an arbitrary sequence of lenses for an input ray of slope r_0' and displacement r_1 is exactly the algebraic sum of the slopes and displacements of the output ray found (*i*) for an input ray of slope r_0' with zero displacement from the axis, and assuming all lens displacements $d_n = 0$, (*ii*) for an input ray of displacement r_1 with zero slope and assuming all lens displacements $d_n = 0$, and (*iii*) for an input ray of zero slope and zero displacement and assuming one lens displacement at a time is nonzero, summing the ray output slopes and displacements thus found over all lens displacements.

The proof is as follows: For the nth lens in a sequence of lenses (see Fig. 1):

$$r_n' = r_{n-1}' - \left(\frac{r_n - d_n}{f_n}\right) \tag{58}$$

where r_n' is the slope of the ray immediately following the nth lens and r_n is the displacement at the nth lens. We note that the angular lens rotation φ_n does not affect the ray propagation, an approximation which implies that

$$(r_n - d_n) \cos \varphi_n \cong (r_n - d_n) \tag{59}$$

or

$$\varphi_n \ll 1.$$

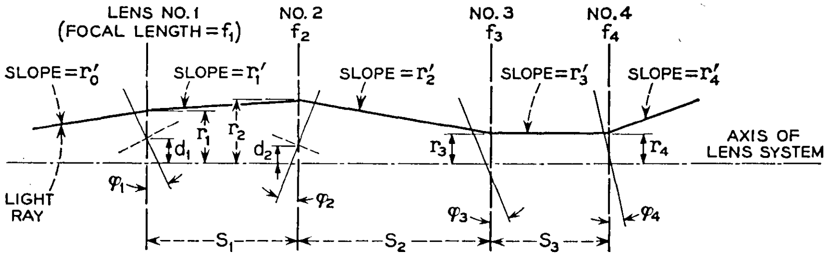


Fig. 15 — Light-ray path in an arbitrary lens system.

We may expand (58) to form

$$r_n' = r_0' - \frac{(r_1 - d_1)}{f_1} - \frac{(r_2 - d_2)}{f_2} \cdots - \frac{(r_n - d_n)}{f_n}. \qquad (60)$$

Similarly, the displacement r_n at the nth lens is

$$r_n = r_{n-1} + s_{n-1}r_{n-1}' \qquad (61)$$

which may be expanded to

$$r_n = r_{n-1} + s_{n-1}\left\{r_0' - \frac{(r_1 - d_1)}{f_1} - \frac{(r_2 - d_2)}{f_2} \cdots - \frac{(r_{n-1} - d_{n-1})}{f_{n-1}}\right\} \qquad (62)$$

which is valid for $n \geqq 2$.

We may examine each term of (60) and (62) and find that

$$r_n' = A_n r_0' + B_n r_1 + \sum_{m=1}^{m=n} \alpha_m \frac{d_m}{f_m} \qquad (63)$$

$$r_n = C_n r_0' + D_n r_1 + \sum_{m=1}^{m=n} \beta_m \frac{d_m}{f_m} \qquad (64)$$

in which A_n, B_n, C_n, D_n and the α_m and β_m are all independent of r_0', r_1 and the d_m.

We have thus proven the above-stated conclusion.

REFERENCES

1. Pierce, J. R., *Theory and Design of Electron Beams*, 2nd ed., D. Van Nostrand, Princeton, 1954.
2. Courant, E. D., Livingston, M. S., and Snyder, H. S., The Strong-Focusing Synchrotron — A New High-Energy Accelerator, Phys. Rev., **88**, Dec. 1, 1952, pp. 1190-1196.
3. Boyd, G. D., and Kogelnik, H., Generalized Confocal Resonator Theory, B.S.T.J., **41**, July, 1962, pp. 1347-1369.

Analysis of a Tubular Gas Lens

By D. MARCUSE and S. E. MILLER

(Manuscript received June 8, 1964)

If a cool gas is blown into a hot tube, it acts as a positive lens which will focus a light beam passing through the tube.

Using a theory presented in Ref. 3, we give curves which show the temperature distribution in the tube as a function of the distance from the tube axis and also as a function of the distance along the axis.

The focusing power of the lens is described by the difference in phase angle between a ray on the tube axis minus a ray at arbitrary distances from this axis and also as the second derivative of the phase angle on the axis of the tube. The phase curves, as a function of distance r from the tube axis, follow very closely an r^2 dependence. Expressions are given for the focal length of the lens.

The power consumption of the lens is discussed, and a figure of merit is defined as focusing power per watt. The gas used for this lens should be selected such that $(n - 1)/k$ is as large as possible (where n is the index of refraction, k the heat conductivity of the gas).

Using CO_2 and a $\frac{1}{4}$-inch ID tube 5 inches long heated 20°C above the incoming gas, a focal length of 5 feet with a power consumption of 0.325 watt is calculated; the focal length is inversely proportional to power consumption within certain limits.

I. INTRODUCTION

A communications system using light as the carrier of intelligence needs an efficient medium to propagate light from transmitter to receiver. Among the several alternatives, the idea of Goubau and Schwering[1] of confining and propagating an electromagnetic wave with a system of lenses appears promising. However, in a lens-waveguide system there is a wide range of possibilities as to what types of lenses to use. Conventional glass lenses present problems, since they may not only absorb light in the glass medium itself, but furthermore present important reflection losses which can be only partially avoided by special techniques such as coating the lens surfaces or making use of the Brewster angle.

Even if such corrective measures are used, there is still residual reflection and scattering of light due to unavoidable surface irregularities.

It appears that most of the problems connected with glass lenses could be overcome if, instead of a high-index medium such as glass, a very low-index focusing medium were used. If the transition from air into a dense medium could be avoided, the problem of light reflection would not exist. Gases present themselves as an obvious choice of a low-index dielectric medium. Their dielectric constant can be influenced by changing their density. A change of density is most easily effected by varying the gas temperature.

D. W. Berreman[2] built a successful gas lens by maintaining a temperature gradient between a hot helix and a cold cylindrical enclosure. Alternatively, D. W. Berreman and S. E. Miller proposed a gas lens formed by blowing a cool gas into a hot tube (Fig. 1). Since the gas heats up first at the wall of the tube and remains cool longer at its center, it has a density distribution of higher-density gas in the center of the tube and decreasing density towards the wall. Since an increase in density is accompanied by an increase in dielectric constant, it is easy to understand that the cool gas flowing through the hot tube acts as a positive lens and tends to focus a light beam traveling along the axis of the tube.

We present in this article some theoretical results of the temperature distribution in the gas and the difference in phase angle between two light beams, one traveling along the tube axis and the other traveling closer to the wall of the tube. This phase difference is a measure of the focusing power of the lens. For an economical lens we want maximum phase shift with a minimum of thermal power. We present curves showing the power consumption of the lens as well as the ratio of phase difference to power consumption. These data allow the construction of an optimum lens. Different gases give different lens properties. A gas is most efficient if the ratio $(n - 1)/k$ is large, where n is the index of refraction of the gas and k is its heat conductivity.

II. TEMPERATURE DISTRIBUTION

The theory of temperature distribution in a cool gas which is blown into a hot tube of constant temperature is presented in Ref. 3.

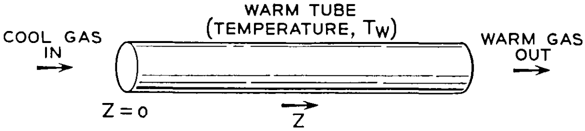


Fig. 1 — Gas lens using forced flow in a tube.

It is assumed that the gas flow is laminar and has the radial velocity distribution of a viscous fluid

$$v(r) = v_0[1 - (r/a)^2] \tag{1}$$

where

r = distance from tube axis
a = radius of tube
v_0 = gas velocity at $r = 0$.

The temperature T of the gas is given as

$$\theta = T_w - T$$

where T_w is the temperature of the wall of the tube. It is normalized with respect to

$$\theta_0 = T_w - T_0$$

with T_0 being the temperature of the cool gas before it enters the hot tube. θ/θ_0 is expanded in terms of functions $R_n(r/a)$, which are shown in Fig. 2 for $n = 0$, 1 and 2. Values of $R_n(r/a)$ are listed in Table I. The temperature depends on the distance z measured from the beginning of the hot tube, the gas velocity v_0, and the following material parameters

k = heat conductivity measured in cal/cm sec deg)
(deg = degrees Kelvin)
ρ = gas density in gram/cm^3
c_p = specific heat at constant pressure in cal/gram.

All these parameters depend somewhat on the temperature but are considered constant in the derivation of the theory. They enter the equations in the combination

$$\sigma = k/av_0\rho c_p\,. \tag{2}$$

TABLE I — R FUNCTIONS OF FIG. 2

x	$R_0(x)$	$F(x)$	$R_1(x)$	$R_2(x)$
0	1	1	1	1
0.1	0.9819	0.9805	0.8923	0.753
0.2	0.9290	0.9261	0.6067	0.206
0.3	0.8456	0.8432	0.2367	−0.290
0.4	0.7382	0.7382	−0.1062	−0.407
0.5	0.6147	0.6175	−0.3399	−0.204
0.6	0.4833	0.4880	−0.4317	0.104
0.7	0.3506	0.3535	−0.3985	0.278
0.8	0.2244	0.2244	−0.3051	0.278
0.9	0.1069	0.1041	−0.1637	0.144
1.0	0	0	0	0

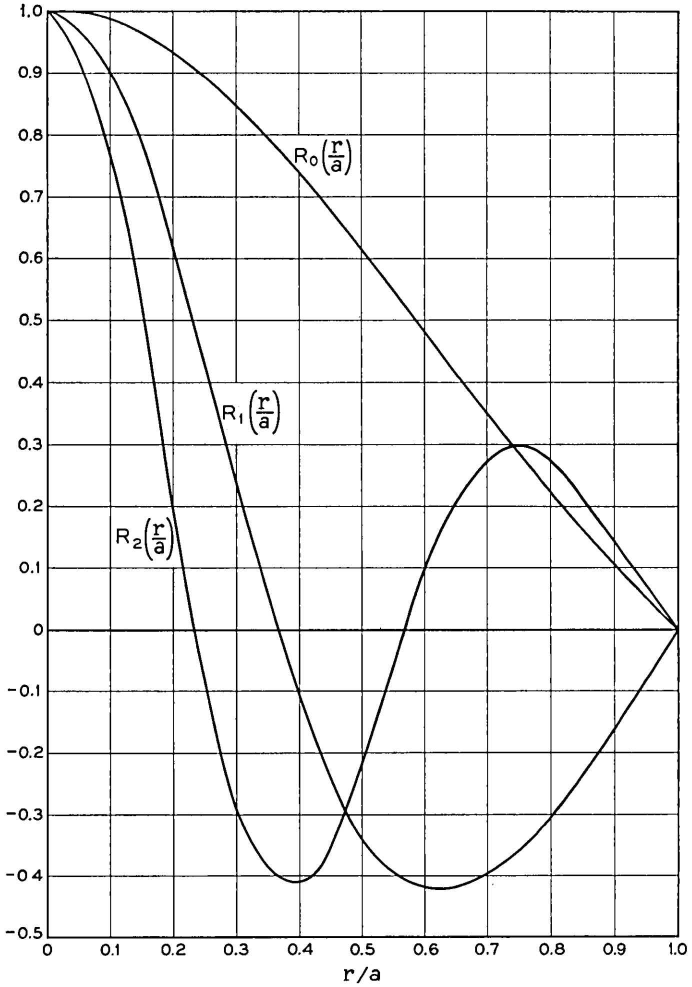


Fig. 2 — Functions R_0, R_1 and R_2 vs r/a.

The first three terms of the infinite series describing the temperature distribution in the tube are

$$\frac{\theta}{\theta_0} = 1.477 \exp\left(-\ 7.316\sigma \frac{z}{a}\right) R_0\left(\frac{r}{a}\right) - 0.810 \exp\left(-44.36\sigma \frac{z}{a}\right) \cdot R_1\left(\frac{r}{a}\right) + 0.385 \exp\left(-106\sigma \frac{z}{a}\right) R_2\left(\frac{r}{a}\right) \pm \cdots. \tag{3}$$

The approximation is fairly poor at $z = 0$. However, the exponential factors in the higher terms of the series drop off very rapidly as z increases so that the approximation is already very good for values of

$$\sigma(z/a) > 0.01.$$

Fig. 3 shows θ/θ_0 at $r = 0$ as a function of $\sigma(z/a)$. It is apparent that for $\sigma(z/a) > 0.05$, the distribution of θ/θ_0 drops off exponentially. Fig. 4 shows the r/a dependence of θ/θ_0 for different values of $\sigma(z/a)$.

To make Figs. 3 and 4 more meaningful, we list in Table II the material parameters for several gases at 20°C and a pressure of 760 mm Hg.

III. POWER CONSUMPTION

The principle of operation of our gas flow lens requires that we heat the cool gas inside the hot tube. Even if we neglect all power losses to the environment, we have to spend a certain amount of heat power to operate our lens. For a subsequent study of lens efficiency we need to know this basic power consumption. At any given length z of the tube we obtain the power absorbed by the gas as

$$\begin{aligned} P(z) &= \int_0^a [T(r,z) - T_0]\rho c_p v(r) 2\pi r \, dr \\ &= 2\pi\rho c_p v_0 \theta_0 \int_0^a r \left[1 - \left(\frac{r}{a}\right)^2\right]\left[1 - \frac{\theta(r,z)}{\theta_0}\right] dr. \end{aligned} \tag{4}$$

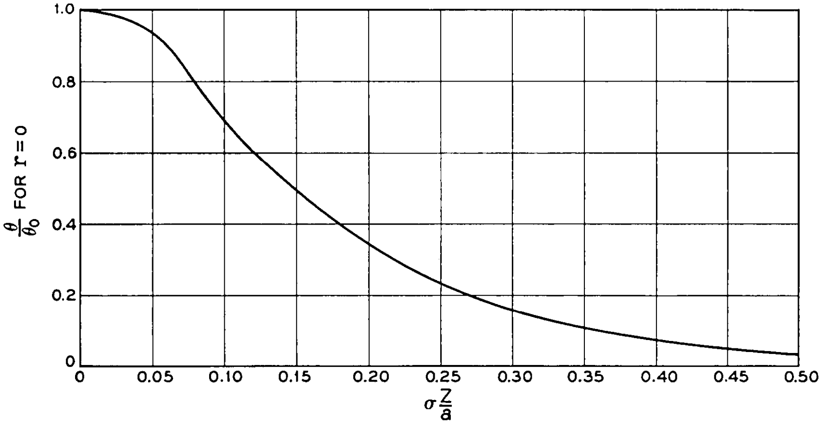


Fig. 3 — Normalized gas temperature on tube axis vs normalized distance along tube.

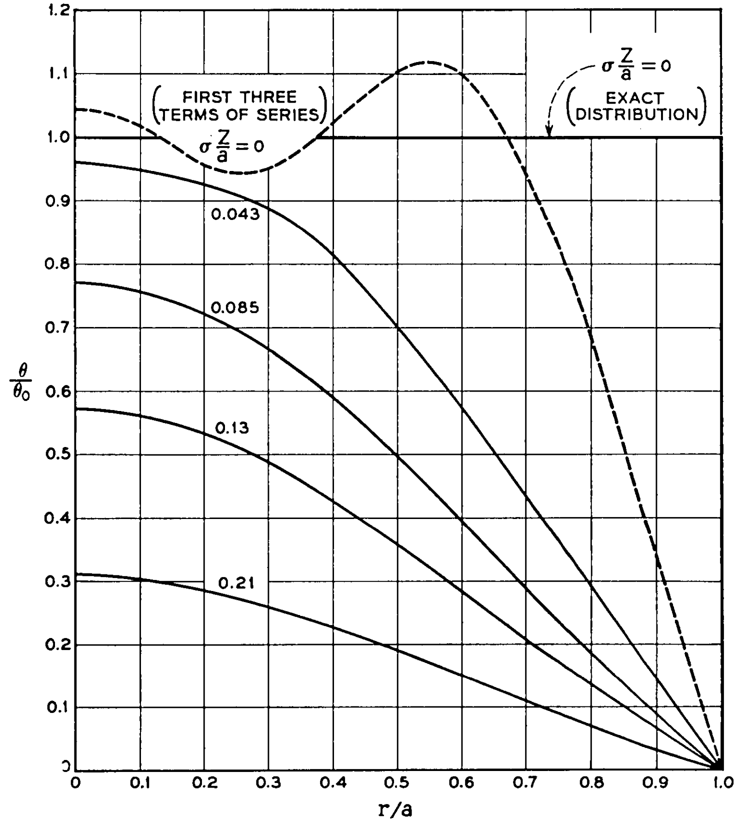


Fig. 4 — Normalized gas temperature vs radial position with longitudinal position as a parameter.

In order to be able to perform the integration easily, we restrict ourselves to values of

$$\sigma(z/a) > 0.05$$

which allows us to express θ/θ_0 by the first term of (3). In addition, we replace $R_0(x)$ by

$$R_0(x) \cong F(x) = 1 - 2.06\, x^2 + 1.06\, x^3. \tag{5}$$

This approximation deviates no more than 2.5 percent from the actual value of $R_0(x)$. The values of $F(x)$ can be compared to those of $R_0(x)$ in Table I. The integration can now be performed easily, and we obtain:

$$P = \frac{\pi}{2} a^2 \rho c_p v_0 \theta_0 \left[1 - 0.820 \exp\left(-7.316\sigma \frac{z}{a} \right) \right]. \tag{6}$$

TABLE II — GAS PARAMETERS VS TEMPERATURE

Gas	k (cal/cm sec deg)	ρ (gram/cm³)	c_p (cal/gram deg)	$av_0\sigma = k/\rho c_p$ (cm²/sec)	$n - 1$	$(n - 1)/k$ (cm sec deg/cal)
CO_2	$3.93\ 10^{-5}$	$1.84\ 10^{-3}$	0.199	0.107	$4.20\ 10^{-4}$	10.7
NH_3	$5.90\ 10^{-5}$	$0.72\ 10^{-3}$	0.523	0.157	$3.48\ 10^{-4}$	5.9
CH_4	$7.80\ 10^{-5}$	$0.67\ 10^{-3}$	0.528	0.220	$4.13\ 10^{-4}$	5.3
Air	$6.28\ 10^{-5}$	$1.21\ 10^{-3}$	0.240	0.216	$2.73\ 10^{-4}$	4.35
H_2	$41.0\ 10^{-5}$	$0.084\ 10^{-3}$	3.39	1.44	$1.23\ 10^{-4}$	0.30
He	$35.0\ 10^{-5}$	$0.166\ 10^{-3}$	1.25	1.69	$0.34\ 10^{-4}$	0.097

Using ρ and c_p from Table II, P is in calories/sec. Fig. 5(a) shows the power consumption as a function of normalized lens length, $\sigma z/a$. An alternative form of (6) brings out the dependence of P on the flow velocity v_0 more clearly:

$$P = \frac{\pi}{2} kz\theta_0 \frac{v_0}{V} \left[1 - 0.820 \exp \left(-7.316 \frac{V}{v_0} \right) \right]. \tag{7}$$

The quantity

$$V = kz/a^2 \rho c_p = \sigma(z/a)v_0 \tag{8}$$

has the dimension of velocity and is characteristic of the gas and the tube geometry. Fig. 5(b) shows the power consumption as a function of normalized gas velocity, v_0/V.

The ratio of V/v_0 can be related to the time $t_0 = z/v_0$ which it takes the gas particles on the axis to traverse the tube of length z with the velocity v_0 and to a time τ which is defined by

$$\frac{1}{\tau} = \frac{\dfrac{dT(0,t)}{dt}}{T_w - T(0,t)}. \tag{9}$$

τ is characteristic of the heat diffusion rate in a gas which rests in a tube whose wall temperature is T_w. At $t = 0$ the gas has the uniform temperature $T_0 < T_w$. Its temperature at a given radius r and time t is $T(r,t)$, so that $T(0,t)$ is the gas temperature at the tube axis at time t. $1/\tau$ is the time rate of temperature rise on the axis per degree of temperature difference between wall and axis.

It is shown in Appendix A that

$$V/v_0 = \sigma z/a = 0.173(t_0/\tau). \tag{10}$$

This shows that V/v_0 expresses the ratio of the time it takes the gas particles (on the tube axis) to flow through the tube of length z to the heat diffusion rate on the tube axis. Equation (10) may be used to replace

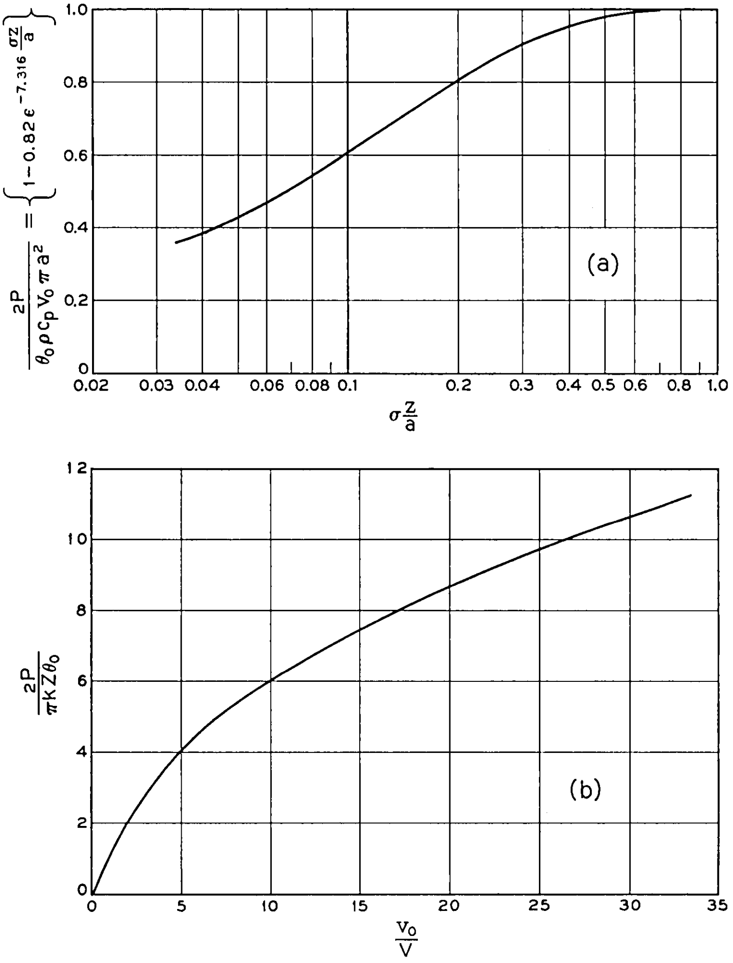


Fig. 5 — (a) Normalized power flow to the gas vs longitudinal position. (b) Normalized power flow to the gas vs normalized gas velocity.

