63-3-5

March 1963

USCEC Report 97-101
EE-22

UNIVERSITY OF SOUTHERN CALIFORNIA

SCHOOL OF ENGINEERING

FINAL REPORT

A SURVEY OF METHODS FOR DYNAMIC SYSTEM IDENTIFICATION AND RESEARCH IN OPTICAL COHERENCE

Robert B. McGhee
Gregory O. Young
Robert S. Macmillan

ELECTRICAL ENGINEERING DEPARTMENT

Engineering

EE-22

Final Report

A SURVEY OF METHODS FOR DYNAMIC SYSTEM IDENTIFICATION

and

RESEARCH IN OPTICAL COHERENCE

Robert B. McGhee
Gregory O. Young
Robert S. Macmillan

ELECTRICAL ENGINEERING DEPARTMENT
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES 7, CALIFORNIA

AF 49(638)-893

March 1963

USCEC Report 97-101

Prepared
for

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
OFFICE OF AEROSPACE RESEARCH
UNITED STATES AIR FORCE
WASHINGTON 25, D. C.

SUMMARY

Part I: A Survey of Methods for Dynamic System Identification

This section of the report summarizes the findings of the first phase of a continuing study concerned with the utilization of computers in the determination of mathematical models for dynamic systems. In this initial phase, a careful review of various types of models which have been previously used or suggested for this purpose has been completed. A discussion of these models and the associated experimental techniques makes up the first part of this section. Following a critical evaluation of the earlier methods, a new formulation of the general model inference problem is presented; this new approach has been named the "parameter space method". The computational advantages of parameter space methods are discussed, and a foundation is laid for the development of explicit computational procedures. The continuation of this research will include the investigation of several specific techniques for achieving the "identification" of nonlinear systems.

Part II: Research in Optical Coherence

A detailed study of certain aspects of the theory of optical coherence has been carried out. A major objective of the study is to obtain criteria for optimizing optical systems employing coherent sources and amplifiers such as lasers. This is the first in a series of studies to investigate the effect of various optical system devices such as antennas, filters, modulators and demodulators, etc., on the coherence properties of optical signals. The present report concerns itself primarily with spatial and temporal filters.

A general coherence function is defined as the autocorrelation function for field strengths. The field strength is a function of three space variables and time. The coherence function is therefore dependent jointly on the space separation and the time separation of the measured field strengths. It is also a function of the origin in space and time if the process is nonstationary in those variables. Limiting cases of pure spatial coherence, pure temporal coherence, pure spatial incoherence, and pure temporal incoherence are considered. The concept of perfect coherence in a particular independent variable is extended to include all deterministic functions which are completely specified for the entire range of that variable.

The general integral equation relating image and object field strengths in all variables is defined. A technique for determining

the spatial and temporal weighting function for a linear optical transducer is suggested. This evaluation is complicated by the fact that the diffraction field for an arbitrary, aperture-field, distribution has only been found (approximately) as the steady-state response to a monochromatic source. Since temporally incoherent or partially coherent sources are common, the weighting function in both space and time must be found. This function may be found approximately from the steady-state, sinusoidal, response by taking the inverse Fourier transform of the spatial response to a sinusoid over the frequency range for which the sinusoidal response function is valid. This inversion is complicated by the fact that the spatial response is a function of the temporal frequency.

The technique of obtaining the spatial and temporal weighting function makes possible the evaluation of both the transient and steady-state, statistical, behavior of far-field diffraction patterns for sources exhibiting partial coherence in both space and time. In particular, information rates may be calculated, and the transducer may be optimized with respect to its variable parameters by maximizing the image information rate with respect to these parameters.

TABLE OF CONTENTS

SUMMARY ii

PART I: A SURVEY OF METHODS FOR DYNAMIC SYSTEM IDENTIFICATION

1. INTRODUCTION 1

2. LINEAR SYSTEM IDENTIFICATION 4

3. FUNCTION SPACE DESCRIPTIONS FOR DYNAMIC SYSTEMS 19

4. PARAMETER SPACE METHODS 33

5. SUMMARY AND CONCLUSIONS 50

REFERENCES 51

PART II: RESEARCH IN OPTICAL COHERENCE 55

LIST OF ILLUSTRATIONS

Figure		Page
1	Experimental Determination of Weighting Functions.	14
2	Continuous Generation of Laguerre Coefficients by a Linear Network.	28
3	Experimental Arrangement Required for the Evaluation of Wiener Coefficients.	31
4	Description of a Transformation in Terms of Wiener Coefficients.	36
5	Parameter Space Description of a Nonlinear System.	38
6	Experimental Determination of the Parameters of a Dynamic System.	43

1. INTRODUCTION

1.1 Purpose and Scope

This report summarizes the findings of the first phase of a continuing research program concerned with the utilization of computers in the determination of mathematical models for dynamic systems from experimental observations. In this initial phase, a careful review of various types of models which have been previously used or suggested for this purpose has been completed. A discussion of these models and the associated experimental techniques makes up the first part of this report. Following a critical evaluation of the earlier methods, a new formulation of the general model inference problem is presented; this new approach has been named the "parameter space method". It is believed that parameter space techniques offer significant advantages for the actual computational determination of models for nonlinear systems in practical situations. The continuation of the research reported herein is aimed at the development of explicit computational algorithms for achieving the identification of nonlinear systems by parameter space operations. This research will provide the basis for future reports dealing with this topic.

1.2 Identification of Dynamic Systems

Since the time of Newton it has been known that the dynamical behavior of mechanical systems is governed by differential equations. In the ensuing centuries since Newton's discoveries, a complete theory of classical mechanics has been constructed which, in principle, permits

the description of the motion of any rigid body system in terms of a set of coupled differential equations. In a similar fashion, electric circuit theory has provided differential equation descriptions for lumped constant electrical devices. More generally, it is now common knowledge that a great many processes involving storage and interchange of mechanical, electrical, and other forms of energy are properly described by complicated sets of differential equations (or partial differential equations if the systems are distributed rather than lumped). Systems of this type are usually referred to as "dynamic" systems in engineering literature (1)[1].

While basic physical theories often permit the form of the differential equation for a dynamic system to be written down, as a rule the parameters of a particular system can be determined only by measurement. For example, the equation for a pendulum with viscous damping can be obtained from the simple moment equation:

$$I\ddot{\theta} + B\dot{\theta} = - M g l \sin \theta \qquad (1)$$

However, before any analysis of the behavior of a particular pendulum may be effected, it is necessary to determine the constants in the equation. Such determination of a specific quantitative model has been variously referred to as the "identification problem" (2), the "characterization problem" (3), and the "parameter estimation" problem (4), depending somewhat upon the methods employed; it can be accomplished only by experimentation.

---

[1] In this report, superscript numbers are used to indicate footnotes while bracketed numbers refer to the list of references at the end of the text.

In the case of a device as simple as a pendulum, the evaluation of the unknown constants is not difficult unless a high degree of precision is required. On the other hand, there are many practical situations in which the measurement of system constants is extremely difficult due either to the complexity of the processes involved or the subtleness of the effects to be measured. Furthermore, even after the system parameters have been established to a sufficient accuracy, analytic solution of the resulting family of differential equations is an extremely difficult task in all but the simplest of situations.

For the reasons just mentioned, engineers and physicists over the past several decades have sought ways of characterizing physical systems other than by a set of differential equations. Such efforts have been quite successful with respect to the important but restrictive class of systems possessing the properties of linearity and time invariance. For more general classes of systems, results have been sparse. If the requirement of time invariance is relaxed while linearity is retained, weighting functions offer an alternative to differential equation descriptions although they may be very difficult to obtain. For general nonlinear systems, it is only by considering a system as an operator effecting a transformation from an input function space to an output space that it has been possible to arrive at alternate descriptions.