V/v_0 in (7) and in the other places where V/v_0 or $\sigma z/a$ appears, resulting (for example) in

$$\exp\left(-7.316\,\frac{\sigma z}{a}\right) = \exp\left(-7.316\,\frac{V}{v_0}\right) = \exp\left(-1.265\,\frac{t_0}{\tau}\right).$$

The parameters lens length z and gas velocity v_0 are important variables for other reasons, and the first two forms of the exponential may be preferred. We may note, however,

$$\tau = 0.173(a^2 \rho c_p / k)$$

and typical values are 0.161 second and 0.08 second for CO_2 and air, respectively, when $a = 0.125$ inch. It is surprising that this time constant is so short.

Another physical meaning one can give V is that it represents that velocity of gas flow along the pipe axis which assures that θ/θ_0 drops from its initial value of one at the beginning of the tube to

$$\theta/\theta_0 = 9.910^{-4} \approx 10^{-3}$$

on the axis at its end.

IV. FOCUSING ACTION

A lens focuses because the optical path length varies for rays traveling at different distances from its axis.

We describe the focusing action of our lens by the phase angle of a ray traveling parallel to the axis of the structure. The phase angle is given by

$$\Phi(r,z) = \beta_0 \int_0^z n(r,x)\, dx. \tag{11}$$

Here, $\beta_0 = 2\pi/\lambda_0$ is the free-space propagation constant of the light beam, and n is the index of refraction of the gas.

$$n(r,x) = 1 + (n_0 - 1) \frac{T_0}{T(r,x)} \tag{12}$$

$$\frac{T_0}{T} = \frac{1}{1 + \frac{\theta_0}{T_0}\left(1 - \frac{\theta}{\theta_0}\right)} \approx 1 - \frac{\theta_0}{T_0}\left(1 - \frac{\theta}{\theta_0}\right). \tag{13}$$

The last step is an approximation for $\theta_0/T_0 \ll 1$. The temperature in (12) has to be expressed in degrees Kelvin.

We decompose $\Phi(r,z)$ into two parts:

$$\Phi(r,z) = \varphi + \psi(r,z). \tag{14}$$

The first part

$$\varphi = \beta_0 \left[1 + (n_0 - 1)\left(1 - \frac{\theta_0}{T_0}\right)\right] z \tag{15}$$

is independent of the position r of the ray in the gas lens, while the second part

$$\begin{aligned}\psi &= \beta_0(n_0 - 1)\frac{\theta_0}{T_0}\int_0^z \frac{\theta(r,x)}{\theta_0}\,dx \\ &= \beta_0 z(n_0 - 1)\frac{\theta_0}{T_0}\cdot\frac{v_0}{V}\left\{0.202R_0\left(\frac{r}{a}\right)\left[1 - \exp\left(-7.316\frac{V}{v_0}\right)\right]\right. \\ &\quad - 0.0183R_1\left(\frac{r}{a}\right)\left[1 - \exp\left(-44.3\frac{V}{v_0}\right)\right] \\ &\quad \left. + 0.00363R_2\left(\frac{r}{a}\right)\left[1 - \exp\left(-106\frac{V}{v_0}\right)\right] + \cdots\right\}\end{aligned} \tag{16}$$

accounts for the different amounts of phase shift in different parts of the lens.

The difference between the phase angle of a ray traveling along the lens axis and the phase angle of a ray traveling at a distance r from the axis is

$$\begin{aligned}\Delta\Phi &= \beta_0 z(n_0 - 1)\frac{\theta_0}{T_0}\frac{v_0}{V}\left\{0.202(1 - R_0)\left[1 - \exp\left(-7.316\frac{V}{v_0}\right)\right]\right. \\ &\quad - 0.0183(1 - R_1)\left[1 - \exp\left(-44.3\frac{V}{v_0}\right)\right] \\ &\quad \left. + 0.00363(1 - R_2)\left[1 - \exp\left(-106\frac{V}{v_0}\right)\right] + \cdots\right\}.\end{aligned} \tag{17}$$

This form of $\Delta\Phi$ shows clearly its dependence on flow velocity for a fixed tube length z. To study the dependence of $\Delta\Phi$ for fixed flow rate and varying length, the following form is preferable.

$$\begin{aligned}\Delta\Phi &= \beta_0(a/\sigma)\,(n_0 - 1)\,(\theta_0/T_0)\,\{0.202(1 - R_0) \\ &\cdot(1 - \exp(-7.316\sigma z/a)) - 0.0183(1 - R_1)\,(1 - \exp(-44.3\sigma z/a)) \\ &\quad + 0.00363(1 - R_2)\,(1 - \exp(-106\sigma z/a)) + \cdots\}.\end{aligned} \tag{18}$$

Fig. 6 shows a plot of $\Delta\Phi$ versus length of lens for $r/a = 0.4$. That means that we compare the phase difference between a ray on the axis and another at distance $r = 0.4a$ from the axis of the lens.

Fig. 7 shows the phase difference at a fixed length z as a function of gas velocity. In this case, $\Delta\Phi$ goes through a maximum which for $r/a = 0.4$ occurs at $v_0/V = 6.9$. The position of this maximum depends somewhat on the radius r of the ray used for phase comparison with the axial ray. Appendix B gives the theory and Table III gives the values of

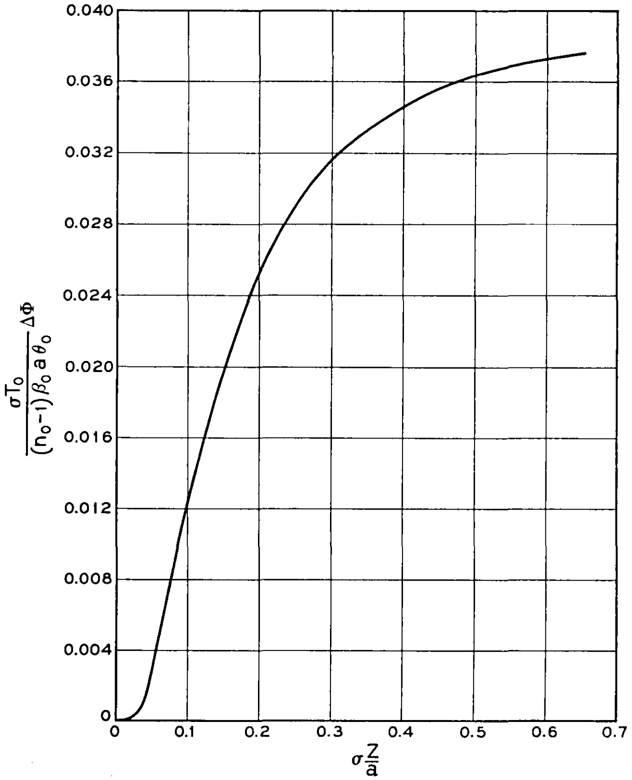


Fig. 6 — Phase difference between ray at $r = 0$ and ray at $r = 0.4a$ vs longitudinal position.

the position of the maximum v_0/V for different values of r/a. Table III shows that the position of the maximum does not change much with the radial position of the reference ray.

We can explain physically this maximum in $\Delta\Phi$ versus v_0 as follows. At very low gas velocity the majority of the gas in the tube is at the same temperature — the temperature of the walls, T_w. It is heated up in a time τ [see (9)] after entering. With little temperature difference between the gas at $r = 0$ and at $r > 0$ there is little $\Delta\Phi$. As v_0 increases, the gas on the axis remains at or near T_0, but that nearer the walls is heated because it flows more slowly [see (1)] and larger $\Delta\Phi$ develops. Beyond some velocity, further increases in velocity cause the gas at $r = 0.4$a (for example) to leave the tube at lower and lower temperatures — i.e., less temperature difference will exist between $r = 0$ and $r = 0.4a$ because the

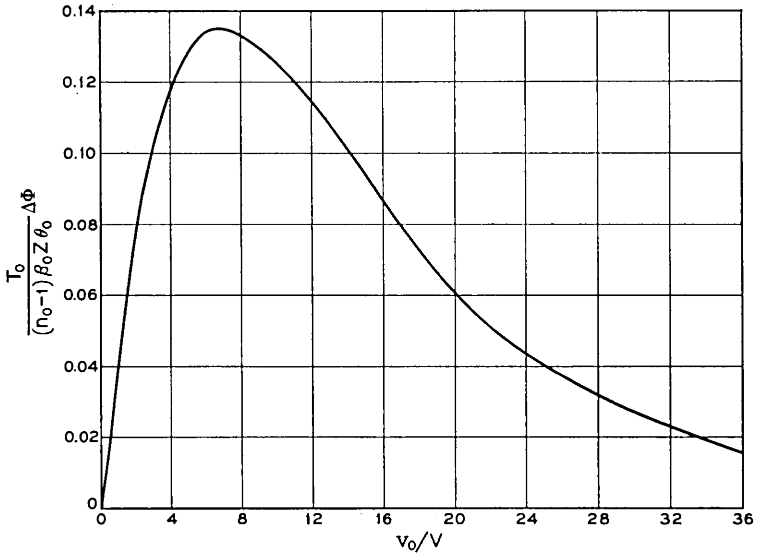


Fig. 7 — Phase difference between ray at $r = 0$ and ray at $r = 0.4a$ vs gas velocity.

transit time of the gas through the tube becomes less than the thermal diffusion time constant τ. Thus, at high velocities the $\Delta\Phi$ decreases. Note that this explanation (and the theory) depends upon laminar flow of the gas. If there is radial mixing of the gas, less $\Delta\Phi$ would be expected than predicted above, and the maximum in $\Delta\Phi$ versus v_0 might not occur.

The second derivative $d^2\Phi/d(r/a)^2$ is a good measure of the effectiveness of the lens for light rays close to its axis. For a glass lens

$$\Phi = n\beta_0 \, d(r)$$

where $d(r)$ is the thickness of the lens as a function of the distance from its axis. The radius of curvature R of the glass lens is given by

$$\frac{1}{R} = \frac{1}{n\beta_0} \frac{d^2\Phi}{dr^2} .$$

In order to be able to take the second derivative of Φ we have to express the functions R_0, R_1, and R_2 by power series with respect to r/a. For

Table III — Maximum v_0/V for Values of r/a

r/a	0.2	0.4	0.6
v_0/V	6.73	6.9	8.26

the second derivative on the axis at $r = 0$, it is sufficient to know the coefficient of $(r/a)^2$ in the expansion. Jakob[3] gives a series expansion of the R-functions.

$$R_n = 1 - \tfrac{1}{4}\beta_n^2(r/a)^2 \pm \cdots \tag{19}$$

with $\beta_0 = 2.705$, $\beta_1 = 6.66$, and $\beta_2 = 10.3$.

We get these values

$$\begin{aligned}\left(\frac{d^2\Phi}{d\left(\frac{r}{a}\right)^2}\right)_{r=0} = \beta_0 z(n_0 - 1)\,\frac{\theta_0}{T_0}\,\frac{v_0}{V}\Big\{0.738\left[1 - \exp\left(-7.316\,\frac{V}{v_0}\right)\right] \\ - 0.405\left[1 - \exp\left(-44.3\,\frac{V}{v_0}\right)\right] \\ + 0.192\left[1 - \exp\left(-106\,\frac{V}{v_0}\right)\right] + \cdots\Big\}.\end{aligned} \tag{20}$$

The maximum of this curve as a function of v_0/V appears at $v_0/V = 6.75$.

Fig. 8(a) is a plot of a normalized value of $d^2\Phi/d(r/a)^2$ as a function of v_0/V, while Fig. 8(b) shows it (with a different normalization) as a function of $\sigma(z/a)$.

The r/a dependence of $\Delta\Phi$ is shown in Fig. 9.

In order to show what values the phase difference might actually assume, and also to compare different gases, we have plotted $\Delta\Phi$ in Fig. 10 for several gases and the following geometry and flow rate:

$2a = 0.25$ inch
$v_0 = 212$ cm/sec, corresponding to 2 liters/minute or 4.77 miles/hr.
$\beta_0 = 1.07 \times 10^5\ \text{cm}^{-1}$, corresponding to $\lambda_0 = 5890\text{Å}$
$T_w = 343°\text{K}$
$T_0 = 293°\text{K}$

and with the values of $n_0 - 1$ and σ as listed in Table II. These curves assume a tube length z so long that no further $\Delta\Phi$ would be realized with a longer z (i.e., outgoing gas at uniform temperature).

It is interesting to compare the r/a dependence of $\Delta\Phi$ to the simple function $c(r/a)^2$. For this purpose we use (17), which is written so that $\Delta\Phi = 0$ at $r = 0$. Fig. 11 gives the normalized value of $\Delta\Phi$ as a function of r/a for several values of the gas velocity v_0/V. For comparison the function $c(r/a)^2$ is shown by dotted lines. The constant c is adjusted so that both curves coincide at $r/a = 0.4$. The actual curves of $\Delta\Phi$ are surprisingly close to the simple square law dependence in all cases. If the

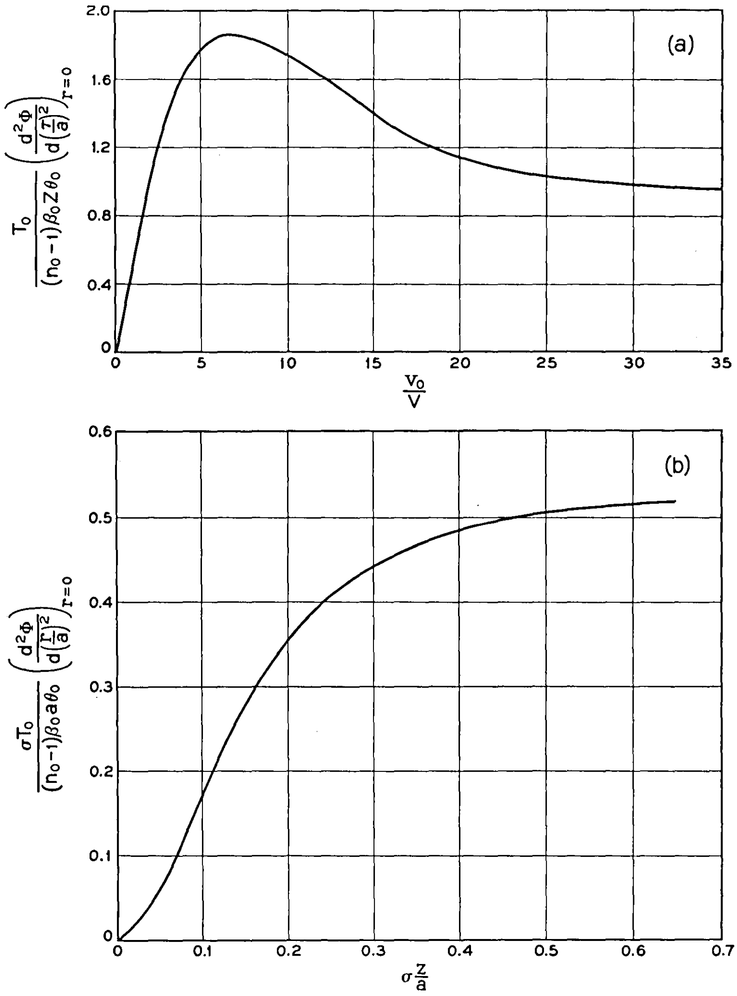


Fig. 8 — (a) Normalized $d^2\Phi/d(r/a)^2$ vs gas velocity; (b) normalized $d^2\Phi/d(r/a)^2$ vs longitudinal position.

gas lens could be treated as a thin lens, it would act very similar to a glass lens with spherically curved surfaces.

However, the gas lens is not thin and the question presents itself: how do different sections of the lens contribute to the over-all focusing effect? Fig. 12 shows the phase difference $\Delta\Phi$ between a ray on the axis at $r = 0$ and a ray at r for a fixed value $v_0/V = 6.9$ for different sections

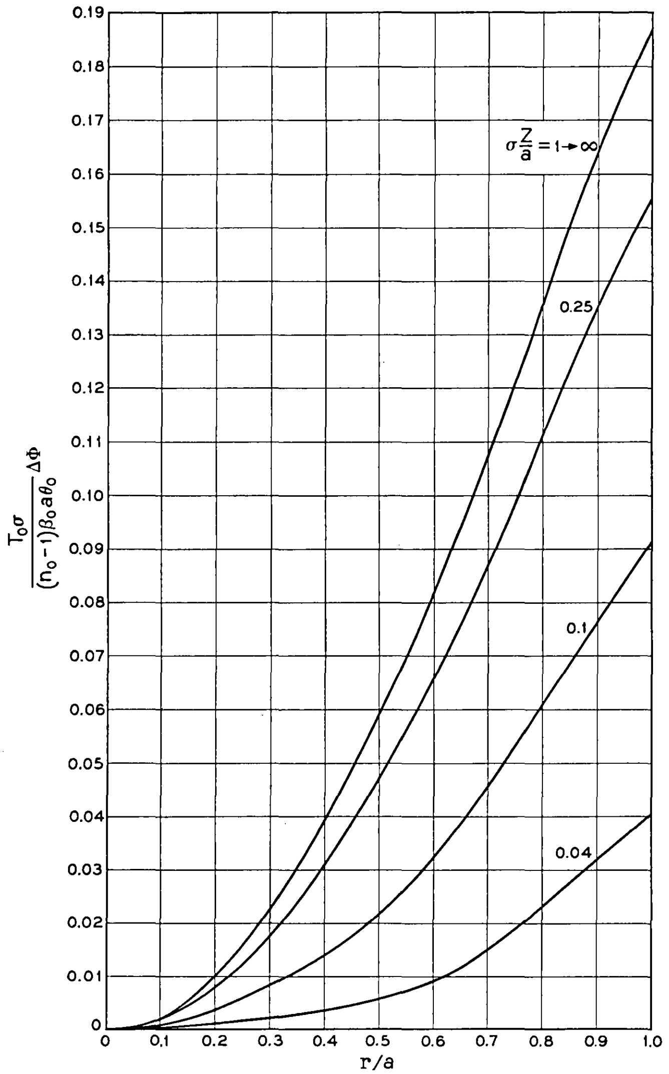


Fig. 9 — Normalized $\Delta\Phi$ vs r/a with longitudinal position as a parameter.

of the lens. The curve showing $\Delta\Phi$ for the section $0 \rightarrow z$ has already been shown in Fig. 11. The other curves show $\Delta\Phi$ in the first $\frac{1}{3}$ of the lens [curve $0 \rightarrow (\frac{1}{3})z$], the second $\frac{1}{3}$[curve $(\frac{1}{3})z \rightarrow (\frac{2}{3})z$], and the last $\frac{1}{3}$[curve $(\frac{2}{3})z \rightarrow z$]. The contributions are surprisingly different at different radii and do not resemble simple $(r/a)^2$ dependences. However, they all add

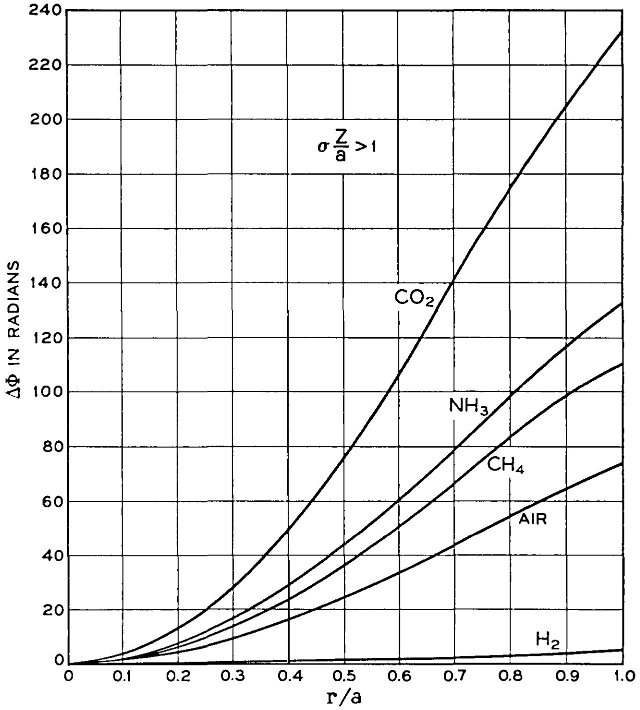


Fig. 10 — $\Delta\Phi$ vs r/a when $a = 0.125$ inch, $v_0 = 212$ cm/sec (corresponding to 2 liters/min), $\lambda_0 = 5890$ Å, $T_W = 343°\text{K}$, $T_0 = 293°\text{K}$ and $\sigma Z/a > 1$.

up to the $0 \rightarrow z$ curve which does resemble the $(r/a)^2$ dependence very closely, as was shown in Fig. 11.

V. FIGURE OF MERIT

The focusing action of the gas lens becomes independent of the length of the lens if $\sigma(z/a) > 1$, as Figs. 6 and 8(b) show. We also know that, for a fixed length of the lens, there is an optimum flow velocity, as shown by Figs. 7 and 8.