The following paragraphs discuss various methods which have been used to describe dynamic systems and suggest a particular characterization to be used as the basis for the development of practical computation procedures for identifying unknown nonlinear systems. The discussion will be confined

to lumped constant systems; i.e., to systems described by ordinary rather than partial differential equations. For the sake of completeness and continuity, the development to follow begins with a review of methods which have been developed exclusively for the characterization of linear systems.

2. LINEAR SYSTEM IDENTIFICATION

2.1 Frequency Response

2.1.1 Introduction

When a system is linear, the "principle of superposition" may be applied to decompose arbitrary forcing functions and initial conditions into component parts whose effects may be more easily analyzed. The individual responses of these components may then be added to produce the total system response. If the system under consideration is time invariant as well as linear, this type of analysis is especially simple and effective. In particular, when a sinusoidal or complex exponential decomposition of the input signal is utilized, Fourier transform techniques may be applied to determine the system output. In this approach, the system under consideration is completely determined by the function, $G(\omega)$, defined by [2]

$$Y(\omega) = G(\omega)\, X(\omega) \tag{2}$$

where $X(\omega)$ and $Y(\omega)$ are Fourier transforms of the input function, $x(t)$, and the output function, $y(t)$, respectively. The function $G(\omega)$ is usually called the system "frequency response" (5,6,7). For a linear time

---

[2] While input and output variables will be treated as scalars in this discussion, no difficulties are experienced if equation 1 is a vector-matrix relationship.

invariant system, $x(t)$ and $y(t)$ are related by

$$Ay(t) = Bx(t) \tag{3}$$

where A and B are linear constant coefficient differential operators; the frequency response, $G(\omega)$, is therefore a rational polynomial in ω in such circumstances.

2.1.2 Experimental Determination of Frequency Response

Within the class of stable, linear, time invariant systems, experimental evaluation of the frequency response function is carried out by a procedure which is (mathematically) independent of the nature of the system. This is a major advantage of the frequency method. No assumptions regarding the system must be made beyond the basic constraint that it be stable, linear, and time invariant. In the usual technique employed to experimentally determine the frequency response, $G(\omega)$, the input, x, is chosen to be the function,

$$x(t) = a \cos \omega_1 t \tag{4}$$

so that after a time sufficient for the decay of transients has elapsed

$$y(t) = b \cos (\omega_1 t + \phi). \tag{5}$$

It is easily demonstrated that the "frequency response" at $\omega = \omega_1$ is then given by

$$G(\omega_1) = \frac{b}{a} e^{j\phi} = \alpha + j\beta \tag{6}$$

The variables $\frac{b}{a}$ and ϕ are commonly called "gain" and "phase shift" respectively and are measurable by standard electronic devices.

The "Nyquist plot" is the locus of $G(\omega)$ in α,β coordinates. Special purpose devices have been constructed to plot Nyquist diagrams automatically since the point by point determination of $G(\omega)$ may be very tedious (8).

2.1.3 Measurement of Frequency Response in the Presence of Interfering Noise

When extremely precise measurement is attempted or uncontrollable sources of interfering noise are present, it is found that the measured gain and phase shift are random variables. To circumvent this difficulty, cross-correlation techniques may be used to discriminate against the noise (9). From equation 6, $y(t)$ may be written

$$y(t) = \alpha a \cos \omega_1 t - a\beta \sin \omega_1 t \tag{7}$$

With noise, $n(t)$ added, this becomes

$$y'(t) = a\alpha \cos \omega_1 t - a\beta \sin \omega_1 t + n(t) \tag{8}$$

Now assuming that $y(t)$ and $n(t)$ are uncorrelated, it follows that

$$\begin{aligned} &\lim_{T \to \infty} \frac{2}{T} \int_0^T y' \cos \omega_1 t \, dt = a\alpha \\ &\lim_{T \to \infty} \frac{2}{T} \int_0^T - y' \sin \omega_1 t \, dt = a\beta \end{aligned} \tag{9}$$

Finite time approximations to these cross-correlation functions can yield arbitrarily precise measurements of α and β in the presence of arbitrarily large noise voltages.

2.1.4 Extension of Frequency Methods to Nonlinear Systems

As a result of the simplicity and practical utility of the frequency response characterization of linear systems, attempts have been made to extend frequency methods to certain types of nonlinear systems. Perhaps the most significant of these is the "describing function" method due to R. J. Kochenberger (10). Kochenberger's method consists of determining the gain and phase characteristics of a nonlinear element with respect to the fundamental component of the output only. In general, the frequency response defined in this way depends not only upon the frequency of the test signal, but also upon its magnitude. Describing functions have found considerable use in the investigation of limit cycles in nonlinear systems. Since the method is approximate, it is of relatively little value in determining the transient behavior of nonlinear systems.

The method for experimentally determining describing functions is the same as for experimentally determining frequency response except that the describing function depends on both a and ω in equation 6 rather than on ω alone.

2.2 Determination of the System Transfer Function

2.2.1 Relationship Between the System Laplace Transform and the Frequency Response Function

If the Laplace transform is applied to both sides of equation 3, the result is

$$A(s)\,Y(s) = B(s)\,X(s) \tag{10}$$

so

$$\frac{Y(s)}{X(s)} = \frac{B(s)}{A(s)} = H(s) \tag{11}$$

The function $H(s)$ is commonly called the system transfer function. Again, since A and B are assumed to be linear constant coefficient differential operators, $H(s)$ is a rational polynomial. As is well known, the frequency response function, $G(\omega)$, may be obtained from $H(s)$ by utilizing the relation

$$G(\omega) = H(j\omega) \tag{12}$$

Furthermore, since $H(s)$ is an analytic function, its behavior along the imaginary axis in the s plane completely specifies its behavior over the entire complex plane. So, conversely, experimentally evaluated frequency response may be used to determine $H(s)$ for stable systems. The process of obtaining $H(s)$ from $G(\omega)$ is called the "approximation problem" in network theory (11,12,13,14).

2.2.2 Direct Measurement of the System Transfer Function

Up to the present time, the frequency response method has been overwhelmingly favored in practice for the experimental characterization of linear time invariant systems. Recently, however, a few investigators have devised methods for measuring pole and zero locations directly by using special test signals. These are sometimes referred to as "time domain" methods (15,16,17,18). While these techniques have the advantage that the approximation problem is by-passed entirely, the test equipment required is generally more complicated. In any event, the final result of the measurement is mathematically equivalent to the frequency response.

2.3 Weighting Function Methods

2.3.1 Relation to the System Transfer Function

The behavior of any linear system with respect to an input or disturbance at a particular point is completely determined by its weighting function or "impulse response" (19). If $h(t,\tau)$ is the weighting function associated with a linear differential operator, L, and $y(t)$ are input and output variables respectively, then the solution to

$$Ly = x \tag{13}$$

is given by

$$y(t) = \int_{-\infty}^{t} h(t,\tau)\, x(\tau)\, d\tau \tag{14}$$

If L is also time invariant, then $h(t,\tau)$ becomes a function of $(t-\tau)$ only and is related to the system transfer function by

$$\mathcal{L}\{h(t-\tau)\} = e^{-\tau s} H(s) \tag{15}$$

where $\mathcal{L}$ denotes the Laplace transform operator (19). Thus the weighting function and frequency response form a Laplace-Fourier transform pair.

While the weighting function bears a simple relationship to frequency response only in the time invariant case, it is nevertheless (in light of equation 14) a perfectly general way of characterizing any linear system. The determination of $h(t,\tau)$ from the operator L is apt to be a formidable task, however. Most often, numerical methods or analog simulation are required to find $h(t,\tau)$.

2.3.2 Direct Measurement of Weighting Functions

If the system under test is initially in a relaxed state, then for $t > 0$, the lower limit in the integral of equation 14 may be replaced by zero. If $x(t)$ is then chosen to be a narrow pulse of unit area centered at $t = t_1 > 0$, the system weighting function can be measured directly. Suppose that $x(t)$ is narrow enough to be sufficiently approximated by a unit impulse function, $\delta(t-t_1)$. Then

$$y(t) = \int_0^t h(t,\tau)\ \delta(t-t_1)\ d\tau = h(t,t_1) \qquad (16)$$

By repeating this experiment with various values of t_1, a family of curves for $h(t,\tau)$ can be constructed.