For practical applications one would like not only to obtain an effective lens but also to do so with a minimum expenditure of power. It is, therefore, interesting to study the lens action, that is, $d^2\Phi/d(r/a)^2$, per unit of applied power. We may introduce the ratio

$$M = \frac{1}{P}\left(\frac{d^2\Phi}{d(r/a)^2}\right)_{r=0} \tag{21}$$

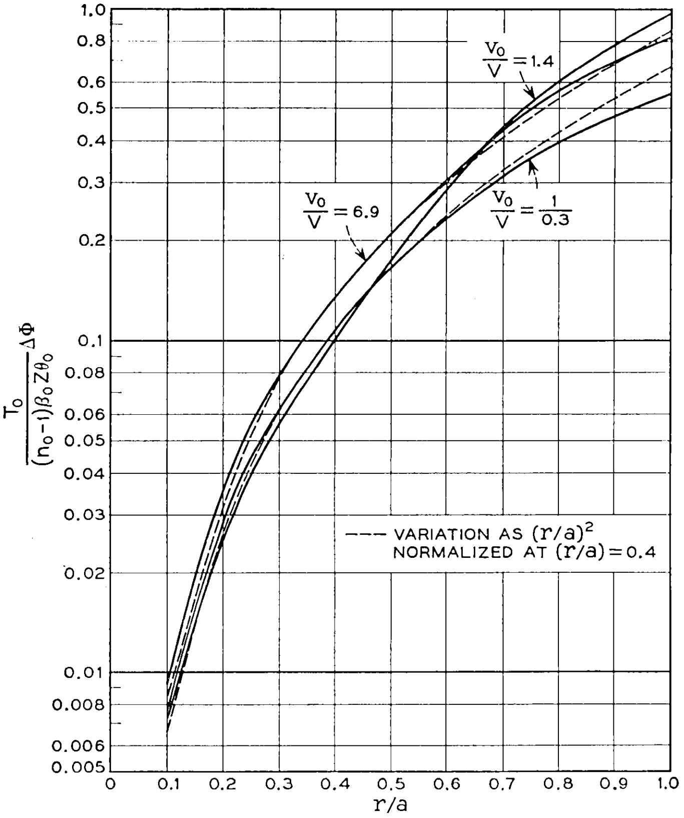


Fig. 11 — Normalized ΔΦ vs r/a with gas velocity as a parameter; dotted curve represents variation as $(r/a)^2$, normalized to ordinate at $r/a = 0.4$.

as the figure of merit of the lens. From (6) and (20) we obtain

$$M = \frac{2\beta_0}{\pi T_0}\cdot\frac{n_0 - 1}{k}\left\{\frac{F}{1 - 0.820\exp\left(-7.316\,\dfrac{V}{v_0}\right)}\right\} \tag{22a}$$

with

$$\begin{aligned} F = {} & 0.738\left[1 - \exp\left(-7.316\,\frac{V}{v_0}\right)\right] - 0.405 \\ & \cdot\left[1 - \exp\left(-\,44.3\,\frac{V}{v_0}\right)\right] + 0.192\left[1 - \exp\left(-106\,\frac{V}{v_0}\right)\right]. \end{aligned} \tag{22b}$$

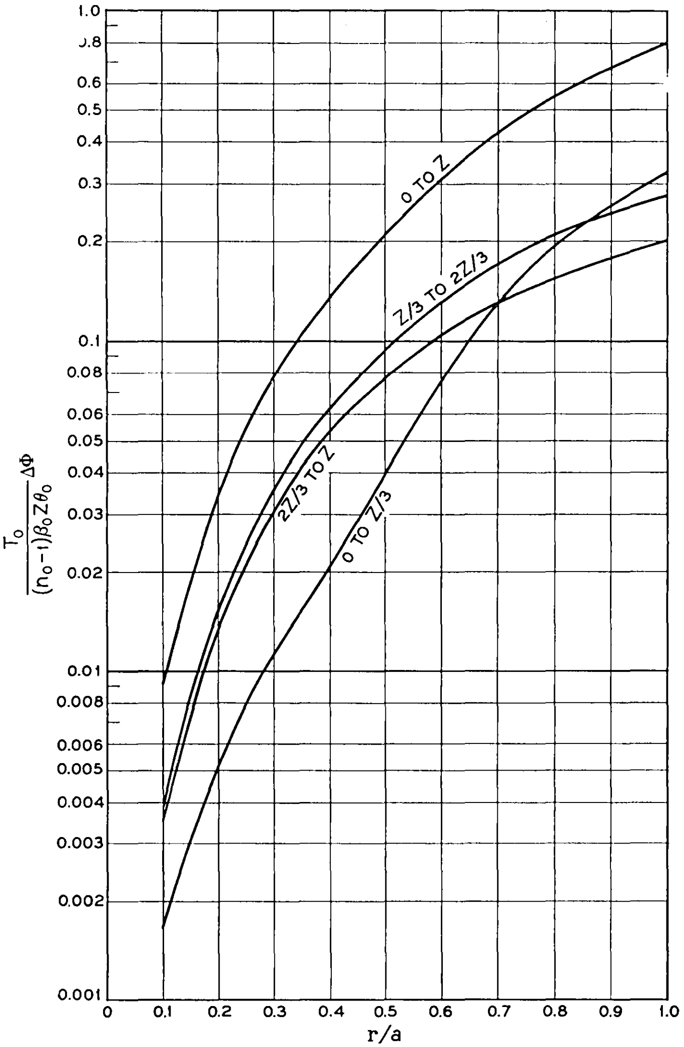


Fig. 12 — $\Delta\Phi$ contributed by first third, second third, and output third of gas lens for $v_0/V = 6.9$.

For a given value of v_0/V, the figure of merit is proportional to $(n_0 - 1)/k$. It is advantageous to make this number as large as possible.

The figure of merit M, given in (22b), is plotted in Fig. 13.

The gases in Table II are arranged in decreasing order of $(n_0 - 1)/k$. Of all the gases listed in that table, carbon dioxide is best suited for a gas

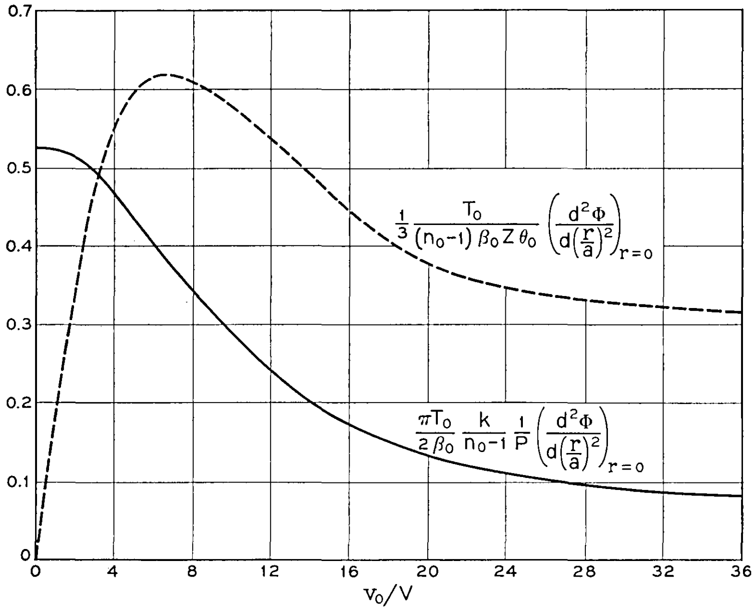


Fig. 13 — Figure of merit given by (22a) and focusing power (as in Fig. 8) vs gas velocity.

lens. This does not mean, however, that gases with larger values of $(n_0 - 1)/k$ cannot be found.

VI. FOCAL LENGTH

We have seen that $\Delta\Phi$ varies nearly as $(r/a)^2$. In the region where the lens is weak, we may treat it as a thin lens and obtain directly a simple expression for focal length.

For any thin lens it may be shown that the focal length f is given by

$$f = \tfrac{1}{2}\beta_0(r^2/\Delta\Phi) \qquad (23)$$

where

$\Delta\Phi$ is the phase shift added on axis as compared to that for a ray at radius r

β_0 = phase constant of the region surrounding the lens.

We obtain $\Delta\Phi$ from (17), which is given by the following for the gas velocity set to maximize $\Delta\Phi$ at $(r/a) = 0.4$ — i.e., at $(v_0/V) = 6.9$:

$$\Delta\Phi = 0.839(r/a)^2(\theta_0/T_0)\beta_0 z(n_0 - 1). \qquad (24)$$

Putting (24) into (23) we obtain the following expression for the focal length of a weak gas lens:

$$f = 0.596 \frac{\lambda^2}{z} \frac{T_0}{\theta_0(n_0 - 1)}. \tag{25}$$

If $a = 0.125$ inch, $z = 5$ inches, $T_0 = 293°\text{K}$, and $\theta_0 = 20°\text{C}$, we find $f \cong 5$ feet using CO_2 as the gas and $f \cong 8$ feet using air as the gas; the power transferred to the gas with CO_2 would be 0.0775 cal/sec = 0.325 watt.

When the gas lens is not weak, one should take into account that the refractive index varies both with radial position and with longitudinal position. Work is under way to analyze this very difficult situation. A simpler approach, and one which should give a first-order answer for gas lenses operated near the velocity producing maximum $\Delta\Phi$ [see (17)], is to assume a medium within the lens

$$n(r,x) = n_a(1 - \tfrac{1}{2}a_2r^2) \tag{26}$$

where

x = distance (within lens)from start of lens
r = radius
n_a = index of refraction on the axis.

For the gas velocity $(v_0/V) = 6.9$ it may be shown that

$$a_2 = \frac{1.68}{n_a a^2} \frac{\theta_0}{T_0} (n_0 - 1). \tag{27}$$

In other unpublished work the authors have shown that the radial position of a ray (or of the axis of a Gaussian beam mode) is

$$r = r_i \cos \sqrt{a_2}x + \frac{r_i'}{\sqrt{a_2}} \sin \sqrt{a_2}x \tag{28}$$

where

r_i = displacement of ray at lens input
r_i' = slope of ray at lens input.

We can use this general result to specify the focal length of a strong (or weak) gas lens with reference to Fig. 14. All input rays with zero slope will converge to a point on the axis a distance d beyond the output face of the lens (Fig. 14), where

$$d = \frac{1}{\sqrt{a_2}} \cot (\sqrt{a_2}t), \tag{29}$$

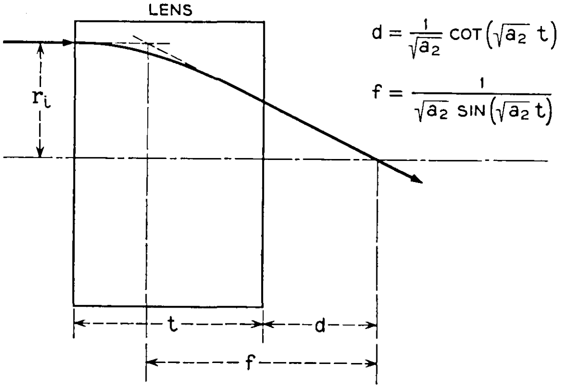

LENS
r_i
$d = \frac{1}{\sqrt{a_2}} \cot(\sqrt{a_2}\, t)$
$f = \frac{1}{\sqrt{a_2} \sin(\sqrt{a_2}\, t)}$
t
d
f

Fig. 14 — Diagram defining focal length and position of the equivalent thin lens for a thick gas lens.

t is the length of the lens, and n_a has been taken as unity. An equivalent thin lens may be located a distance f back from the focal point, where

$$f = \frac{1}{\sqrt{a_2} \sin(\sqrt{a_2} t)} . \tag{30}$$

This expression for focal length is valid up to $(\sqrt{a_2} t) = \pi/2$, at which point $d = 0$. For $(\sqrt{a_2} t) > \pi/2$ the rays cross within the lens, per (28). For $(\sqrt{a_2}\, t) \ll 1$ it may be shown that (30) passes into (25) and the location of the equivalent thin lens is in the center of the distributed lens.

E. A. J. Marcatili[4] has solved Maxwell's equations for a medium characterized by (26) and has found the normal modes for a sequence of lenses composed of segments of such a medium. This work relates closely to a sequence of gas lenses described in this paper. Experiments with tubular thermal gas lenses are reported by A. C. Beck.[5]

VII. CONCLUSION

When a cool gas is blown through a warmer tube, the gas at the axis has a lower temperature than that near the walls. Thus the density and refractive index is larger at the axis and a converging lens is formed. If the tube were at a lower temperature than the input gas, a diverging lens would be formed.

There is an optimum gas velocity for maximizing the focusing power of such lenses, and expressions are given for this velocity. It turns out that the optimum transit time for gas through the tube is approximately the time constant for temperature changes in a gas at rest in the tube, which for typical gases (air and carbon dioxide) is about 0.1 second in a $\frac{1}{4}$-inch ID tube. Although not discussed in this paper, it is found that a $\frac{1}{4}$-inch tube 6 inches long yields (at the optimum velocity for focusing power) a Reynolds number well below that at which turbulence is expected.

Expressions are given for focal length and a figure of merit expressed as focusing power per watt of power transferred to the moving gas.

The best gas is one with a maximum $(n - 1)/k$, where n is the refractive index and k is the heat conductivity.

APPENDIX A

Derivation of (10)

The relation (10)

$$V/v_0 = 0.173(t_0/\tau) \tag{31}$$

can be derived as follows.

The time development of a cool gas resting in a tube of wall temperature T_w can be described as follows

$$T = T_w - 2(T_w - T_0) \sum_{n=1}^{\infty} e^{-\lambda_n t} \frac{J_0\left(w_n \frac{r}{a}\right)}{w_n J_1(w_n)} \tag{32}$$

with

$$\lambda_n = (k/a^2 \rho c_p) w_n^2 \quad \text{and} \quad J_0(w_n) = 0.$$

At $t = 0$, (32) becomes $T(r,0) = T_0$, which is constant throughout the tube's cross section.

As time progresses the exponents $\lambda_n t$ become large, so that very soon the first term of the series is the only contributing factor. Neglecting all the terms except the first, we get

$$\frac{1}{\tau} = \frac{\frac{dT(0,t)}{dt}}{T_w - T(0,t)} \cong \lambda_1 = \frac{k}{a^2 \rho c_p} w_1^2 = 5.79 \frac{V}{z} \tag{33}$$

and substituting $z = v_0 t_0$ we get (31). Since we neglected all but the first term in the series, (33) represents the asymptotic value which $1/\tau$ assumes after the initial transients have died down.

To obtain a feeling for the accuracy of the approximation involved in deriving (33), we write down the ratio of the second to the first term in the sum (32) for $r = 0$:

$$\frac{w_1 J_1(w_1)}{w_2 J_1(w_2)} \exp[-(\lambda_2 - \lambda_1)t] = 0.666 \exp\left(-4.26 \frac{t}{\tau}\right).$$

This ratio is 10^{-2} for $t/\tau = 0.986$ and is 10^{-1} for $t/\tau = 0.45$. The approximation is excellent for times $t \geqq \tau$ and is quite good for $t > 0.5\tau$.

For the special example used in Fig. 10 we get for

$$\begin{aligned} CO_2 &: \tau = 0.161 \text{ sec} \\ \text{air} &: \tau = 0.08 \text{ sec.} \end{aligned}$$

APPENDIX B

The Maximum of (17)

We seek an expression for the value of v_0/V which brings $\Delta\Phi$, equation (17), to a maximum:

$$\begin{aligned} \frac{d(\Delta\Phi)}{d(v_0/V)} &= \beta_0 z(n_0 - 1) \frac{\theta_0}{T_0} \\ &\cdot \left[0.202(1 - R_0) \left\{ 1 - \exp\left(-7.316 \frac{V}{v_0}\right) \left(1 + 7.316 \frac{V}{v_0}\right) \right\} \right. \\ &- 0.0183(1 - R_1) \left\{ 1 - \exp\left(-44.3 \frac{V}{v_0}\right) \left(1 + 44.3 \frac{V}{v_0}\right) \right\} \\ &\left. + 0.00363(1 - R_2) \left\{ 1 - \exp\left(-106 \frac{V}{v_0}\right) \left(1 + 106 \frac{V}{v_0}\right) \right\} \right]. \end{aligned} \tag{34}$$

We set (34) equal to zero, and noting that the second and third exponentials are small, we neglect them (to be justified by the solutions thus obtained) yielding

$$\begin{aligned} \exp\left(-7.316 \frac{V}{v_0}\right) &\left(1 + 7.316 \frac{V}{v_0}\right) \\ &= 1 - 0.0906 \frac{(1 - R_1)}{(1 - R_0)} + 0.018 \frac{(1 - R_2)}{(1 - R_0)}. \end{aligned} \tag{35}$$

Equation (35) gives the approximate value of v_0/V at the maximum of $\Delta\Phi$ for any chosen radius, r/a. Additional terms in (34) can be taken if more accuracy is desired, which was done in computing the v_0/V for $r/a = 0.2$, as given in the body of the paper.

REFERENCES

1. Goubau, G., and Schwering, F., On the Guided Propagation of Electromagnetic Wave Beams, I.R.E. Trans. **AP-9,** May, 1961, pp. 248–256.
2. Berreman, D. W., B.S.T.J., this issue, pp. 1469 and 1476.
3. Jakob, M., *Heat Transfer*, Vol. 1, John Wiley, New York, 1949, pp. 451–464.
4. Marcatili, E. A. J., Wave Optics in a Periodic Sequence of Arbitrarily Thick Lens-Like Focusers, to be published.
5. Beck, A. C., Thermal Gas Lens Measurements, B.S.T.J., this issue, p. 1818.

Hollow Metallic and Dielectric Waveguides for Long Distance Optical Transmission and Lasers

By E. A. J. MARCATILI and R. A. SCHMELTZER

(Manuscript received June 12, 1964)

The field configurations and propagation constants of the normal modes are determined for a hollow circular waveguide made of dielectric material or metal for application as an optical waveguide. The increase of attenuation due to curvature of the axis is also determined.

The attenuation of each mode is found to be proportional to the square of the free-space wavelength λ *and inversely proportional to the cube of the cylinder radius a. For a hollow dielectric waveguide made of glass with* $\nu = 1.50$, $\lambda = 1\mu$, *and* $a = 1$ *mm, an attenuation of 1.85 db/km is predicted for the minimum-loss mode,* EH_{11}. *This loss is doubled for a radius of curvature of the guide axis* $R \approx 10$ *km. Hence, dielectric materials do not seem suitable for use in hollow circular waveguides for long distance optical transmission because of the high loss introduced by even mild curvature of the guide axis. Nevertheless, dielectric materials are shown to be very attractive as guiding media for gaseous amplifiers and oscillators, not only because of the low attenuation but also because the gain per unit length of a dielectric tube containing He-Ne "masing" mixture at the right pressure can be considerably enhanced by reducing the tube diameter. In this application, a small guide radius is desirable, thereby making the curvature of the guide axis not critical. For* $\lambda = 0.6328\mu$ *and optimum radius* $a = 0.058$ *mm, a maximum theoretical gain of 7.6 db/m is predicted.*

It is shown that the hollow metallic *circular waveguide is far less sensitive to curvature of the guide axis. This is due to the comparatively large complex dielectric constant exhibited by metals at optical frequencies. For a wavelength* $\lambda = 1\mu$ *and a radius* $a = 0.25$ *mm, the attenuation for the minimum loss* TE_{01} *mode in an aluminum waveguide is only 1.8 db/km. This loss is doubled for a radius of curvature as short as* $R \approx 48$ *meters. For* $\lambda = 3\mu$ *and* $a = 0.6$ *mm, the attenuation of the* TE_{01} *mode is also 1.8 db/km. The radius of curvature which doubles this loss is approximately 75 meters. The* *straight guide loss for the EH_{11} mode for $\lambda = 1\mu$ and $a = 0.25$ mm is 57 db/km and is increased to 320 db/km for $\lambda = 3\mu$ and $a = 0.6$ mm.*

In view of the low-loss characteristic of the TE_{01} mode in metallic waveguides, the high-loss discrimination of noncircular electric modes, and the relative insensitivity to axis curvature, the hollow metallic *circular waveguide appears to be very attractive as a transmission medium for long distance optical communication.*

I. INTRODUCTION

During recent years the potentially large frequency range made available to communications by the development of the optical maser has stimulated much interest in efficient methods for long distance transmission of light. The most promising contenders for long distance optical transmission media consist of sequences of lenses or mirrors, highly reflective hollow metallic pipes, and dielectric waveguides.[1-10]

In this paper we present an analysis of the field configurations and propagation constants of the normal modes in a hollow circular waveguide which, because of its simplicity and low loss, may become an important competitor. The guiding structure considered here may consist of an ordinary metallic pipe of precision bore whose inner surface is highly reflective, or of a hollow dielectric pipe — i.e., one in which the metal is replaced with dielectric. Although the transmission characteristics of metallic waveguides are well known for microwave frequencies, this theory is invalidated for operation at optical wavelengths, because the metal no longer acts as a good conductor but rather as a dielectric having a large dielectric constant. In the subsequent analysis, therefore, both the dielectric and metallic guide are considered as special cases of a general hollow circular waveguide having an external medium made of arbitrary isotropic material whose optical properties are characterized by a finite complex refractive index. If the free-space wavelength is much smaller than the internal radius of the tube, the energy propagates not in the external medium but essentially within the tube, bouncing at grazing angles against the wall. Consequently, there is little energy loss due to refraction. The refracted field is partially reflected by the external surface of the tube and may, in general, interfere constructively or destructively with the field inside the tube, decreasing or increasing the attenuation. Because of the difficulty of controlling the interference paths, it seems more convenient to eliminate the effect completely by introducing sufficient loss in the dielectric or, in the case of a glass dielectric, by frosting the external surface. The field in the hole of the tube is then unaffected by wall thickness. We shall therefore simplify the analysis of the hollow circular waveguide by assuming infinite wall thickness, as depicted in Fig. 1.

This structure will be shown to be attractive as a low-loss transmission medium for long distance optical communication as well as for optical gaseous amplifiers and oscillators. It is known, for example, that in a tube containing a He-Ne mixture such that the product of radius and pressure is roughly a constant, the gain per unit length is inversely proportional to the radius of the tube.[11] On the other hand, we find in this paper that the attenuation of the normal modes is inversely proportional to the cube of the radius. Hence there is an optimum tube radius for which the net gain per unit length is a maximum. Furthermore, because the guidance is continuous, there is no need for periodic focusing. Consequently, no restriction need be imposed on the length of the amplifying or oscillating tube.

We begin by analyzing an idealized guide having a straight axis and a cylindrical wall. The results are then extended to include the effects of mild curvature of the guide axis by finding a perturbation correction for field configurations and propagation constants of the idealized straight guide.

II. MODAL ANALYSIS OF THE GENERAL STRAIGHT CIRCULAR WAVEGUIDE

Consider a waveguide consisting of a circular cylinder of radius a and free-space dielectric constant ϵ_0 embedded in another medium of dielectric or metal having a complex dielectric constant ϵ. The magnetic permeability μ_0 is assumed to be that of free space for both media. We are interested in finding the field components of the normal modes of the waveguide and in determining the complex propagation constants of these modes.

The problem is substantially simplified if it is assumed that

$$ka = 2\pi a/\lambda \gg |\nu| u_{nm} \tag{1}$$

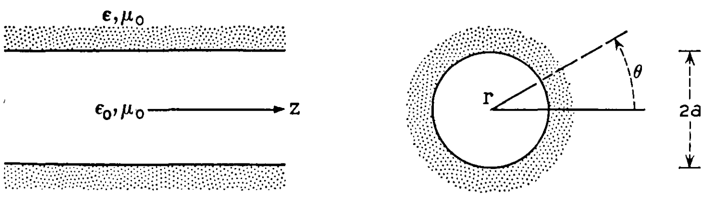


Fig. 1 — Hollow dielectric waveguide.

and

$$|\,(\gamma/k) - 1\,| \ll 1 \tag{1}$$

where $k = \omega\sqrt{\epsilon_0\mu_0} = 2\pi/\lambda$ is the free-space propagation constant; u_{nm} is the mth root of the equation $J_{n-1}(u_{nm}) = 0$, and n and m are integers that characterize the propagating mode; $\nu = \sqrt{\epsilon/\epsilon_0}$ is the complex refractive index of the external medium; and γ is the axial propagation constant of the mode under consideration. The first inequality states that the radius a is much larger than the free-space wavelength λ. In the case of metalization of the external medium, $|\,\nu\,|$ may be quite large but is finite at optical frequencies. The second inequality restricts our analysis to low-loss modes, which are those whose propagation constants γ are nearly equal to that of free space.

The field components of the natural modes of the most general circular cylindrical structure with arbitrary isotropic internal and external media have been determined by Stratton.[12] This structure supports three types of modes: first, transverse circular electric modes whose only field components are E_θ, H_r and H_z; second, transverse circular magnetic modes whose components are H_θ, E_r and E_z; and third, hybrid modes with all the electric and magnetic components present. The approximate field components of these modes are written below. They have been derived using the inequalities (1) and neglecting terms with powers of λ/a larger than one. The superscripts i and e refer to the internal and external media, respectively.

1. Circular electric modes TE_{0m} $(n = 0)$

$$\left.\begin{array}{l} \left.\begin{array}{l} E_{\theta 0m}{}^{i} = J_1(k_i r) \\ H_{r0m}{}^{i} = -\sqrt{\dfrac{\epsilon_0}{\mu_0}}\, J_1(k_i r) \\ H_{z0m}{}^{i} = -i\sqrt{\dfrac{\epsilon_0}{\mu_0}}\, \dfrac{u_{0m}}{ka}\, J_0(k_i r) \end{array}\right\} \exp i(\gamma z - \omega t) \\ \left.\begin{array}{l} E_{\theta 0m}{}^{e} = -1 \\ H_{r0m}{}^{e} = \sqrt{\dfrac{\epsilon_0}{\mu_0}} \\ H_{z0m}{}^{e} = -i\sqrt{\nu^2 - 1}\,\sqrt{\dfrac{\epsilon_0}{\mu_0}} \end{array}\right\} \begin{array}{l} i\,\dfrac{u_{0m}}{k\sqrt{ar(\nu^2 - 1)}}\, J_0(u_{0m}) \\ \exp i[k_e(r - a) + \gamma z - \omega t] \end{array} \end{array}\right\}. \tag{2}$$

2. Circular magnetic modes TM_{0m} $(n = 0)$

$$\left.\begin{aligned}
&\left.\begin{aligned}
E_{r0m}{}^{i} &= J_1(k_i r)\\
E_{z0m}{}^{i} &= i\,\frac{u_{0m}}{ka}\,J_0(k_i r)\\
H_{\theta 0m}{}^{i} &= \sqrt{\frac{\epsilon_0}{\mu_0}}\,J_1(k_i r)
\end{aligned}\right\}\exp i(\gamma z - \omega t)\\
&\left.\begin{aligned}
E_{r0m}{}^{e} &= -\frac{1}{\nu^2}\\
E_{z0m}{}^{e} &= \sqrt{\nu^2 - 1}\\
H_{\theta 0m}{}^{e} &= -\sqrt{\frac{\epsilon_0}{\mu_0}}
\end{aligned}\right\}\; i\,\frac{u_{0m}J_0(u_{0m})}{k\sqrt{ar(\nu^2 - 1)}}\exp i\,[k_e(r-a) + \gamma z - \omega t]
\end{aligned}\right\}. \tag{3}$$

3. Hybrid modes EH_{nm} $(n \neq 0)$

$$\left.\begin{aligned}
&\left.\begin{aligned}
E_{\theta nm}{}^{i} &= \left[J_{n-1}\,(k_i r) + \frac{iu_{nm}{}^{2}}{2nka}\sqrt{\nu^2 - 1}\,J'_n(k_i r)\right]\\
&\qquad\qquad\qquad\qquad\cdot\cos n(\theta + \theta_0)\\
E_{rnm}{}^{i} &= \left[J_{n-1}\,(k_i r) + \frac{iu_{nm}}{2kr}\sqrt{\nu^2 - 1}\,J_n(k_i r)\right]\\
&\qquad\qquad\qquad\qquad\cdot\sin n(\theta + \theta_0)\\
E_{znm}{}^{i} &= -i\,\frac{u_{nm}}{ka}\,J_n(k_i r)\,\sin n(\theta + \theta_0)
\end{aligned}\right\}\exp i(\gamma z - \omega t)\\
&H_{\theta nm}{}^{i} = \sqrt{\frac{\epsilon_0}{\mu_0}}\,E_{rnm}{}^{i}\\
&H_{rnm}{}^{i} = -\sqrt{\frac{\epsilon_0}{\mu_0}}\,E_{\theta nm}{}^{i}\\
&H_{znm}{}^{i} = -\sqrt{\frac{\epsilon_0}{\mu_0}}\,E_{znm}{}^{i}\,\mathrm{ctn}\,n(\theta + \theta_0)\\
&\left.\begin{aligned}
E_{\theta nm}{}^{e} &= \cos n(\theta + \theta_0)\\
E_{rnm}{}^{e} &= \sin n(\theta + \theta_0)\\
E_{znm}{}^{e} &= -\sqrt{\nu^2 - 1}\,\sin n(\theta + \theta_0)
\end{aligned}\right\}\begin{aligned} &i\,\frac{u_{nm}}{k\sqrt{ar(\nu^2 - 1)}}\,J_n(u_{nm})\\ &\cdot\exp i[k_e(r - a) + \gamma z - \omega t]\end{aligned}\\
&H_{\theta nm}{}^{e} = \nu^2\sqrt{\frac{\epsilon_0}{\mu_0}}\,E_{rnm}{}^{e}\\
&H_{rnm}{}^{e} = -\sqrt{\frac{\epsilon_0}{\mu_0}}\,E_{\theta nm}{}^{e}\\
&H_{znm}{}^{e} = -\sqrt{\frac{\epsilon_0}{\mu_0}}\,E_{znm}{}^{e}\,\mathrm{ctn}\,n(\theta + \theta_0)
\end{aligned}\right\} \tag{4}$$

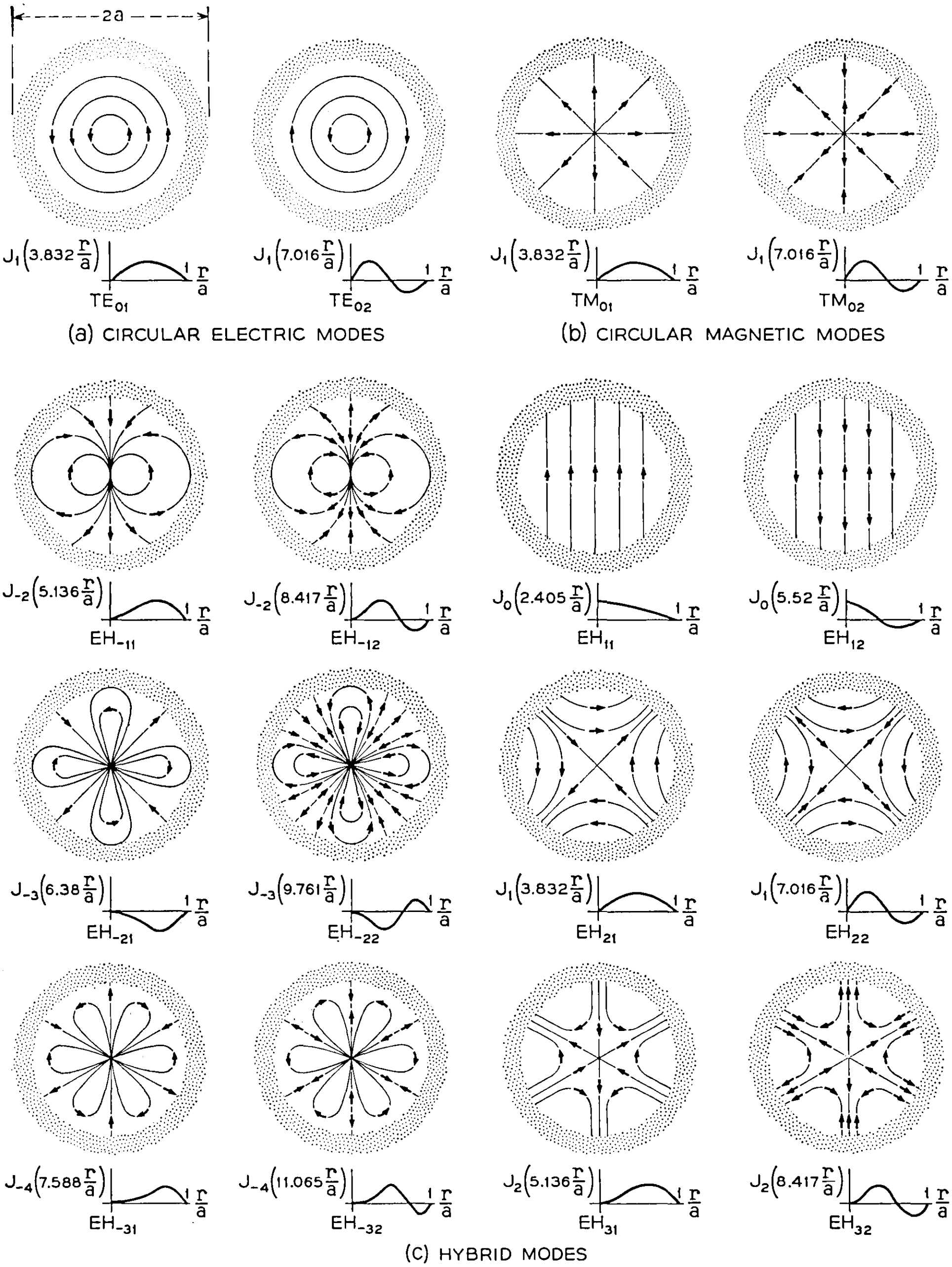

2a
$J_1(3.832\frac{r}{a})$ 1 $\frac{r}{a}$
TE_{01}
$J_1(7.016\frac{r}{a})$ 1 $\frac{r}{a}$
TE_{02}
$J_1(3.832\frac{r}{a})$ 1 $\frac{r}{a}$
TM_{01}
$J_1(7.016\frac{r}{a})$ 1 $\frac{r}{a}$
TM_{02}
(a) CIRCULAR ELECTRIC MODES
(b) CIRCULAR MAGNETIC MODES
$J_{-2}(5.136\frac{r}{a})$ 1 $\frac{r}{a}$
EH_{-11}
$J_{-2}(8.417\frac{r}{a})$ 1 $\frac{r}{a}$
EH_{-12}
$J_0(2.405\frac{r}{a})$ 1 $\frac{r}{a}$
EH_{11}
$J_0(5.52\frac{r}{a})$ 1 $\frac{r}{a}$
EH_{12}
$J_{-3}(6.38\frac{r}{a})$ 1 $\frac{r}{a}$
EH_{-21}
$J_{-3}(9.761\frac{r}{a})$ 1 $\frac{r}{a}$
EH_{-22}
$J_1(3.832\frac{r}{a})$ 1 $\frac{r}{a}$
EH_{21}
$J_1(7.016\frac{r}{a})$ 1 $\frac{r}{a}$
EH_{22}
$J_{-4}(7.588\frac{r}{a})$ 1 $\frac{r}{a}$
EH_{-31}
$J_{-4}(11.065\frac{r}{a})$ 1 $\frac{r}{a}$
EH_{-32}
$J_2(5.136\frac{r}{a})$ 1 $\frac{r}{a}$
EH_{31}
$J_2(8.417\frac{r}{a})$ 1 $\frac{r}{a}$
EH_{32}
(C) HYBRID MODES

Fig. 2 — Electric field lines of modes in hollow dielectric waveguides: (a) circular electric modes, (b) circular magnetic modes, (c) hybrid modes.

where the complex propagation constant γ satisfies the relationships

$$\begin{aligned} k_i^2 &= k^2 - \gamma^2 \\ k_e^2 &= \nu^2 k^2 - \gamma^2 \end{aligned} \tag{5}$$

and u_{nm} is the mth root of the equation

$$J_{n-1}(u_{nm}) = 0. \tag{6}$$

As usual, $|\, n \,|$ is the number of periods of each field component in the θ direction, and m is both the order of the root of (6) and the number of maxima and minima of each component counted in the radial direction within the internal medium. The constant θ_0 appearing in (4) will become of interest later on when we study the waveguide with curved axis, because it will admit any orientation of the transverse electric field relative to the plane of curvature of the guide axis.

For $n = 0$, the modes are either transverse electric TE_{0m} (2), or transverse magnetic TM_{0m} (3). The lines of electric field of the TE_{0m} modes are transverse concentric circles centered on the z axis. The lines of magnetic field are in planes containing the z axis. Similarly, the lines of magnetic field of the TM_{0m} modes are transverse concentric circles centered on the z axis with the electric field contained in radial planes. The electric field lines of the modes TE_{01}, TE_{02}, TM_{01} and TM_{02} are shown in Figs. 2(a) and 2(b); each vector represents qualitatively the intensity and direction of the local field.

For $n \neq 0$, the modes are hybrid, EH_{nm} (4); therefore, the magnetic and electric field are three-dimensional with relatively small axial field components in the internal medium. Thus the hybrid modes are almost transverse.

Let us examine the projection of these three-dimensional field lines on planes perpendicular to the axis z of the waveguide. The differential equations for the projected lines of electric field in both media are

$$\begin{aligned} \frac{1}{r}\frac{dr}{d\theta} &= \frac{E_{rnm}{}^i}{E_{\theta nm}{}^i} \\ \frac{1}{r}\frac{dr}{d\theta} &= \frac{E_{rnm}{}^e}{E_{\theta nm}{}^e}. \end{aligned} \tag{7}$$

$E_{rnm}{}^i$ as well as $E_{\theta nm}{}^i$ contain two terms as given in (4). Both are necessary to satisfy the boundary conditions. If we neglect the second term, however, no substantial error is introduced except very close (a few wavelengths) to the boundary, where the second term dominates as the first tends to zero. With this simplification, the differential equations (7) in both media become identical

$$(1/r)(dr/d\theta) = \tan n\theta.$$

Upon integrating, one obtains an equation for the locus of the projected electric field lines

$$(r/r_0)^n \cos n\theta = 1 \tag{8}$$

where r_0 is a constant of integration that individualizes the member of the family of lines. The electric field of an EH_{nm} mode is different from that of EH_{-nm} mode.

The projection of the magnetic field lines is determined in a similar way. These equations are

$$(r/r_0)^n \sin n\theta = 1 \tag{9}$$

for the internal medium and

$$(r/r_0)^{n\nu^2} \sin n\theta = 1$$

for the external medium.

The projections of the internal electric (8) and magnetic (9) field lines are identical for any given mode except for a rotation of $\pi/(2n)$ radians around the z axis. In Fig. 2(c) the lines of the electric field in the internal medium are depicted for the first few hybrid modes. Again the vectors represent qualitatively the field intensities and directions.

What happens at the boundary? Consider, for example, the projected electric lines of mode EH_{11}, as shown in Fig. 3(a). These field lines satisfy (8), an equation which is valid everywhere except near the boundary. The boundary conditions are violated in Fig. 3(a) because there is continuity not only of the tangential electric component but also of the normal component. The internal normal component must be ν^2 times larger than the external one. Consequently, the electric field line must be discontinuous. This result is shown qualitatively in Fig. 3(b).

A three-dimensional representation of the field lines is far more complicated than the two-dimensional one depicted in Fig. 2. As a typical example, the electric field lines of the EH_{22} mode are shown in Fig. 4 in a three-dimensional perspective.

The propagation constants of the TE_{0m}, TM_{0m} and EH_{nm} $(n \neq 0)$ modes are determined below (21). It is found that the hybrid mode $\mathrm{EH}_{-|n|,m}$ is degenerate (same propagation constant) with the $\mathrm{EH}_{|n|+2,m}$; i.e., for every hybrid mode with negative azimuthal index there is a degenerate hybrid mode with positive aximuthal index. The

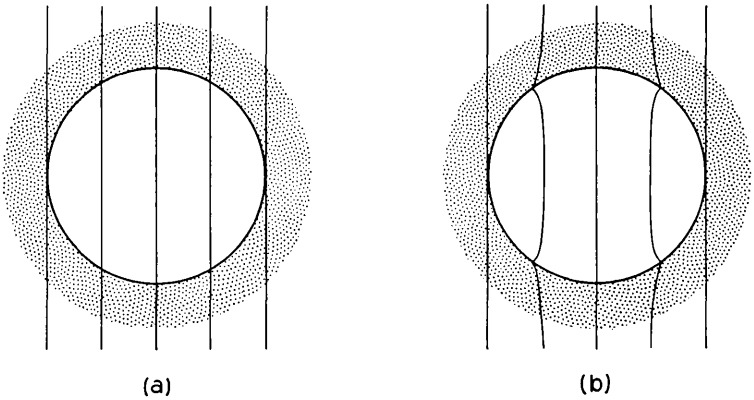


Fig. 3 — (a) Electric field lines of EH_{11} mode violating boundary conditions; (b) same EH_{11} mode with electric field lines qualitatively corrected.

transverse modes TE_{0m} and TM_{0m} and the hybrid modes EH_{1m} and EH_{2m} have no degenerate counterpart.

If the field components of the degenerate $EH_{-|n|,m}$ and $EH_{|n|+2,m}$ modes (4) are added, we obtain new composite modes whose electric and magnetic field lines project as straight lines on a plane perpendicular to the z axis. Some of those composite modes are shown in Fig. 5.

It should be noted that if the refractive index of the external medium,

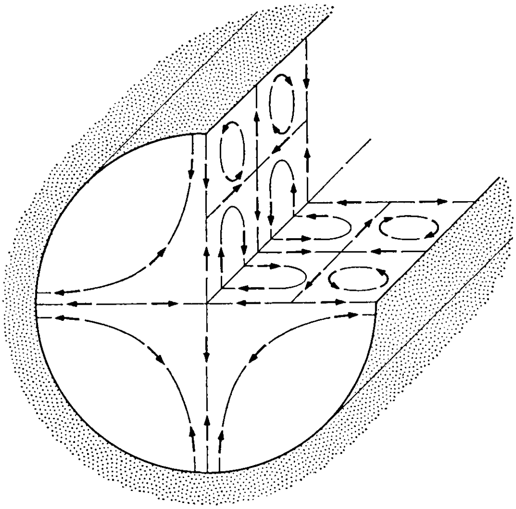


Fig. 4 — Cutaway view of electric field lines of EH_{22} mode. The axial period is grossly exaggerated.

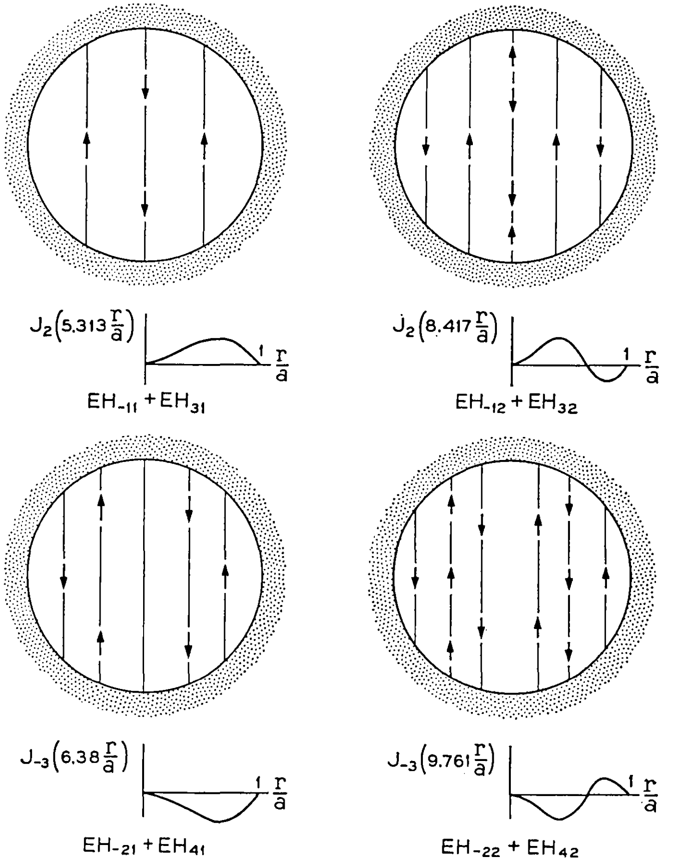


Fig. 5 — Electric field lines of composite modes $\mathrm{EH}_{-|n|,m} + \mathrm{EH}_{|n|+2,m}$.