When L is time invariant, it is sufficient to make a single experiment with $t_1 = 0$ since $h(t,\tau)$ depends only on $t-\tau$. In this respect, weighting function measurement of system dynamics is markedly superior to the frequency method which is a point by point procedure. On the other hand, the equipment needed to record transient behavior is usually more complicated and restricted in applicability than that used in determining frequency response.

E. Miskin and R. A. Haddad have described an adaptive system which uses integrals of impulse functions (i.e., steps, ramps, and parabolas) for both identification and control of an object whose weighting function is initially unknown (20). This method permits the control and identification functions to proceed simultaneously but it is restricted to situations where measurement noise is insignificant.

2.3.3 Determination of Weighting Functions by Cross-Correlation

All of the methods mentioned thus far for characterizing linear systems share a common weakness. In order to make measurements on the device in question, it is necessary to remove it from its normal function and apply special test signals. That is, the methods described so far involve essentially laboratory or "bench test" types of measurements. There are many important situations in which it is either impossible or undesirable to create such controlled conditions. In such cases, cross-correlation of input and output may be used to discriminate against the unwanted effects.

The application of cross-correlation to frequency testing has already been discussed in this report. In 1950, J. B. Wiesner and Y. W. Lee pointed out that cross-correlation can also be used to advantage in the measurement of weighting functions (21). Following their suggestion, suppose that the input to a system, $x(t)$, is a wide-band stationary Gaussian process. Then the autocorrelation function of this process is given by

$$\phi_{xx}(\tau) = E\{x(t)\, x(t+\tau)\} = N_o\, \delta(\tau) = \phi_{xx}(-\tau) \qquad (17)$$

where N_o is the (two sided) spectral density of the input variable, $x(t)$. If the cross-correlation between the input, x, and the output, y, is defined by

$$\phi_{yx}(t,\tau) = E\{y(t)\, x(t+\tau)\} \qquad (18)$$

then

$$\begin{aligned}\phi_{yx}(t,-\tau) &= E\{y(t)\, x(t-\tau)\} \\ &= E\left\{x(t-\tau) \int_0^t h(t,\alpha)\, x(\alpha)\, d\alpha\right\} \\ &= \int_0^t E\{x(t-\tau)\, x(\alpha)\}\, h(t,\alpha)\, d\alpha \\ &= \int_0^t N_o\, \delta(\alpha-t+\tau)\, h(t,\alpha)\, d\alpha \\ &= N_o\, h(t,t-\tau) \end{aligned} \tag{19}$$

or

$$h(t,t-\tau) = \frac{\phi_{yx}(t,-\tau)}{N_o} \tag{20}$$

In the special situation when ϕ_{yx} is not time dependent, then equation 20 reduces to

$$h(\tau) = h(-\tau) = \frac{\phi_{yx}(-\tau)}{N_o} \tag{21}$$

which may be replaced by a time average:

$$h(\tau) = \lim_{T \to \infty} \frac{1}{T} \int_0^T y(t)\, x(t-\tau)\, dt \tag{22}$$

This is the result originally pointed out by Lee and Weisner (21) and the one most often used in weighting function methods for system identification. The input process, $x(t)$, may represent either normal operating signals alone or such signals plus a deliberately added wideband test signal. In the event that a special test signal is added, it is only this signal which is cross-correlated with the system output.