ν, is very close to unity, then for each value of m, the TE_{0m} , TM_{0m} and EH_{2m} modes also become degenerate (17), (21) and the sum of the components of TE_{0m} (2) and EH_{2m} (4) yields a new composite mode, as shown in Fig. 6. This mode, together with those in Fig. 5 and the EH_{1m} of Fig. 2(c), form a complete set that closely resembles the set found for interferometers with plane circular mirrors or for sequences of circular irises.[1]

Let us now consider the field intensity distribution outside and inside the hollow dielectric waveguide. The external field (2), (3) and (4) has the radial dependence

$$\frac{\exp\left[ik_e(r - a)\right]}{\sqrt{r}} \cdot$$

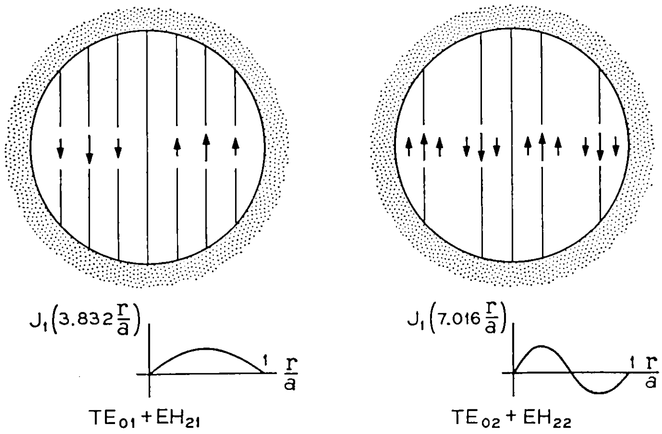


Fig. 6 — Electric field lines of composite modes $TE_{0m} + EH_{2m}$.

From (5) and (20) we obtain, neglecting terms of order $(\lambda/a)^2$ and higher, $k_e = k\sqrt{\nu^2 - 1}$. The radial dependence is then

$$\frac{\exp\left[ik\sqrt{\nu^2 - 1}(r - a)\right]}{\sqrt{r}}.$$

If the dielectric is lossy, the refractive index ν has a positive imaginary part. The external electric and magnetic fields then oscillate with period of the order of $\lambda / \mid \sqrt{\nu^2 - 1} \mid$ and decay exponentially in the radial direction. The maximum field intensities in the external medium occur at the boundary $r = a$. Being proportional to λ/a, these maxima are small.

The field intensity inside the hollow waveguide is more interesting. Again if we substitute γ (20) into (2), (3) and (4) and neglect terms of the order λ/a, only the internal transverse components remain.

For TE_{0m} modes

$$E_{\theta 0m}{}^i = -\sqrt{\frac{\mu_0}{\epsilon_0}}\, H_{r0m}{}^i = J_1\left(u_{0m}\,\frac{r}{a}\right). \tag{10}$$

For TM_{0m} modes,

$$E_{r0m}{}^i = \sqrt{\frac{\mu_0}{\epsilon_0}}\, H_{\theta 0m}{}^i = J_1\left(u_{0m}\,\frac{r}{a}\right). \tag{11}$$

For EH_{nm} modes,

$$\begin{aligned} E_{\theta nm}{}^{i} &= -\sqrt{\frac{\mu_0}{\epsilon_0}}\, H_{rnm}{}^{i} = J_{n-1}\left(u_{nm}\frac{r}{a}\right)\cos n\theta \\ E_{rnm}{}^{i} &= \sqrt{\frac{\mu_0}{\epsilon_0}}\, H_{\theta nm}{}^{i} = J_{n-1}\left(u_{nm}\frac{r}{a}\right)\sin n\theta. \end{aligned} \tag{12}$$

The field components of each mode have approximately the same radial dependence, varying as Bessel functions of the first kind, and tending to negligibly small values at the boundary (6). This approximate radial dependence (10), (11) and (12) is reproduced under each mode pattern in Figs. 2(a), 2(b) and 2(c).

III. PROPAGATION CONSTANTS FOR THE GENERAL CIRCULAR CYLINDRICAL GUIDE

In this section we shall determine the propagation constants γ, of the TE_{0m}, TM_{0m} and EH_{nm} modes in the straight hollow guide at optical wavelengths. The propagation constants are the roots of the following characteristic equation for the general circular cylindrical structure.[12] They are related to k_i and k_e by expressions (5).

$$\left[\frac{J_n'(k_ia)}{J_n(k_ia)} - \frac{k_i}{k_e}\frac{H_n^{(1)'}(k_ea)}{H_n^{(1)}(k_ea)}\right]\left[\frac{J_n'(k_ia)}{J_n(k_ia)} - \frac{\nu^2k_i}{k_e}\frac{H_n^{(1)'}(k_ea)}{H_n^{(1)}(k_ea)}\right] = \left[\frac{n\lambda}{kk_ia}\right]^2\left[1 - \left(\frac{k_i}{k_e}\right)\right]^2. \tag{13}$$

This equation is simplified substantially when the approximations in (1) are introduced. Since $k_ea \gg 1$, the asymptotic value of the Hankel functions may be used

$$\frac{H_n^{(1)'}(k_ea)}{H_n^{(1)}(k_ea)} \approx i + 0(1/k_ea), \qquad k_ea \gg 1. \tag{14}$$

Since

$$\frac{\nu^2}{k_ea} \approx \frac{\nu^2}{(\nu^2 - 1)^{\frac{1}{2}}}\left(\frac{\lambda}{2\pi a}\right) \ll 1 \tag{15}$$

powers of ν^2/k_ea larger than one shall be neglected. The characteristic equation then simplifies to

$$J_{n-1}(k_ia) = i\nu_n(k_i/k)J_n(k_ia) \tag{16}$$

where

$$\nu_n = \begin{cases} \dfrac{1}{\sqrt{\nu^2 - 1}} & \text{for TE}_{0m} \text{ modes } (n = 0) \\ \dfrac{\nu^2}{\sqrt{\nu^2 - 1}} & \text{for TM}_{0m} \text{ modes } (n = 0) \\ \dfrac{\frac{1}{2}(\nu^2 + 1)}{\sqrt{\nu^2 - 1}} & \text{for EH}_{nm} \text{ modes } (n \neq 0). \end{cases} \tag{17}$$

To solve the characteristic equation for k_ia we notice that because of (1) and (5), the right-hand side of (16) is close to zero. Using a perturbation technique and keeping only the first term of the perturbation,

$$k_ia \approx u_{nm}(1 - i\nu_n/ka) \tag{18}$$

where u_{nm} as before is the mth root of the equation

$$J_{n-1}(u_{nm}) = 0. \tag{19}$$

The validity of (18) is assured provided that the order of the mode is low enough so that $|\nu_n| u_{nm} \ll ka$. The propagation constants γ can then be obtained from (5)

$$\gamma \approx k\left[1 - \frac{1}{2}\left(\frac{u_{nm}\lambda}{2\pi a}\right)^2\left(1 - \frac{i\nu_n\lambda}{\pi a}\right)\right]. \tag{20}$$

The phase constant and attenuation constant of each mode are the real and imaginary parts of γ, respectively,

$$\begin{aligned} \beta_{nm} &= \operatorname{Re}(\gamma) = \frac{2\pi}{\lambda}\left\{1 - \frac{1}{2}\left[\frac{u_{nm}\lambda}{2\pi a}\right]^2\left[1 + \operatorname{Im}\left(\frac{\nu_n\lambda}{\pi a}\right)\right]\right\} \\ \alpha_{nm} &= \operatorname{Im}(\gamma) = \left(\frac{u_{nm}}{2\pi}\right)^2 \frac{\lambda^2}{a^3} \operatorname{Re}(\nu_n). \end{aligned} \tag{21}$$

IV. PROPAGATION CONSTANTS FOR STRAIGHT DIELECTRIC GUIDES

For guides made of dielectric material, ν_n is usually real and independent of λ, so that the phase and attenuation constants are

$$\left.\begin{aligned}
\beta_{nm} &= \frac{2\pi}{\lambda}\left\{1 - \frac{1}{2}\left(\frac{u_{nm}\lambda}{2\pi a}\right)^2\right\} \\
\alpha_{nm} &= \left(\frac{u_{nm}}{2\pi}\right)^2 \frac{\lambda^2}{a^3}\begin{cases} \dfrac{1}{\sqrt{\nu^2 - 1}}, & \text{for } TE_{0m} \text{ modes } (n = 0) \\[2ex] \dfrac{\nu^2}{\sqrt{\nu^2 - 1}}, & \text{for } TM_{0m} \text{ modes } (n = 0) \\[2ex] \dfrac{\frac{1}{2}(\nu^2 + 1)}{\sqrt{\nu^2 - 1}}, & \text{for } EH_{nm} \text{ modes } (n \neq 0) \end{cases}
\end{aligned}\right\}. \tag{22}$$

The phase constant of modes in hollow dielectric waveguides have the same frequency dependence as modes in perfectly conducting metallic waveguides when operating far from cutoff; both transmission media are then similarly dispersive.

The attenuation constants are proportional to λ^2/a^3. Consequently, the losses can be made arbitrarily small by choosing the radius of the tube a sufficiently large relative to the wavelength λ.

The refractive index ν affects the attenuation of each of the three types of modes (22) in different ways. This fact is reasonable on physical grounds. TE_{0m} modes can be considered to be composed of plane wavelets, each impinging at grazing angle on the interface between the two media with polarization perpendicular to the plane of incidence. It is known from the laws of refraction that the larger the value of ν, the smaller the refracted power.

TM_{0m} modes may also be thought of as consisting of plane wavelets, but with the electric field of each now contained in the plane of incidence. For ν very close to unity, there is little reflection and the refracted loss is high; as the value of ν is allowed to become large, each wavelet gets close to the Brewster angle of incidence and again the refracted loss is high. The minimum occurs for $\nu = \sqrt{2}$.

EH_{nm} modes are composed of both types of plane wavelets. Therefore, as is reasonable from the above argument, the attenuation constant α_{nm} has a ν dependence which is an average of those of TE_{0m} and TM_{0m} modes. The value of ν that minimizes α_{nm} is $\nu = \sqrt{3} = 1.73$.

The attenuation constants (22) are proportional to u_{nm}^2. Some values of u_{nm} (19) are presented in Table I. For a fixed value of n the attenuation constant increases with m. This statement is not true for m fixed and n variable.

Comparing the attenuation constants (22) of the different modes, we find that the mode with lowest attenuation is TE_{01} if $\nu > 2.02$ and EH_{11} if $\nu < 2.02$. Most glasses have a refractive index $\nu \approx 1.5$, and

TABLE I — SOME VALUES OF u_{nm}

n/m	1	2	3	4
1	2.405	5.52	8.654	11.796
2 or 0	3.832	7.016	10.173	13.324
3 or −1	5.136	8.417	11.62	14.796
4 or −2	6.380	9.761	13.015	16.223

consequently for hollow glass tube EH_{11} should be preferred. The attenuation of this mode ($8686\alpha_{11}$ in db/km) has been plotted in Fig. 7 as a function of λ/a for $\nu = 1.50$ using λ as a parameter. Typically, for a wavelength $\lambda = 1\mu$ and radius $a = 1$ mm, the attenuation of the EH_{11} mode is 1.85 db/km ($\approx$3 db/mile). If the radius of the guide is doubled, the attenuation is reduced to 0.231 db/km.

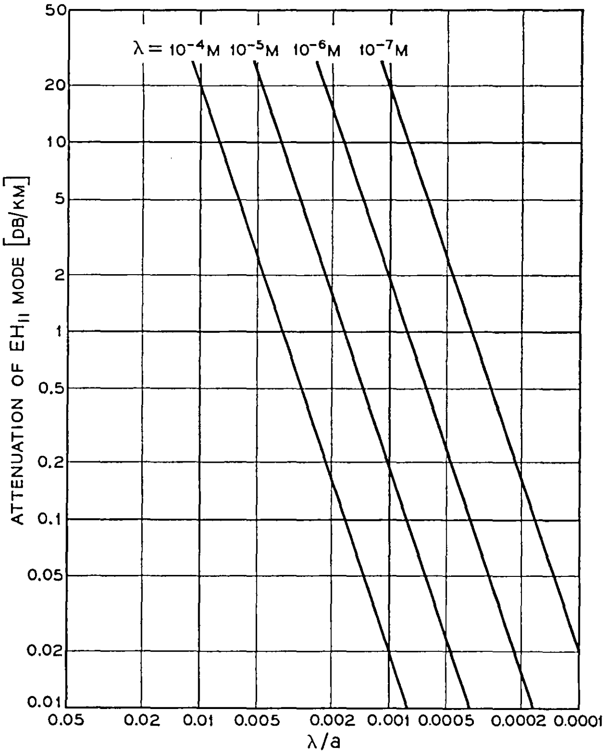


Fig. 7 — Attenuation of EH_{11} modes ($1.85\ \lambda^2/a^3$) versus wavelength/radius ($\nu = 1.5$).

V. HOLLOW DIELECTRIC WAVEGUIDE FOR OPTICAL MASER AMPLIFIERS AND OSCILLATORS

A mode traveling in a hollow dielectric waveguide filled with "masing" material experiences a net gain which is given by the difference between the amplification due to the active medium and the loss due to leakage through the walls. It has been shown[11] that in a tube filled with the right mixture of He and Ne at the proper pressure, the gain G is inversely proportional to the radius a of the tube. Then

$$G = (A/a) \text{ db/m} \tag{23}$$

where the radius a is measured in meters and the constant A is

$$A = 0.00066 \text{ db.}$$

On the other hand, we have found that the transmission loss of the EH_{11} mode in the hollow waveguide with a refractive index $\nu = 1.50$ is $L = 8.686\alpha_{11}$. From (22)

$$L = B(\lambda^2/a^3) \text{ db/m} \tag{24}$$

where the constant B is

$$B = 1.85 \text{ db.}$$

The net gain per unit length is then

$$G - L = (A/a) - B(\lambda^2/a^3) \tag{25}$$

passing through a maximum at the value of the radius for which

$$\partial(G - L)/\partial a = 0.$$

The optimum radius and the maximum net gain are respectively

$$\begin{aligned} a_{\text{opt}} &= \sqrt{3\,\frac{B}{A}}\,\lambda = 91.7\lambda \\ (G - L)_{\max} &= \frac{2}{3^{\frac{3}{2}}}\,\frac{A^{\frac{3}{2}}}{B^{\frac{1}{2}}}\,\frac{1}{\lambda} = 4.81\,\frac{10^{-6}}{\lambda} \text{ db/m.} \end{aligned} \tag{26}$$

For the He-Ne mixture, $\lambda = 0.6328\ 10^{-6}$ m. Consequently

$$\begin{aligned} a_{\text{opt}} &= 0.058 \text{ mm} \\ (G - L)_{\max} &= 7.6 \text{ db/m.} \end{aligned} \tag{27}$$

Although the diameter of the tube is quite small, the gain per unit length is sufficiently large as to make hollow dielectric amplifiers and oscillators attractive for experimentation.

Present-day confocal He-Ne masers employ tubes whose approximate length and radius are 1 m and 3 mm respectively. The gain per passage (23) is 0.22 db ($\approx$5 per cent). If a hollow dielectric waveguide with an optimum radius 0.058 mm were used, the same gain would be achieved with a length of only $0.22/7.6 = 29$ mm. This presents an excellent possibility for a very compact maser.

Even for radii larger than the optimum, the hollow dielectric waveguide is still attractive. For example with $a = 0.25$ mm, the gain is 2.6 db/m, a value far larger than the gain 0.22 db/m obtained for the 3-mm radius tube commonly used for masers.

Nevertheless, for long-wavelength masers the optimum values (26) are not practical. Consider for example a tube containing an active material which amplifies at $\lambda = 10^{-4}$ m. Let us assume that the constant A is still 0.00066. Then from (26), the optimum radius and maximum gain are

$$\begin{aligned} a_{\text{opt}} &= 9.14 \text{ mm} \\ (G - L)_{\max} &= 0.0481 \text{ db/m}. \end{aligned} \tag{28}$$

The gain is very small. It could be enhanced by reducing the radius and by increasing the refractive index ν of the walls to a value much larger than 1.5. This can be accomplished if metal is used instead of dielectric, as is shown in the next section.

VI. ATTENUATION CONSTANTS FOR THE STRAIGHT METALLIC GUIDE

In order to discuss the attenuation characteristics of metallic waveguides, we shall need to have some quantitative information about the behavior of metals at optical frequencies. We examine as a typical example the optical properties of aluminum, even though this may not be the most suitable metal. The dispersion characteristics of the conductivity and relative dielectric constants of aluminum have been studied extensively by Hodgson,[13] Beattie and Conn,[14] and Schulz.[15] The data used below have been taken from a compilation of the results of these studies,[16] and is presented graphically in Fig. 8. It is evident from these dispersion curves that the dielectric constant for aluminum is much larger than for ordinary dielectrics and increases monotonically with wavelength in the range $0.3\mu < \lambda < 4.0\mu$.

The circular electric modes have the lowest loss in metallic waveguides, while the circular magnetic and hybrid modes are rapidly attenuated even for a wavelength as short as 0.3μ. The attenuation constant α_{01} for the lowest-loss TE_{01} mode is plotted in Fig. 9 for wave-

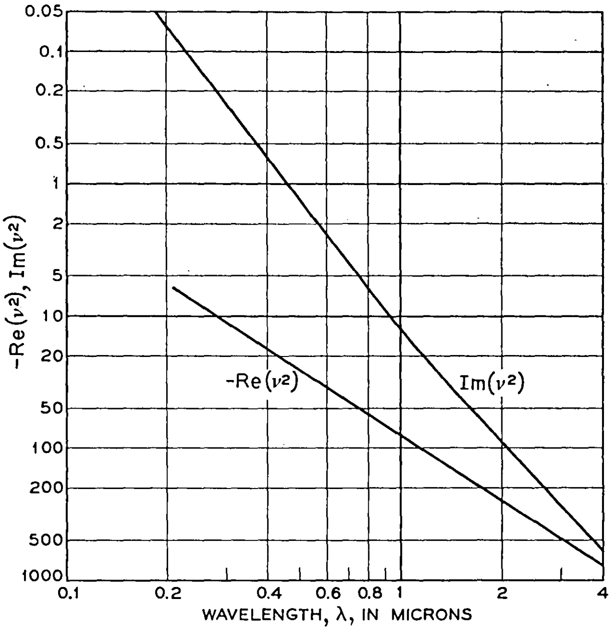


Fig. 8 — Dispersion curve for aluminum $\nu^2 = \epsilon/\epsilon_0 = \mathrm{Re}(\nu^2) + i\,\mathrm{Im}(\nu^2)$ versus wavelength $\lambda(\mu)$.

lengths in the range $0.3\mu < \lambda < 4.0\mu$ for $a = 0.25$ mm, 0.50 mm and 1 mm. These data show a considerable improvement over that corresponding to the lowest-loss mode EH_{11} for the dielectric guide. We saw that for a hollow glass dielectric waveguide, the EH_{11} mode has a loss of 1.8 db/km for a radius $a = 1$ mm and wavelength $\lambda = 1\mu$. The attenuation for the TE_{01} mode for the aluminum guide with the same radius and wavelength is only 0.028 db/km. For a wavelength $\lambda = 1\mu$ and a radius $a = 0.25$ mm, the minimum-loss TE_{01} mode for the aluminum waveguide has an attenuation constant $\alpha_{01} = 1.8$ db/km. The same attenuation is achieved for $\lambda = 3\mu$ and $a = 0.6$ mm. The attenuation constant for the TE_{02} mode under the last two conditions is $\alpha_{02} = 6.05$ db/km. For a wavelength $\lambda = 1\mu$ and $a = 0.25$ mm, the straight guide losses for the TM_{01} and EH_{11} modes are approximately 145 db/km and 57 db/km, respectively.

VII. FIELD CONFIGURATION AND ATTENUATION OF MODES IN THE CURVED GUIDE

In order to achieve a more realistic evaluation of the hollow circular waveguide for long distance optical transmission, it is necessary to

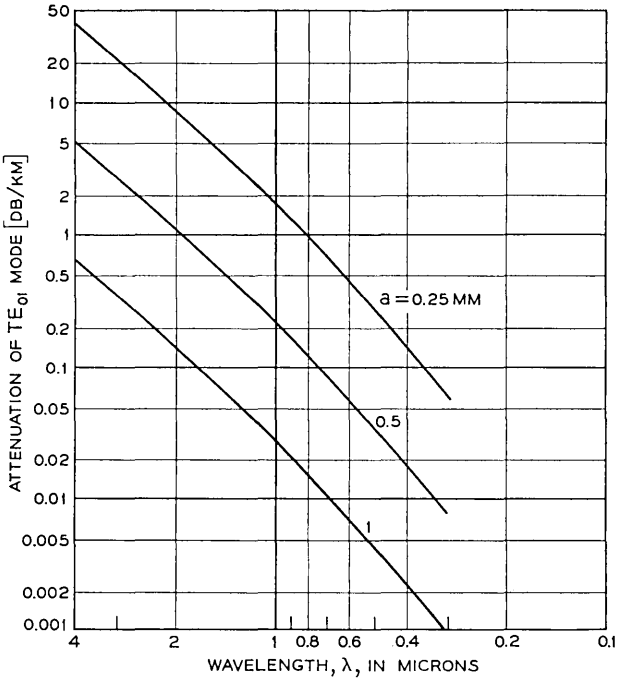


Fig. 9 — Attenuation of TE_{01} mode, α_{01} versus wavelength $\lambda(\mu)$, for aluminum guide.

evaluate the effects of mild curvature of the guide axis. This is most easily accomplished by determining a perturbation correction for both the field configuration and the attenuation constants for the idealized straight guide whose characteristics have been described above.

VIII. FORMULATION OF THE PROBLEM

Consider the toroidal system (r, θ, z) with metric coefficients

$$\begin{aligned} e_r &= 1 \\ e_\theta &= r \\ e_z &= 1 + r/R \sin \theta \end{aligned} \tag{29}$$

as depicted in Fig. 10. In this system of coordinates, a differential length is given by

$$ds = (e_r^2 dr^2 + e_\theta^2 d\theta^2 + e_z^2 dz^2)^{\frac{1}{2}} \tag{30}$$

where R is the radius of curvature of the toroidal system and is chosen

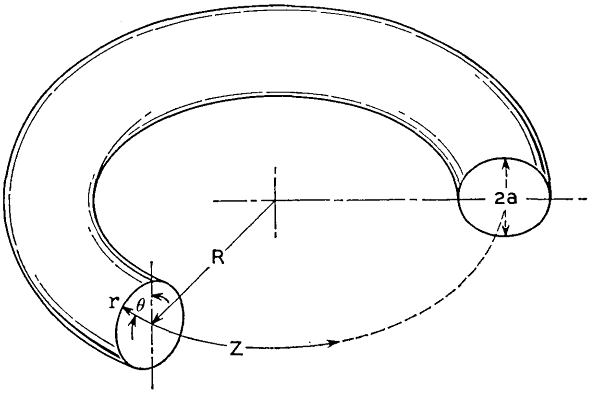


Fig. 10 — The curved hollow dielectric waveguide and the associated toroidal coordinate system (r, θ, Z).

equal to the radius of curvature of the guide axis, so that the guide wall is located at $r = a$, and the axis of the guide coincides with the curved z-axis. In this toroidal coordinate system, Maxwell's equations are

$$\begin{aligned}
\frac{\partial}{\partial\theta}\{(1 + r/R\sin\theta)\mathcal{H}_z\} - i\gamma_c r\mathcal{H}_\theta + i\omega\epsilon r(1 + r/R\sin\theta)\mathcal{E}_e &= 0\\
i\gamma_c\mathcal{H}_r - \frac{\partial}{\partial r}\{(1 + r/R\sin\theta)\mathcal{H}_z\} + i\omega\epsilon(1 + r/R\sin\theta)\mathcal{E}_\theta &= 0\\
\frac{\partial}{\partial r}(r\mathcal{H}_\theta) - \frac{\partial}{\partial\theta}\mathcal{H}_r + i\omega\epsilon r\mathcal{E}_z &= 0\\
\frac{\partial}{\partial\theta}\{(1 + r/R\sin\theta)\mathcal{E}_z\} - i\gamma_c\mathcal{E}_\theta - i\omega\mu r(1 + r/R\sin\theta)\mathcal{H}_r &= 0\\
i\gamma_c\mathcal{E}_r - \frac{\partial}{\partial r}\{(1 + r/R\sin\theta)\mathcal{E}_z\} - i\omega\mu(1 + r/R\sin\theta)\mathcal{H}_\theta &= 0\\
\frac{\partial}{\partial r}(r\mathcal{E}_\theta) - \frac{\partial}{\partial\theta}\mathcal{E}_r - i\omega\mu r\mathcal{H}_z &= 0
\end{aligned} \tag{31}$$

where we have omitted the common factor

$$\exp i(\gamma_c z - \omega t)$$

in which γ_c is the propagation constant along the curved z-axis.

The toroidal system (r, θ, z) and the curved waveguide degenerate into a cylindrical system and a straight guide, respectively, as R approaches infinity. Maxwell's equations for the straight guide are therefore obtained from (31) by letting $R \to \infty$.

$$\begin{aligned}
\frac{\partial}{\partial\theta} H_z - i\gamma r H_\theta + i\omega\epsilon r E_r &= 0 \\
i\gamma H_r - \frac{\partial}{\partial r} H_z + i\omega\epsilon E_r &= 0 \\
\frac{\partial}{\partial r}(rH_\theta) - \frac{\partial}{\partial\theta} H_r + i\omega\epsilon r E_r &= 0 \\
\frac{\partial}{\partial\theta} E_z - i\gamma r E_\theta - i\omega\epsilon r H_r &= 0 \\
i\gamma E_r - \frac{\partial}{\partial r} E_z - i\omega\mu H_\theta &= 0 \\
\frac{\partial}{\partial r}(rE_\theta) - \frac{\partial}{\partial\theta} E_r - i\omega\mu r H_z &= 0
\end{aligned} \tag{32}$$

where γ is the propagation constant for the straight guide, and the superscript i and subscripts nm are suppressed.

IX. SOLUTION FOR THE CURVED GUIDE

We proceed to solve (31) for the field vectors $\vec{\mathcal{E}}$, $\vec{\mathcal{H}}$ and obtain the propagation constant γ_c for the curved guide as functions of the field vectors $\vec{E}$, $\vec{H}$ and the propagation constant γ of the straight guide. The latter quantities are known [(2), (3), (4) and (20)]. We introduce a parameter

$$\sigma = \frac{k}{\gamma - k} \cdot \frac{a}{R} \approx 2 \left(\frac{2\pi a}{u_{nm}\lambda} \right)^2 \frac{a}{R} \cdot \tag{33}$$

The range of interest is that for which the radius of curvature R is so large that $\sigma \ll 1$.

Using a first-order perturbation technique, the solution of (31) is

$$\begin{aligned}
\mathcal{E}_\theta &= (1 + \sigma r/a \sin\theta) E_\theta \\
\mathcal{E}_r &= (1 + \sigma r/a \sin\theta) E_r \\
\mathcal{E}_z &= (1 + \sigma r/a \sin\theta) E_z + (i\sigma/ka)(E_r \sin\theta + E_\theta \cos\theta) \\
\mathcal{H}_\theta &= (1 + \sigma r/a \sin\theta) H_\theta \\
\mathcal{H}_r &= (1 + \sigma r/a \sin\theta) H_r \\
\mathcal{H}_z &= (1 + \sigma r/a \sin\theta) H_z + (i\sigma/ka)(H_r \sin\theta + H_\theta \cos\theta).
\end{aligned} \tag{34}$$

The effect of curvature of the guide axis is to make unsymmetrical the transverse field configuration of the straight guide. Each transverse component is enhanced in the half cross section farthest from the center of curvature.

To a first-order perturbation of σ, the propagation constants of the curved and straight guide are identical; i.e., $\gamma_c \approx \gamma$. Nevertheless, knowing the field components of the mildly curved structure, it is possible to calculate its attenuation constants $\alpha_{nm}(R) = \operatorname{Re} \gamma_c$.

X. ATTENUATION CONSTANTS $\alpha_{nm}(R)$

The mean radial power flowing into the dielectric per unit length at the surface of the guide is

$$P_r = \frac{1}{2} \int_0^{2\pi} \operatorname{Re} [\mathcal{E}_\theta \mathcal{H}_z{}^* - \mathcal{E}_z \mathcal{H}_\theta{}^*] \underset{r=a}{} [1 + a/R \sin \theta] a \, d\theta. \qquad (35)$$

The power flow in the axial z direction within the internal medium $r < a$ is

$$P_z = \frac{1}{2} \int_0^a \int_0^{2\pi} \operatorname{Re} [\mathcal{E}_r \mathcal{H}_\theta{}^* - \mathcal{E}_\theta \mathcal{H}_r{}^*] r \, d\theta \, dr \qquad (36)$$

and decreases along z at a rate equal to the radial flow per unit length P_r ; i.e.,

$$\frac{dP_z}{dz} = -2\alpha_{nm}(R) P_z = -P_r \qquad (37)$$

where $\alpha_{nm}(R)$ is the attenuation constant of the mode under consideration for the curved hollow dielectric waveguide. Consequently

$$\alpha_{nm}(R) = \tfrac{1}{2}(P_r/P_z). \qquad (38)$$

To compute P_r we substitute the known field quantities into (35). This yields

$$P_r = \operatorname{Re} \sqrt{\frac{\epsilon_0}{\mu_0}} \frac{u_{nm}{}^2 J_n{}^2(u_{nm})}{2k^2 a \sqrt{\nu^2 - 1}} \int_0^{2\pi} | 1 + \sigma \sin \theta |^2 (1 + a/R \sin \theta)$$

$$\left\{ \begin{matrix} 1 \\ \nu^2 \\ \nu^2 \sin^2 n(\theta + \theta_0) + \cos^2 n(\theta + \theta_0) \end{matrix} \right\} d\theta \quad \begin{matrix} \text{for } \mathrm{TE}_{0m} \text{ modes} \\ \text{for } \mathrm{TM}_{0m} \text{ modes} \\ \text{for } \mathrm{EH}_{nm} \text{ modes.} \end{matrix} \qquad (39)$$

Terms with powers of $\lambda/(2\pi a)$ larger than two have been neglected. Upon integrating,

$$P_r = \pi \operatorname{Re} \sqrt{\frac{\epsilon_0}{\mu_0}} \frac{u_{nm}{}^2 J_n{}^2(u_{nm})}{k^2 a \sqrt{\nu^2 - 1}}$$

$$\left\{\begin{array}{c} (1 + \tfrac{1}{2}\sigma^2) \\ \nu^2(1 + \tfrac{1}{2}\sigma^2) \\ \dfrac{1}{2}(\nu^2 + 1)\left[1 + \dfrac{1}{2}\sigma^2\left(1 + \dfrac{\delta_n(\pm 1)}{2}\,\dfrac{\nu^2 - 1}{\nu^2 + 1} \right.\right. \\ \qquad\qquad\qquad \left.\left. \cdot \cos 2\theta_0\right)\right] \end{array}\right\} \begin{array}{l} \text{for } \mathrm{TE}_{0m} \text{ modes} \\ \text{for } \mathrm{TM}_{0m} \text{ modes} \\ \\ \text{for } \mathrm{EH}_{nm} \text{ modes} \end{array} \tag{40}$$

where

$$\delta_n(\pm 1) = \begin{cases} 1, & n = \pm 1 \\ 0, & n \neq \pm 1. \end{cases} \tag{41}$$

The power P_z flowing radially in the guide is obtained by substituting (34) into (36) and integrating

$$P_z = \frac{\pi a^2}{2}\sqrt{\frac{\epsilon_0}{\mu_0}}\, J_n^{\,2}(u_{nm})\left\{1 + \frac{\sigma^2}{6}\,[1 + 2n(n - 2)/u_{nm}^{\,2}]\right\}. \tag{42}$$

Hence

$$\begin{aligned} \alpha_{nm}(R) = \alpha_{nm}(\infty)&\left\{1 + \frac{4}{3}\left(\frac{2\pi a}{u_{nm}\lambda}\right)^4\left(\frac{a}{R}\right)^2\right. \\ &\left.\cdot\left[1 - \frac{n(n-2)}{u_{nm}^{\,2}} + \frac{3}{4}\,\delta_n(\pm 1)\,\frac{\operatorname{Re}\sqrt{\nu^2 - 1}}{\operatorname{Re}\dfrac{\nu^2 + 1}{\sqrt{\nu^2 - 1}}}\cos 2\theta_0\right]\right\} \end{aligned} \tag{43}$$

where $\alpha_{nm}(\infty) = \alpha_{nm}$ is the attenuation constant for modes in the straight guide ($R = \infty$) given by (21). The attenuation constant $\alpha_{nm}(R)$ can also be written in the following form

$$\alpha_{nm}(R) = \alpha_{nm}(\infty) + (a^3/\lambda^2R^2)\operatorname{Re}V_{nm}(\nu) \tag{44}$$

where

$$\begin{aligned} V_{nm}(\nu) = \frac{4}{3}&\left\{\begin{array}{c} \dfrac{1}{\sqrt{\nu^2 - 1}} \\ \dfrac{\nu^2}{\sqrt{\nu^2 - 1}} \\ \dfrac{\tfrac{1}{2}(\nu^2 + 1)}{\sqrt{\nu^2 - 1}} \end{array}\right\}\left(\frac{2\pi}{u_{mn}}\right)^2 \\ &\cdot\left\{1 - \frac{n(n-2)}{u_{nm}^{\,2}} + \frac{3}{4}\,\delta_n(\pm 1)\left(\frac{\nu^2 - 1}{\nu^2 + 1}\right)\cos 2\theta_0\right\}. \end{aligned} \tag{45}$$

The values of $\mathrm{Re}V_{nm}(\nu)$ are always positive. Some of them have been calculated in Table II for a refractive index $\nu = 1.50$.

The attenuation constant of any mode consists of two terms (44). The first coincides with that of the straight guide and is proportional to $u_{nm}^2\lambda^2/a^3$; the second term represents an increase in attenuation due to curvature of the guide axis and is proportional to $a^3/\lambda^2R^2u_{nm}^2$. Therefore the lower the straight guide attenuation constant (small $u_{nm}^2\lambda^2/a^3$), the larger the loss due to bends and vice versa. From (43) or (45) we find that only for the $\mathrm{EH}_{\pm1,m}$ modes, the orientation of the field with respect to the plane of curvature influences the attenuation. If $\theta_0 = 0$, the electric field in the center of the guide is in the plane of curvature and the attenuation is a maximum. For $\theta_0 = \pm\pi/2$, the electric field is normal to the plane of curvature and the attenuation is a minimum. The ratio of maximum to minimum is mild, however. For the lowest attenuation mode EH_{11} and $\nu = 1.50$, it is

$$\frac{V_{nm}(\theta_0 = 0)}{V_{nm}(\theta_0 = \pi/2)} = 1.65. \tag{46}$$

If $|\nu| \gg 1$, that ratio is

$$\frac{V_{nm}(\theta_0 = 0)}{V_{nm}(\theta = \pi/2)} = 4.6.$$

From equation (43) we find that the radius of curvature which doubles the straight guide attenuation is

$$R_0 = \frac{2}{\sqrt{3}}\left(\frac{2\pi}{u_{mn}}\right)^2\frac{a^3}{\lambda^2} \cdot\left[1 - \frac{n(n-2)}{u_{nm}^2} + \frac{3}{4}\,\delta_n(\pm1)\;\frac{\mathrm{Re}\,\dfrac{\sqrt{\nu^2-1}}{\nu^2+1}}{\mathrm{Re}\,\dfrac{\nu^2+1}{\sqrt{\nu^2-1}}}\;\cos 2\vartheta_0\right]^{\frac{1}{2}}. \tag{47}$$

This value of R_0 is only approximate since (43) was derived by assuming $\sigma \ll 1$.

XI. EFFECT OF CURVATURE ON ATTENUATION OF MODES IN THE HOLLOW DIELECTRIC WAVEGUIDE

For a straight hollow glass waveguide with $\nu = 1.5$ and a radius $a = 1$ mm operating typically at a wavelength $\lambda = 1\mu$, the attenuation of the lowest-loss mode EH_{11} is $\alpha_{11} = 1.85$ db/km. This loss is doubled for a radius of curvature $R_0 \approx 10$ km. For long distance optical transmission a radius of curvature of at least a few hundred meters would

TABLE II — SOME VALUES OF $V_{nm}(\nu)$

n/m		1	2	3	4
-1		2.57 (1 + 0.326 cos $2\theta_0$)	1.034 (1 + 0.301 cos $2\theta_0$)	0.553 (1 + 0.295 cos $2\theta_0$)	0.347 (1 + 0.293 cos $2\theta_0$)
0	TE	3.22	0.955	0.455	0.265
	TM	7.22	2.145	1.022	0.596
1		15.5 (1 + 0.246 cos $2\theta_0$)	2.60 (1 + 0.279 cos $2\theta_0$)	1.034 (1 + 0.284 cos $2\theta_0$)	0.554 (1 + 0.286 cos $2\theta_0$)
2		5.22	1.55	0.735	0.432
3		2.57	1.034	0.553	0.347
4 or -2		1.51	0.737	0.430	0.287

be tolerable. Therefore hollow dielectric waveguides do not seem suitable for long distance optical transmission.

On the other hand, the curvature in hollow dielectric waveguides for application in gaseous amplifiers and oscillators is not critical. For example, if $a = 0.25$ mm and $\lambda = 1\mu$, the straight guide attenuation is 0.12 db/meter. The radius of curvature which doubles this quantity for the lossiest polarization — i.e., with the electric field at the center of the guide contained in the plane of curvature — is approximately 150 meters, a value well within the limits of laboratory precision. Consequently, the hollow dielectric waveguide does remain very attractive as a guiding medium for optical amplifiers and oscillators where a small guide radius is desirable, thereby making the guide less sensitive to curvature of the axis.

XII. EFFECT OF CURVATURE ON ATTENUATION OF MODES IN THE METALLIC GUIDE

The attenuation constants $\alpha_{0m}(R)$ for the lowest-loss TE_{0m} modes in the curved metallic guide are given by

$$\alpha_{0m}(R) \approx \alpha_{0m}(\infty)\left\{1 + \frac{4}{3}\left(\frac{2\pi a}{\lambda u_{0m}}\right)^4 \left(\frac{a}{R}\right)^2\right\} \tag{48}$$

where $\alpha_{0m}(\infty)$ is the attenuation constant for the TE_{0m} mode in the straight guide, $R = \infty$. For a radius $a = 0.25$ mm and wavelength $\lambda = 1\mu$, the straight guide loss for the lowest-loss TE_{01} mode, $\alpha_{01}(\infty) =$ 1.8 db/km, is doubled for a radius of curvature of only $R_0 \approx 48$ meters.

For $\lambda = 3\mu$ and $a = 0.6$ mm, the straight TE_{01} loss is also 1.8 db/km and the radius of curvature that doubles that loss is 75 *m.*

XIII. CONCLUSIONS

The hollow dielectric waveguide at optical wavelengths supports a complete set of normal modes that are either circular electric, circular magnetic or hybrid. They resemble the modes found in a sequence of circular irises not only in field configuration but also in loss discrimination among them. For hollow metallic waveguides the mode discrimination is far larger.

The field configuration and propagation constants have been determined. The attenuation is practically independent of the loss tangent of the dielectric but depends essentially on the refraction mechanism at the wall. Assuming refractive index of the dielectric, 1.5 for hollow dielectric waveguides, the EH_{11} mode exhibits the lowest power attenuation, viz., 1.85 (λ^2/a^3) db/m. For a wavelength $\lambda = 1\mu$ and a tube radius $a = 1$ mm, the attenuation is only 1.85 db/km.

The hollow dielectric waveguide does not, however, seem suitable for long distance optical transmission because of the high loss introduced by even mild curvature of the guide axis. Nevertheless it remains very attractive as a guiding medium for optical amplifiers and oscillators, since here a small radius of the guide is desirable. Consequently, curvature of the guide axis is not critical. Filled with "masing" material, the hollow dielectric waveguide provides not only guidance but also gain which is almost inversely proportional to the radius. For the right He-Ne mixture, the maximum theoretical gain attainable is 7.6 db/m provided that the radius is 0.058 mm. But even if the radius is 0.25 mm, the predicted gain is still large, viz. 2.6 db/m.

The metallic waveguide is superior to the hollow dielectric waveguide for use in long distance optical transmission. Because of the relatively large dielectric constant exhibited by aluminum at optical frequencies, the attenuation constant for the lowest-loss mode TE_{01} is comparatively small and less sensitive to curvature of the guide axis. For a radius $a = 0.25$ mm and a wavelength $\lambda = 1\mu$, the attenuation constant for TE_{01} modes in the straight aluminum guide is only 1.8 db/km, which is doubled for a radius of curvature of about 48 meters. For $a = 0.6$ mm and $\lambda = 3\mu$, the TE_{01} straight guide loss is also 1.8 db/km but is doubled if the radius of curvature of the waveguide axis is 75m.

We have considered some of the theoretical problems of the hollow dielectric or metallic waveguide. The results are promising. Nevertheless, the usefulness of these guides has yet to be proven experimentally, and furthermore the attenuation constants discussed here do not include scattering losses due to surface imperfections.

XIV. ACKNOWLEDGMENTS

It is a pleasure to thank R. Kompfner and S. E. Miller for their suggestions.

REFERENCES

1. Fox, A. G., and Li, Tingye, Resonant Modes in a Maser Interferometer, B.S.T.J., **40**, March, 1961, p. 453.
2. Boyd, G. D., and Gordon, J. P., Confocal Multimode Resonator for Millimeter through Optical Wavelength Masers, B.S.T.J., **40**, March, 1961, p. 489.
3. Boyd, G. D., and Kogelnik, H., Generalized Confocal Resonator Theory, B.S.T.J., **41**, July, 1962, p. 1347.
4. Goubau, G., and Schwering, F., On the Guided Propagation of Electromagnetic Wave Beams, Trans. I.R.E., **AP-9**, May 1961, p. 248.
5. Eaglesfield, C. C., Optical Pipeline: A Tentative Assessment, The Inst. of Elect. Engineers, January, 1962, p. 26.
6. Simon, J. C., and Spitz, E., Propagation Guidée de Lumière Cohérente, J. Phys. Radium, **24**, February, 1963, p. 147.
7. Goubau, G., and Christian, J. R., Some Aspects of Beam Waveguides for Long Distance Transmission at Optical Frequencies, IEEE Trans. on Microwave Theory and Techniques, **MTT-12**, March, 1964, pp. B.S.T.J., 212–220.
8. Marcuse, D., and Miller, S. E., Analysis of a Tubular Gas Lens, B.S.T.J., this issue, p. 1759.
9. Berreman, D. W., A Lens or Light Guide Using Convectively Distorted Thermal Gradients in Gases, B.S.T.J., this issue, p. 1469.
10. Berreman, D. W., A Gas Lens using Unlike, Counter-Flowing Gases, B.S.T.J., this issue, p. 1476.
11. Gordon, E. I., and White, A. D., Similarity Laws for the He-Ne Gas Maser, Appl. Phys. Letters, **3**, December 1, 1963, p. 199.
12. Stratton, J. A., *Electromagnetic Theory*, McGraw-Hill Book Co., New York and London, 1941, p. 524.
13. Hodgson, J. N., Proc. Phys. Soc. (London), **B68**, 1955, p. 593.
14. Beattie, J. R., and Conn, G. K. T., Phil. Mag., **7**, 1955, pp. 46, 222, and 989.
15. Schulz, L. G., J. Opt. Soc. Am., **41**, 1951, p. 1047; **44**, 1954 p. 357.
16. Givens, M. Parker, Optical Properties of Metals, *Solid State Physics*, **6**, Acad. Press Inc., New York, 1958, p. 313.

Contributors to This Issue

VACLAV E. BENEŠ, A.B., 1950, Harvard College; M.A. and Ph.D., 1953, Princeton University; Bell Telephone Laboratories, 1953—. Mr. Beneš has been engaged in mathematical research on stochastic processes, traffic theory, and servomechanisms. In 1959–60 he was visiting lecturer in mathematics at Dartmouth College. He is the author of *General Stochastic Processes in the Theory of Queues* (Addison-Wesley, 1963). Member, American Mathematical Society, Association for Symbolic Logic, Institute of Mathematical Statistics, Society for Industrial and Applied Mathematics, Mind Association and Phi Beta Kappa.