2.3.4 Experimental Evaluation of Cross-Correlation Functions for Time Invariant Linear Systems

Following the original publication of Lee and Weisner, a sizable number of investigators have explored various techniques for the experimental evaluation of $\phi_{yx}(\tau)$. Despite the considerable originality of some of these methods, nearly all of them contain the basic elements of delay, multiplication, and averaging as shown in Figure 1. Since the averaging time, T, must be finite, the output of the process identifier is $\hat{h}(\tau)$, an estimator of $h(\tau)$. The mean square deviation of $\hat{h}(\tau)$ from $h(\tau)$ may be made as small as desired by increasing the averaging time sufficiently. The next few paragraphs discuss some specific hardware mechanizations of equation 22 as well as some minor variations on the basic scheme.

J. A. Aseltine, et al. have suggested that a binary test signal offers particular advantages in the experimental determination of weighting functions (22). If an input is used which switches from one level to the other in the manner of a Poisson process, then the delay time, τ, is easily realized by either a delay line or a shift register while the multiplication required is replaced by a simple on-off electronic gate. While a Gaussian random process has most often been used as a probing signal for weighting function determination, equation 21 remains valid for any input with an autocorrelation function which is sharp compared to $h(\tau)$. Consequently, a binary signal is entirely satisfactory as a test signal so long as the average number of switchings per second considerably exceeds the bandwidth of the system under test.

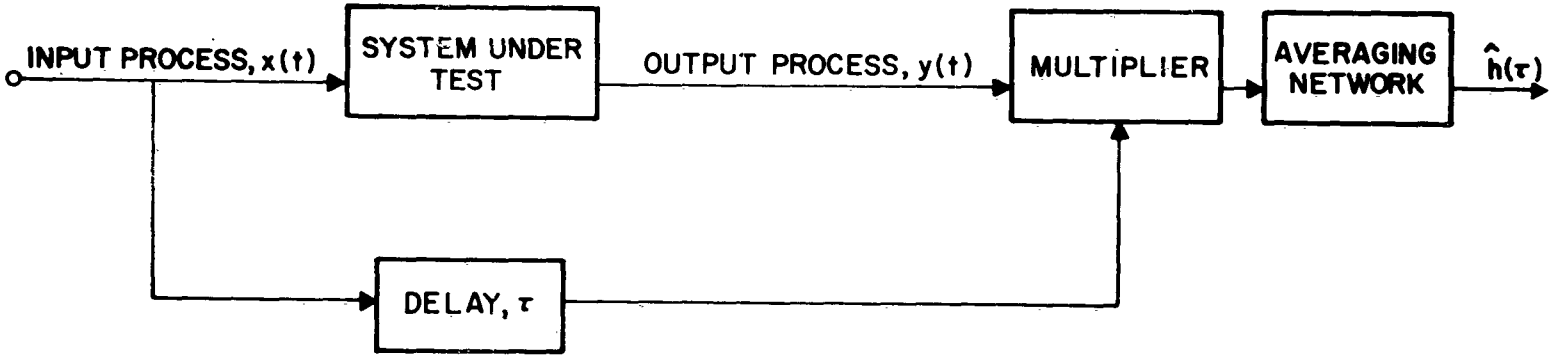

FIG. 1 EXPERIMENTAL DETERMINATION OF WEIGHTING FUNCTIONS.

G. L. Turin has discussed the replacement of explicit elements of Figure 1 by a filter matched to a particular test signal (23). When this is done, the physical equipment involved is greatly simplified. Moreover, $h(t)$ is obtained as a continuous function of time rather than on a point by point basis as in the conventional cross-correlation method. On the other hand, Turin's method does not permit long averaging times to be used to discriminate against outputs resulting from extraneous noise or normal operating signals. W. W. Lichtenberger has suggested that this difficulty can be circumvented by averaging the filter response over a number of measurements (24). However, this again yields a point-wise estimate of $h(\tau)$ unless an analog memory is available to store successive samples of $h(\tau)$ in their entirety.

Several investigators have used multiple sinusoidal test signals followed by syncronous detection in place of a Gaussian input (25,26,27,28). In these schemes it is system coefficients which are to be determined directly rather than the weighting function. Unlike simple frequency response testing, these methods use averaging following the syncronous detection and are therefore able to function in the presence of interfering signals. Despite the fact that the test signals are sinusoidal, this type of measurement is a correlation method which fits into the framework of Figure 1.