MARTIN R. BIAZZO, Bell Telephone Laboratories, 1951—. He attended Rutgers University College. His first assignment at Bell Laboratories was on PCM systems, primarily in the area of ferrite core transformers and PCM repeaters. His present interest is in the measurement of optical losses with the application of electronic techniques on optical lasers.

JAMES O. CAPPELLARI, JR., B.S., 1953, Marshall University; B.S.A.E., 1956, M.S.A.E., 1957, Ph.D., 1961, Purdue University; Instructor, School of Aeronautical and Engineering Sciences, Purdue, 1957–1961; Bell Telephone Laboratories, 1961–1962, Bellcomm, Inc., 1962—. While at Bell Laboratories he was engaged in attitude control, rigid body dynamics, and orbital mechanics problems connected with Project Telstar. Since going to Bellcomm, Inc., he has been concerned with a wide variety of trajectory analysis studies connected with Project Apollo. Member, A.I.A.A. and Sigma Xi.

G. F. FOXHALL, B.S. (Physics), Worcester Polytechnic Institute, 1961, M.S. (Physics), University of Illinois, 1962; Bell Telephone Laboratories, 1961—. He has investigated the effects of gaseous ambients on silicon surfaces, worked on the development of a germanium bridge rectifier polarity guard, and studied metal-semiconductor contacts. Member, Sigma Xi.

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, he 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 simulations. Member, Acoustical Society of America, IEEE, Sigma Xi and Tau Beta Pi.

John P. Hyde, A.B., Princeton University, 1959; M.S., Northwestern University, 1960; Bell Telephone Laboratories, 1960—. He has worked on machine aids to design, and especially computer aids to sequential circuit synthesis. He is presently engaged in further development of the ALPAK system and other aspects of computer algebra.

James F. Kaiser, E.E., University of Cincinnati, 1952; S.M., 1954, and Sc.D., 1959, Massachusetts Institute of Technology; faculty of the Massachusetts Institute of Technology, 1956–1960; Bell Telephone Laboratories, 1959—. He has been concerned with problems of data processing, digital filter design, and system simulation. Member, IEEE, Association for Computing Machinery, Society for Industrial and Applied Mathematics, Eta Kappa Nu, Sigma Xi and Tau Beta Pi.

B. K. Kinariwala, B.S., 1951, Benares University (India); M.S., 1954, and Ph.D., 1957, University of California; Bell Telephone Laboratories, 1957—. He was first engaged in research in circuit theory involving, in particular, active and time-varying networks. More recently, he has been concerned with problems in digital communication systems. Member, IEEE and Sigma Xi.

David A. Kleinman, S.B., 1946, and S.M., 1947, Massachusetts Institute of Technology; Ph.D., 1952, Brown University; Brookhaven National Laboratory, 1949–53; Bell Telephone Laboratories, 1953—. Mr. Kleinman has worked in the areas of neutron scattering in solids, semiconductor electronics, electron energy bands, and the infrared properties of crystals, and is currently working on problems related to the optical maser. Member, American Physical Society.

John A. Lewis, B.S., 1944, Worcester Polytechnic Institute, Sc. M., 1948, Brown University, Ph.D., 1950, Brown University, Bell Telephone Laboratories, May, 1951—. A member of the mathematical physics department, he has been engaged in theoretical investigation of problems of fluid dynamics, piezoelectric vibrations, heat transfer, and satellite attitude control. He is currently studying problems of hypersonic flow. Member, American Mathematical Society, Society for Industrial and Applied Mathematics, l'Unione Matematica Italiana and the Society for Natural Philosophy.

E. A. J. Marcatili, Aeronautical Engineer, 1947, and E.E., 1948, University of Córdoba (Argentina); Research staff, University of Córdoba, 1947–54; Bell Telephone Laboratories, 1954—. He has been engaged in theory and design of filters in multimode waveguides and in waveguide systems research. More recently he has concentrated on the study of optical transmission media. Member, IEEE.

Dietrich Marcuse, Diplom Vorpruefung, 1952 and Dipl. Phys., 1954; Berlin Free University; D.E.E., 1962, Technische Hochschule, Karlsruhe, Geımany; Siemens and Halske (Germany), 1954–1957; Bell Telephone Laboratories, 1957—. At Siemens and Halske Mr. Marcuse was engaged in transmission research, studying coaxial cable and circular waveguide transmission. At Bell Laboratories he has been engaged in studies of circular electric waveguides and work on gaseous masers. Member, IEEE.

Stewart E. Miller, B.S. and M.S., 1941, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1941—. He first worked on coaxial carrier repeaters and later shifted to microwave radar systems development. At the close of World War II he returned to coaxial carrier repeater development until 1949, when he joined the radio research department. There his work has been in the fields of circular electric waveguide communication, microwave ferrite devices, and other components for microwave radio systems. As Director, Guided Wave Research Laboratory, he heads a group engaged in research on communication techniques for the millimeter wave and optical regions. Fellow, IEEE.

A. J. Rack, B.S., 1930, University of Illinois; M.A., 1935, Columbia University; Bell Telephone Laboratories, 1930—. He has been engaged in the application of circuits in the communication field, including studies of tube noise, feedback amplifiers, transistor circuits and PCM. At present he is investigating the field of optical loss measurement.

Irwin W. Sandberg, B.E.E., 1955, M.E.E., 1956, and D.E.E., 1958, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1958—.

He has been concerned with analysis of military systems, particularly radar systems, and with synthesis and analysis of active and time-varying networks. He is currently involved in a study of the signal-theoretic properties of nonlinear systems. Member, IEEE, Society for Industrial and Applied Mathematics, Eta Kappa Nu, Sigma Xi and Tau Beta Pi.

Robert A. Schmeltzer, B.S., 1955, The Cooper Union; M.S., 1958, The Stevens Institute of Technology; Eng.Sc.D. 1962, New York University; Bell Telephone Laboratories, 1963—. He has been engaged in theoretical research in low-loss guided wave systems for long distance optical transmission and lasers. Member, American Mathematical Society, IEEE and Eta Kappa Nu.

Lewis C. Thomas, B.E.E., Cornell University, 1949, M.S. in E.E., c.s.l., 1958, Newark College of Engineering; Bell Telephone Laboratories, 1949—. He has worked on the Nike missile systems, pulse code modulation systems, and data transmission systems. His recent work has included attitude and orbital mechanics studies for Project Telstar and communication satellite system studies. He is on the lecture staff of the American Museum–Hayden Planetarium. Member, IEEE, A.I.A.A., Royal Astronomical Society of Canada and Eta Kappa Nu; honorary member, Epsilon Pi Tau.

B. S. T. J. BRIEFS

A Condition for the $\mathcal{L}_\infty$-Stability of Feedback Systems Containing a Single Time-Varying Nonlinear Element

By I. W. SANDBERG

(Manuscript received May 14, 1964)

In the automatic control literature, a feedback system is frequently said to be stable if, regardless of the initial state of the system, each bounded input applied at $t = 0$ produces a bounded output. The purpose of this brief is to present a sufficient condition for the feedback system of Fig. 1 to be stable in this sense.

Our discussion is restricted to cases in which g_1, f, u, and v (in Fig. 1) denote real-valued measurable functions of t defined for $t \geqq 0$. The block labeled ψ is assumed to represent a memoryless time-varying element that introduces the constraint $u(t) = \psi[f(t),t]$, in which $\psi(x,t)$ is a function of x and t with the properties that $\psi(0,t) = 0$ for $t \geqq 0$ and there exist a positive constant β and a real constant α such that

$$\alpha \leqq \frac{\psi(x,t)}{x} \leqq \beta, \qquad t \geqq 0$$

for all real $x \neq 0$.

The block labeled $\mathbf{K}$ represents the linear time-invariant portion of the forward path, and is assumed to introduce the constraint

$$v(t) = \int_0^t k(t - \tau)u(\tau)d\tau - g_2(t), \qquad t \geqq 0 \tag{1}$$

in which k and g_2 are real-valued measurable functions such that

$$\int_0^\infty |k(t)|\,dt < \infty, \qquad \sup_{t \geqq 0} |g_2(t)| < \infty.$$

The function g_2 takes into account the initial conditions at $t = 0$. We do not require that u and v be related by a differential equation (or by a system of differential equations).

Assumption: We shall assume that the response $v(t)$ is well defined and such that for all finite $y > 0$

$$\sup_{0 \leqq t \leqq y} |v(t)| < \infty,$$

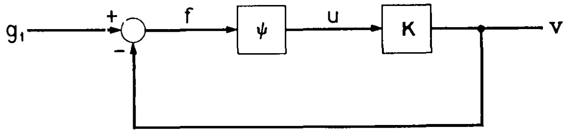


Fig. 1 — Nonlinear feedback system.

for each initial-condition function g_2 that meets the conditions stated above and each input g_1 such that

$$\sup_{t \geqq 0} | g_1(t) | < \infty .$$

Definition: We shall say that the feedback system of Fig. 1 is "$\mathcal{L}_\infty$-stable" if and only if there exists a positive constant ρ with the property that the response v satisfies

$$\sup_{t \geqq 0} | v(t) | \leqq \rho \sup_{t \geqq 0} | g_1(t) + g_2(t) | + \sup_{t \geqq 0} | g_2(t) |$$

for every initial-condition function g_2 that meets the conditions stated above and every input g_1 such that

$$\sup_{t \geqq 0} | g_1(t) | < \infty .$$

Clearly, if the system is $\mathcal{L}_\infty$-stable, then the response is bounded whenever the input is bounded.

Theorem: Suppose that

(*i*) *with* $K(s) = \int_0^\infty k(t) e^{-st} dt$ *for* $Re[s] \geqq 0$,

$1 + \frac{1}{2}(\alpha + \beta)K(s) \neq 0$ *for* $Re[s] \geqq 0$, *and*

(*ii*) *with* $h(\cdot)$ *the inverse Laplace transform of*

$$\frac{K(s)}{1 + \frac{1}{2}(\alpha + \beta)K(s)},^*$$

$$\tfrac{1}{2}(\beta - \alpha) \int_0^\infty | h(t) | \, dt < 1.$$

Then the feedback system of Fig. 1 is $\mathcal{L}_\infty$*-stable.*

It is of interest to note that condition (*ii*) is satisfied whenever condition (*i*) is met, $\alpha > 0$, $h(t) \geqq 0$ for $t \geqq 0$, and

$$\int_0^\infty k(t) \, dt > 0,$$

* Condition (*i*) and our assumption regarding $k(\cdot)$ imply that $h(\cdot)$ exists, and that its modulus is integrable on $[0,\infty)$ [see Ref. 1].

for then

$$\tfrac{1}{2}(\beta - \alpha) \int_0^\infty |h(t)|\, dt = \frac{\tfrac{1}{2}(\beta - \alpha)K(0)}{1 + \tfrac{1}{2}(\alpha + \beta)K(0)} < 1.$$

The theorem can be proved with the techniques discussed in Refs. 2 and 3. More specifically, consider the relation between f and $(g_1 + g_2)$:

$$g_1(t) + g_2(t) = f(t) + \int_0^t k(t - \tau)\psi[f(\tau),\tau]d\tau, \qquad t \geqq 0 \qquad (2)$$

and suppose that

$$\sup_{t \geqq 0} | g_1(t) + g_2(t) | < \infty.$$

Arguments very similar to those used to prove Theorem 1 of Ref. 2 and Theorem I of Ref. 3 show that if (a) f satisfies (2), (b)

$$\sup_{0 \leqq t \leqq y} | f(t) | < \infty$$

for all finite $y > 0$, and (c) conditions (*i*) and (*ii*) of our theorem are satisfied, then there exists a positive constant ρ_1 (which does not depend upon g_1 or g_2) such that

$$\sup_{t \geqq 0} | f(t) | \leqq \rho_1 \sup_{t \geqq 0} | g_1(t) + g_2(t) | .$$

Our theorem is a direct consequence of this fact, in view of (1) and the relation between u and f.

REFERENCES

1. Paley, R. E., and Wiener, N., Fourier Transforms in the Complex Domain, Publ. Am. Math. Soc., Providence, R. I., 1934, pp. 60–61.
2. Sandberg, I. W., On the $\mathfrak{L}_2$-Boundedness of Solutions of Nonlinear Functional Equations, B.S.T.J., this issue, p. 1581.
3. Sandberg, I. W., Signal Distortion in Nonlinear Feedback Systems, B.S.T.J., **42,** Nov., 1963, p. 2533.

Thermal Gas Lens Measurements

By A. C. BECK

(Manuscript received June 12, 1964)

The refractive index of a gas is an inverse function of its temperature under isobaric conditions. Therefore a cool gas flowing into a heated tube will have a lower refractive index near the heated walls than in the center, and the combination becomes a convex lens which will focus a light beam transmitted through the tube.[1,2]

A simple arrangement was built to get some information about the behavior of such a device. It is sketched in Fig. 1. The gas flowed through a 5-inch long electrically heated brass tube with a $\frac{1}{4}$-inch inside diameter. The tube was mounted in a large polyfoam cylinder to reduce external heat losses. A single-mode light beam from a helium-neon gas laser oscillating at $\lambda = 0.63$ micron was collimated at a diameter of about $\frac{1}{8}$ inch with glass lenses and then sent along the axis of the heated tube, through which a gas was flowing. The light was intercepted on a screen about 10 to 20 feet away. When this system, acting like a convex lens, is placed in such a collimated beam, the light goes through a focus at the lens focal length, and then expands to form a much larger area of light on the screen. A Foucault knife edge cutting the beam at the focus was used for measuring the focal length of the lens.

The reciprocal of the lens focal length in meters (often called the focusing power or convergence of the lens, and usually expressed in diopters) is plotted as a function of wall temperature rise on Fig. 2. Measurements were made with air and with carbon dioxide at the indicated flow rates. Helium was found to give very small effects, as expected.[2]

It will be noted from Fig. 2 that the lens power increases more slowly with flow rates at the higher values. At somewhat higher flow rates it flattens off and stops increasing, but at these very high flow rates, gas turbulence sets in and laminar flow is no longer present, so accurate measurements become impossible. At the flow rates shown on Fig. 2, constant and steady light patterns and focal lengths are observed when heated gas emerging from the tube is kept out of the light beam. This indicates that when laminar flow exists, the gas is very stable in a tube as small as this one; therefore steady optical lens effects are obtained. At higher temperatures the lens power continues to increase, at least up to the temperatures obtainable with this equipment. At the highest lens powers there is an effect that appears like spherical aberration,

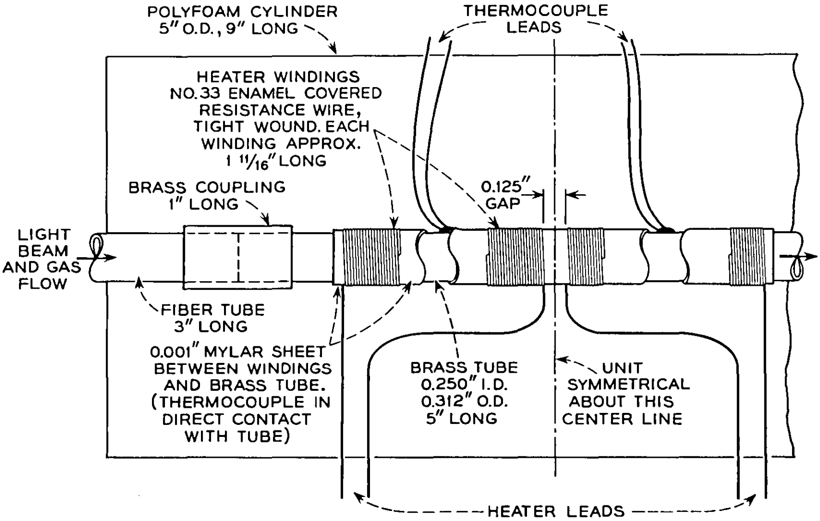

POLYFOAM CYLINDER
5" O.D., 9" LONG
THERMOCOUPLE LEADS
HEATER WINDINGS NO.33 ENAMEL COVERED RESISTANCE WIRE, TIGHT WOUND. EACH WINDING APPROX. 1 11/16" LONG
BRASS COUPLING 1" LONG
0.125" GAP
LIGHT BEAM AND GAS FLOW
FIBER TUBE 3" LONG
0.001" MYLAR SHEET BETWEEN WINDINGS AND BRASS TUBE. (THERMOCOUPLE IN DIRECT CONTACT WITH TUBE)
BRASS TUBE 0.250" I.D. 0.312" O.D. 5" LONG
UNIT SYMMETRICAL ABOUT THIS CENTER LINE
HEATER LEADS

Fig. 1 — Experimental arrangement for thermal gas lens.

shown by a concentric pattern of light rings beginning to appear on the screen,[3,4] but this is very slight in the ranges shown on Fig. 2.

The brass tube was heated with direct current, and the input electrical power was measured for all points. The electrical power required with no gas flow was also measured as a function of temperature. The heat power taken away by the flowing gas was the difference between these two values at the measured temperature rise. From this information, a "figure of merit," defined as the reciprocal of the focal length in meters divided by the watts of heat extracted by the flowing gas (diopters per watt), can be calculated.

Representative data, and some comparisons with theoretical values, are summarized in Table I. The theory, notation, and gas constants are given by Marcuse and Miller.[2]

Agreement between calculated and measured values is fair for heat power consumption of the flowing gas, but not as good for focusing power, particularly at large flow rates of the heavy gas.

The focusing power and figure of merit of these lenses can be improved by using longer sections of tubing, since this one was not optimized. They may also be improved by using other gases having a higher $(n - 1)/k$ ratio, where n is the refractive index and k is the heat conductivity of the gas, and by operating at higher gas pressures.

Thermal gas lenses have short focal lengths with low power consumptions, and the effects are remarkably steady. The focusing power can

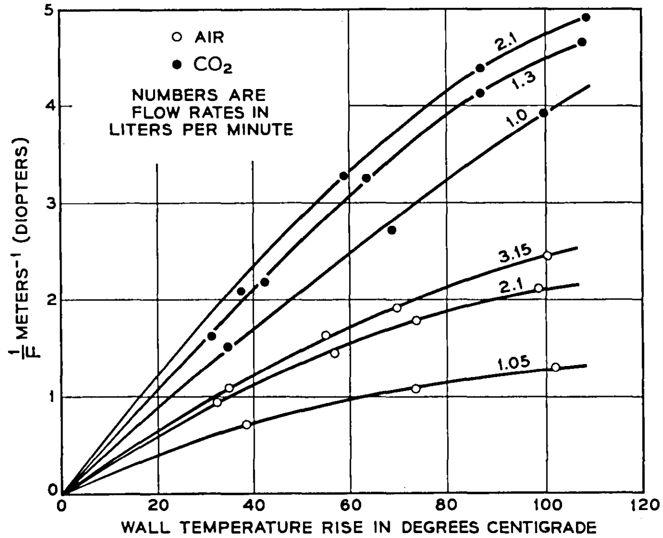


Fig. 2 — Lens focusing power versus wall temperature rise.

be adjusted easily by changing the gas flow rate. Such devices may find applications to long distance optical communication systems.

The assistance of W. E. Whitacre in constructing and setting up the equipment is gratefully acknowledged.

Table I — Representative Data and Theoretical Values

Gas	Flow Rate (liters/min)	V (cm/sec)	v_0 (cm/sec)	v_0/V	1/F (diopters)		Power (watts)		Figure of merit (diopters/watt)	
					calc.	meas.	calc.	meas.	calc.	meas.
Air	1	27.24	105	3.87	1.67	1.3	1.74	1.8	0.96	0.72
Air	2	27.24	211	7.74	1.95	2.1	2.75	2.9	0.71	0.72
CO_2	1	13.52	105	7.80	3.00	3.8	1.73	2.0	1.73	1.90
CO_2	2	13.52	211	15.6	2.02	4.7	2.49	2.9	0.81	1.62

REFERENCES

1. Berreman, D. W., A Lens or Light Guide Using Convectively Distorted Thermal Gradients in Gases, B.S.T.J., **43,** July, 1964, Part 1, p. 1469.
2. Marcuse, D., and Miller, S. E., Analysis of a Tubular Gas Lens, B.S.T.J., this issue, p. 1759.
3. Born, Max, and Wolf, Emil, *Principles of Optics,* Pergamon Press, 1959, pp. 472–477.
4. Beck, A. C., Gas Mixture Lens Measurements, B.S.T.J., this issue, p. 1821.

Gas Mixture Lens Measurements

By A. C. BECK

(Manuscript received June 15, 1964)

Different gases have different refractive indexes. If two different gases are introduced into different parts of a straight hollow tube and allowed to continue in laminar flow, they diffuse into each other slowly but remain sufficiently separated so that a refractive index gradient is maintained for some distance. It has been suggested by A. R. Hutson, G. E. Conklin, and the author that various optical components such as prisms and lenses can be obtained by using such gaseous structures. These components have very low losses and reflections, with no solid surfaces to cause matching or cleaning problems. For these reasons they may have important applications in optical communications systems.[1,2,3]

A simple structure, sketched in Fig. 1, was built to test this principle and to get some information about the performance of a gas mixture lens. A cylindrical porous-walled tube with an inside diameter of $\frac{1}{4}$ inch and an exposed length of 3 inches was mounted inside a larger phenolic cylinder with its ends blocked to form a gas reservoir. A single-mode light beam from a helium-neon gas laser oscillating at $\lambda = 0.63$ micron was collimated at a diameter of about $\frac{1}{8}$ inch with glass lenses and then sent along the axis of the porous tube through which one gas was flowing. The second gas was introduced into the tube through the porous walls. The light beam was then intercepted on a screen about 10 to 20 feet away. If there were no optical refractive effects on this beam, a spot about $\frac{1}{8}$ inch in diameter was seen at these distances, since the beam was collimated. A diverging lens increases the spot size without the beam going through a focus between the lens and the screen. A converging lens causes the beam to become smaller, go through a focus, then expand, forming a much larger spot on the screen at distances of many focal lengths. When measuring converging lenses, a Foucault knife edge cutting the beam was used to determine the focal point and thus the focal length of the lens.

With this system, diverging lenses are obtained if the gas of higher refractive index is introduced through the walls of the porous tube while the gas with the lower index flows down the tube parallel to the light beam. Conversely, sending the less refractive gas through the porous wall of the tube through which the more refractive gas and the light beam are passing gives a converging or positive lens. The reciprocal of the lens focal length in meters (often called the focusing power or con-

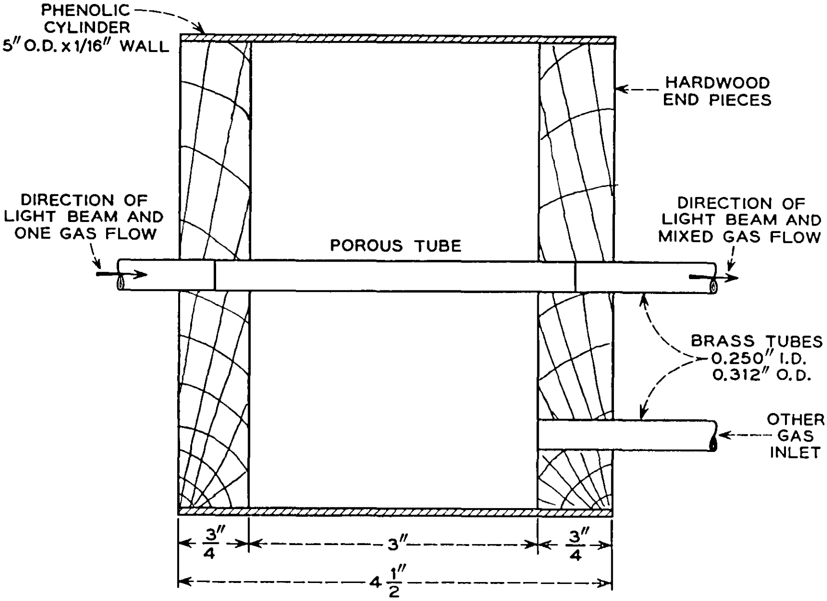

PHENOLIC CYLINDER 5" O.D. x 1/16" WALL
HARDWOOD END PIECES
DIRECTION OF LIGHT BEAM AND ONE GAS FLOW
POROUS TUBE
DIRECTION OF LIGHT BEAM AND MIXED GAS FLOW
BRASS TUBES 0.250" I.D. 0.312" O.D.
OTHER GAS INLET
3/4"
3"
3/4"
4 1/2"

Fig. 1 — Experimental arrangement for gas mixture lens.

vergence of the lens, and usually expressed in diopters) is plotted for various gas flow rates in Figs. 2 and 3. Three different gas mixtures are shown, although other gas mixtures might have even greater focusing power. With carbon dioxide-helium and air-helium, focusing power increased with increased flow rate in the range of Fig. 2, but Fig. 3 shows that this is not the case for the carbon dioxide-air mixture in the range shown there, since focusing power at first increased, reached a maximum, then decreased as the carbon dioxide flow rate was increased. For these curves, Dextilose paper, No. 17V, was rolled up with about five thicknesses to form the porous tube. Other materials work equally well or better, porous ceramics, sintered metals, fritted glass and fine copper screen showing similar results.

At the indicated flow rates, the patterns and focal lengths are constant and steady when the mixed gases emerging from the tube are kept out of the light beam. This shows that the gas mixture is quite stable when laminar flow exists in the tube. The tube continued beyond the porous region exposed to the second gas for about $1\frac{1}{2}$ inches, but greater focusing power than that shown here was obtained with longer exit tube lengths. This indicates that the gases were still in laminar flow and were not completely mixed in this short length, so that focusing continued with

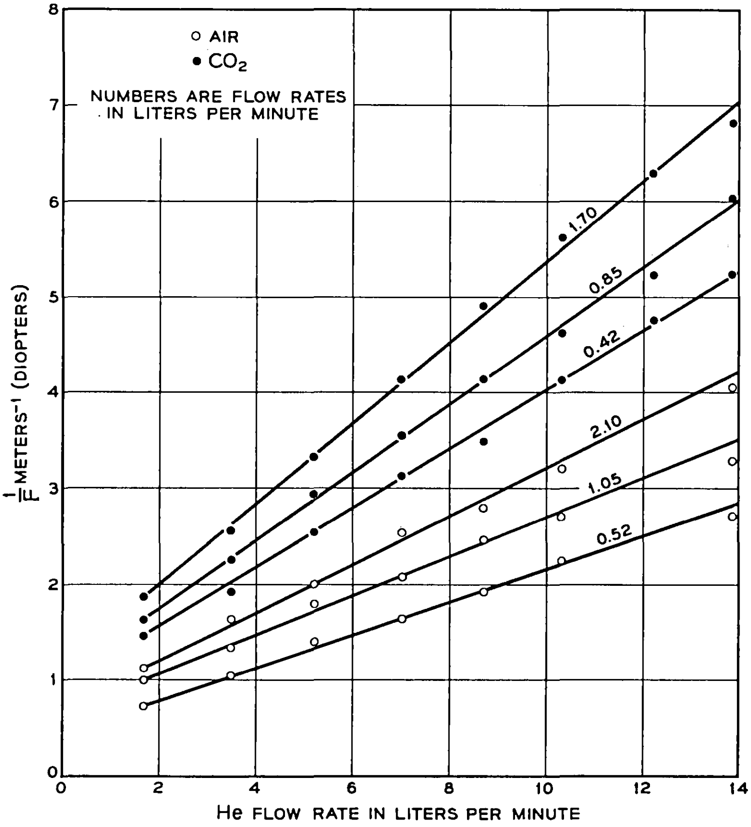


Fig. 2 — Lens focusing power versus gas flow rates, air-helium and carbon dioxide-helium.

more length. At higher flow rates, particularly of the gas first introduced into the system, turbulence appeared. This broke up the light beam and gave unsteady and flickering patterns on the screen.

Some photographs of patterns observed with gas mixture arrangements are reproduced on Fig. 4. At the upper left, the image of the beam is shown with no gases flowing in the tube. The diameter of this spot on a screen at any distance was about $\frac{1}{8}$ inch. At the upper right is the enlarged spot formed by carbon dioxide flowing down the tube and helium introduced through the porous tube walls at rather high flow rates so that a short focal length (approximately 6 inches) was obtained. This pattern, of course, increased in size as the screen was moved farther away. At lower flow rates, giving longer focal lengths, the light pattern is quite uniform, but as the focal length is made short, concentric rings

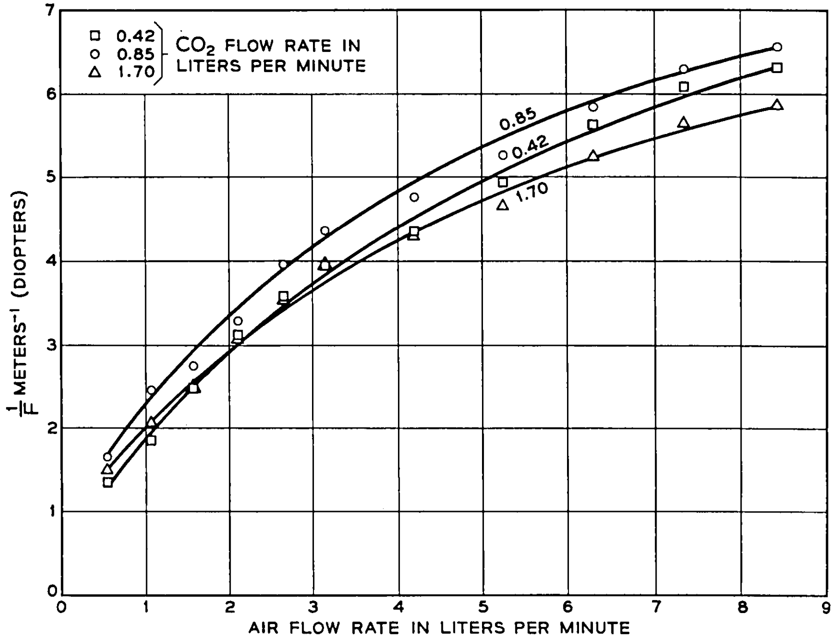
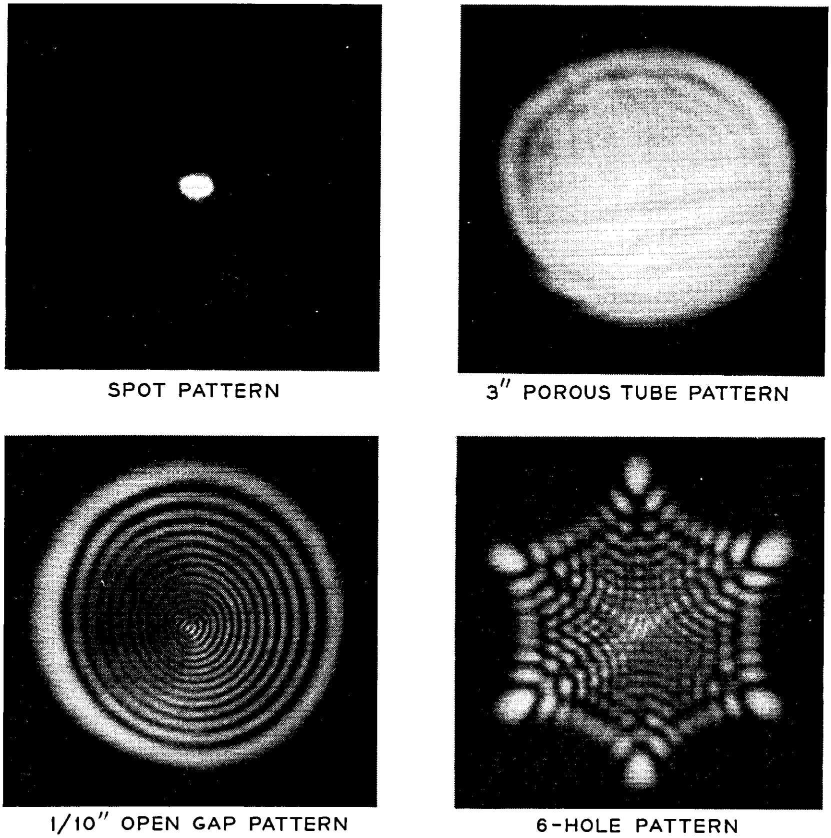


Fig. 3 — Lens focusing power versus gas flow rates, carbon dioxide-air.

of light begin to appear, as can just be seen here. These rings are similar to those which appear when spherical aberration forms a light interference pattern.[4] With a 3-inch porous tube, this effect is slight until very short focal lengths are produced. However, when the porous tube was replaced by two solid brass tubes in line with a gap of about $\frac{1}{10}$ inch between them at the center of the second gas reservoir, the light pattern shown in the photograph at the lower left was obtained. When the porous tube was replaced by a solid brass tube having six $\frac{1}{10}$-inch diameter holes drilled symmetrically every 60 degrees around its circumference in a plane located at the center of the second gas reservoir, the pattern at the lower right was observed. The use of tubes with different numbers of equally spaced holes produced patterns similar to this, having an n-fold symmetry where n is the number of holes. The steadiness and stability of these patterns of light interference are graphic indications of the stability of the gas flow, and the stability of gas lenses.

There are ways of separating the mixed gases, if desired, by permeable membranes or other means so that they can be mixed again in a later lens farther along a transmission system. Complete separation is not necessary to give lens action.



SPOT PATTERN

3″ POROUS TUBE PATTERN

1/10″ OPEN GAP PATTERN

6-HOLE PATTERN

Fig. 4 — Observed patterns.

Gas mixture lenses can have short focal lengths with moderate gas flows. The effects are remarkably steady and constant. The focal length is easily changed by changing the gas flow rates. Such lenses may have possible uses in optical transmission systems.

The assistance of W. E. Whitacre in constructing and setting up the equipment is gratefully acknowledged.

REFERENCES

1. Berreman, D. W., A Lens or Light Guide Using Convectively Distorted Thermal Gradients in Gases, B.S.T.J., **43,** July, 1964, Part 1, p. 1469.
2. Berreman, D. W., A Gas Lens Using Unlike, Counter-Flowing Gases, B.S.T.J., **43,** July, 1964, Part 1, p. 1476.
3. Marcuse, D., and Miller, S. E., Analysis of a Tubular Gas Lens, B.S.T.J., this issue, p. 1759.
4. Born, Max, and Wolf, Emil, *Principles of Optics,* Pergamon Press, 1959, pp. 472–477.

A Circle Diagram for Optical Resonators

By J. P. GORDON

(Manuscript received June 16, 1964)

A graphical representation of the relationships between the parameters of Hermite Gaussian light beams has been introduced recently[1,2] by S. A. Collins, Jr. T. Li has pointed out[3] that Collins' chart has two equivalent forms. Collins' chart relates the spot radius and phase front curvature at any position on the beam to the position and spot radius of the beam waist. In this note we point out that a similar chart can be made which relates the curvature parameters[4,5] of any two phase fronts along a Gaussian beam and the spot radii on those phase fronts. The curvature parameters are defined as $g_1 = 1 - d/R_1$, $g_2 = 1 - d/R_2$, where R_1 and R_2 are the radii of curvature of the phase fronts, and d is their separation. This new chart directly relates mirror curvatures and spot radii in a spherical mirror resonator.

The equations on which the chart is based are[5,6]

$$\frac{g_1}{g_2} = \frac{w_2^2}{w_1^2} \tag{1}$$

$$w_1 w_2 = \frac{\lambda d}{\pi}(1 - g_1 g_2)^{-\frac{1}{2}}, \tag{2}$$

where w_1 and w_2 are the spot radii on the two phase fronts. If we eliminate w_2 and g_2, respectively, from these two equations we get

$$\left(g_1 - \frac{1}{2g_2}\right)^2 + \left(\frac{\lambda d}{\pi w_1^2}\right)^2 = \left(\frac{1}{2g_2}\right)^2 \tag{3}$$

and

$$\left(\frac{\lambda d}{\pi w_1^2} - \frac{\pi w_2^2}{2\lambda d}\right)^2 + g_1^2 = \left(\frac{\pi w_2^2}{2\lambda d}\right)^2. \tag{4}$$

On a graph whose Cartesian coordinates are $\lambda d/\pi w_1^2$ and g_1, these two equations represent circles of diameters $1/g_2$ and $(\pi w_2^2/\lambda d)$, respectively, as shown in Fig. 1. The point of intersection of the two circles gives the values of w_1 and g_1 which satisfy both (3) and (4). A more complete chart, similar to Collins' chart, can be used to read off spot radii in a resonator whose mirror curvatures are known, or to find mirror curvatures from measurements of the beam spots. As does Collins' chart, this new chart has two equivalent forms, which in our case differ only in having the subscripts 1 and 2 interchanged.

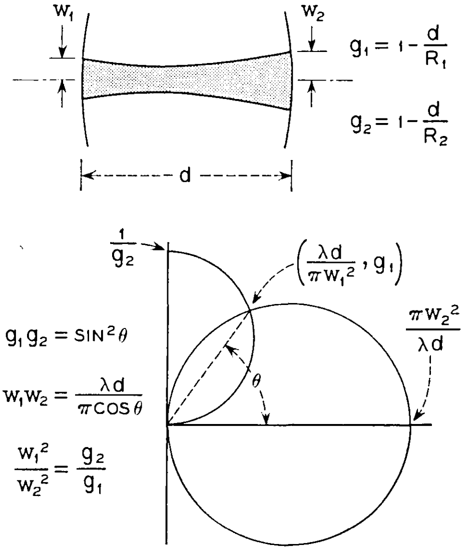

w_1
w_2
$g_1 = 1 - \frac{d}{R_1}$
$g_2 = 1 - \frac{d}{R_2}$
d
$\frac{1}{g_2}$
$\left(\frac{\lambda d}{\pi w_1^2}, g_1\right)$
$\frac{\pi w_2^2}{\lambda d}$
θ
$g_1 g_2 = \mathrm{SIN}^2\theta$
$w_1 w_2 = \frac{\lambda d}{\pi \cos\theta}$
$\frac{w_1^2}{w_2^2} = \frac{g_2}{g_1}$

FIG. 1 — Circle diagram for spherical mirror resonators. Cartesian coordinates of the important points are indicated, along with some geometrical relationships.

REFERENCES

1. Collins, S. A., Jr., J. Opt. Soc. Am., **53**, 1963, p. 1339.
2. Collins, S. A., Jr., Appl. Opt., to be published.
3. Li, T., Appl. Opt., to be published.
4. Fox, A. G., and Li, T., Proc. IEEE, **51**, 1963, p. 80.
5. Boyd, G. D., and Kogelnik, H., B.S.T.J., **41**, July, 1962, p. 1347.
6. Gordon, J. P., and Kogelnik, H., B.S.T.J., to be published.

Gas Pumping in Continuously Operated Ion Lasers

By E. I. GORDON and E. F. LABUDA

(Manuscript received June 26, 1964)

Gas ion lasers[1] operate at discharge currents of several amperes in small-bore tubing. Under these conditions the discharge acts to pump gas from the cathode to the anode[2] and pressure differences in excess of 10:1 can be established in less than one minute of discharge operation. Since the optimum pressure range for laser operation is narrowly defined relative to the range of pressures existing in the discharge tube (see Fig. 1), laser action usually deteriorates or goes out shortly after turn-on.

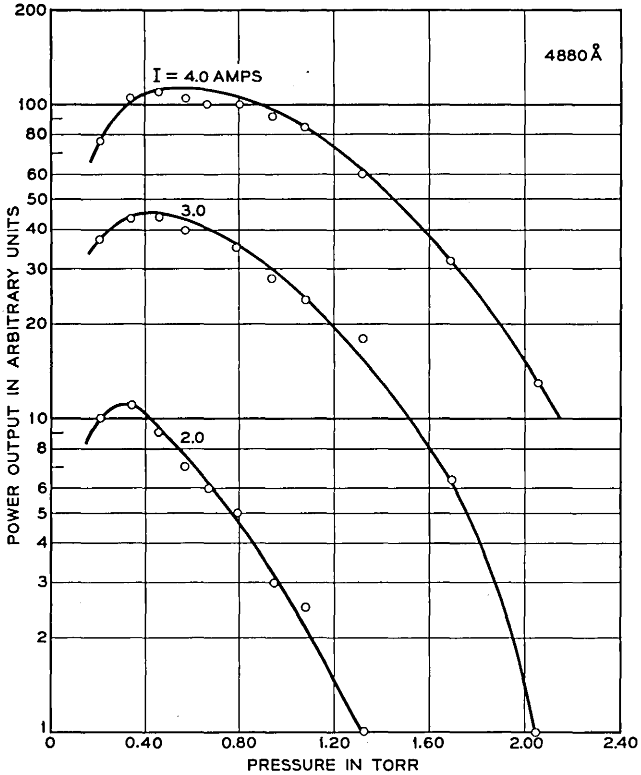


Fig. 1 — Laser output vs pressure at various discharge currents.

When the discharge is turned off for several seconds to allow the pressure to equalize and then turned on again, the output returns to its initial value only to deteriorate again.

By placing a connecting tube of high gas flow conductance between the anode and cathode as shown in Fig. 2, the pressure difference between the anode and cathode can be virtually eliminated. Tubes operated with the connecting tube show no deterioration over long periods of time. The tube length and bore are chosen so that the sustaining voltages for the two paths are comparable.

The connecting tube is fashioned in the form of a helix to relieve any strains that might develop from differential expansion and because a helix provides a convenient means of getting a long length of tubing into a small volume.

The connecting tube also serves to eliminate gas separation or cata-

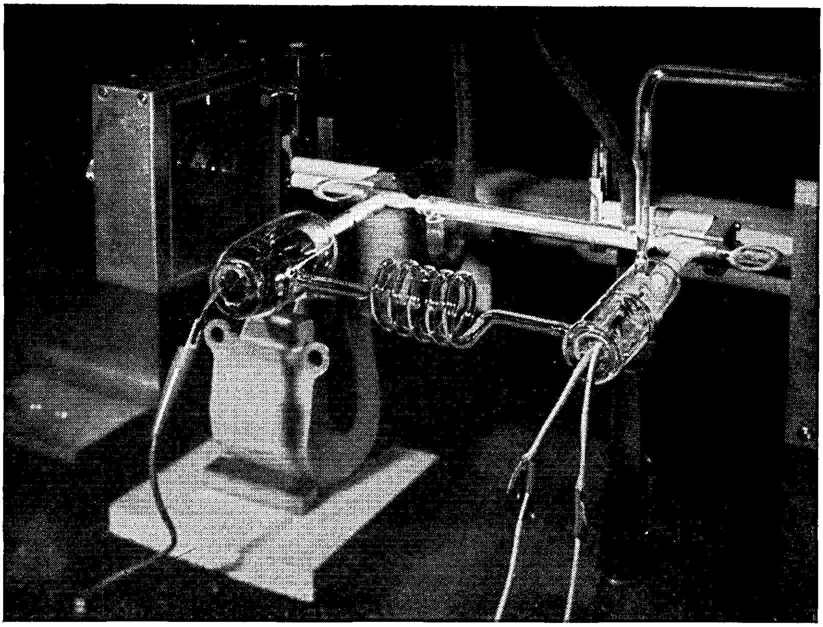


Fig. 2 — Photograph of laser illustrating the connecting tube.

phoresis that occurs when using mixed gases. This is of particular importance in dc-excited laser discharges such as helium-xenon.[3] Cataphoresis is no special problem in the helium-neon discharges in the current range at which they are operated, and the side tube is unnecessary.

We are indebted to J. T. Bannon, who constructed the experimental tubes.

REFERENCES

1. Gordon, E. I., Labuda, E. F., and Bridges, W. B., Continuous Visible Laser Action in Singly Ionized Argon, Krypton and Xenon, Appl. Phys. Lett., **4**, May 15, 1964, p. 178.
2. Francis, Gordon, *Handbuch Der Physik*, XXII, ed. S. Flugge, p. 198.
3. Bridges, W. B., High Optical Gain at 3.5μ in Pure Xenon, Appl. Phys. Lett., **3**, Aug. 1, 1963, p. 45. See also Faust, W. L., McFarlane, R. A., Patel, C. K. N., and Garrett, C. G. B., Gas Maser Spectroscopy in the Infrared, Appl. Phys. Lett., **1**, Dec., 1962, p. 4.