Quite a number of investigations have been directed toward the problem of determining weighting functions by correlation methods utilizing normal operating records without the introduction of special test signals. The first significant work in this direction appears to have been accomplished by T. P. Goodman and J. B. Reswick (29). They constructed a special piece

of equipment involving a tapped delay line and manually set weighting potentiometers to match the cross-correlation properties of a system under test. The unknown weighting function evaluated at evenly spaced time intervals appeared as potentiometer settings at the end of an iterative manual adjustment process. R. E. Kalman has demonstrated an automatic iterative computational technique for obtaining a least squares estimate of a z-transformed version of the system weighting function during normal operation (30). P. Joseph et al. have proposed a variation in Kalman's method which allows initial conditions to be considered (31). R. B. Kerr and W. H. Surber have examined the relationship between time of observation and accuracy of weighting function estimation (32). V. V. Solodovnikov and A. S. Uskov have discussed transform methods for the solution of the equation

$$\phi_{yx}(\tau) = \int_0^\infty \phi_{xx}(\tau-t)\, h(t)\, dt \qquad (23)$$

for $h(t)$ (33). This equation arises when the input correlation function is not extremely narrow compared to $h(t)$ and the averaging time of Figure 1 is allowed to become infinite.

There have been many other publications dealing with cross-correlation since the original observation of Lee and Weisner. Since the present discussion is intended simply to place the methods proposed in this report in perspective, the examination of cross-correlation techniques has not been exhaustive. The bibliographies attached to the publications referenced provide many other sources of information concerning this topic.

2.3.5 Identification of Time Varying Linear Systems by Cross-Correlation

The experimental determination of time varying weighting functions using equation 20 is seriously hampered by the fact that expected value operators cannot be replaced by time averages. Rather, it is necessary to estimate $\phi_{yx}(t-\tau)$ by averaging the results of many separate experiments. Furthermore, since $h(t,\tau)$ is a function of two variables, these repeated experiments must be carried out for each value of τ to be considered. Thus the total number of experiments involved is very large. Moreover, the system must be restored to a reference condition $(t=0)$ at the start of each experiment. It may not be feasible to achieve this type of operation in many systems of practical concern. Despite these difficulties, equation 20 appears to offer promise of useful application in special situations. The necessary storage and averaging could be accomplished by a digital machine possessing analog inputs and outputs. As far as is known to the writer, no application of this approach has been made to date.

2.3.6 Impulse Response Measurement by Regression Analysis

Due to the force of tradition more than anything else, analog measuring and computing devices have been assumed almost exclusively in the investigation of practical procedures for impulse response measurement. Mathematically, this means that the techniques used have been restricted to processes of analysis to the virtual exclusion of algebraic methods. By contrast, classical statistics is concerned mainly with discrete data obtained from repeated experiments so algebraic methods predominate in that subject. Thus, when conventional statistical methods are applied to

the measurement of a system impulse response, a fresh viewpoint results. An example of the statistical approach has been provided by M. J. Levin in an investigation dealing with the estimation of impulse response at discrete values of time by linear regression analysis (34). In Levin's approach, the convolution integral

$$y(t) = \int_0^t h(\tau)\, x(t-\tau)\, d\tau \qquad (24)$$

is replaced by an approximating convolution summation

$$y(NT) = \sum_{n=0}^{N} h(nT) \left\{ x\left[(N-n)T\right] \right\} T \qquad (25)$$

In this expression, x and y are physically measurable while h(nT) is unknown. Since the equation is linear in the unknown weighting function values, a set of N+1 simultaneous equations resulting from application of equation 25 may be solved for h(nT), n=0, 1, ··· N, by matrix inversion. However, since measurement errors are invariably present, Levin suggests that redundant data be taken and h(nT) be determined by least squares regression analysis. It is shown in his paper that this results in estimates which are optimum in several senses.

By substituting equation 25 for equation 24 the difficult problem of finding an inverse operator is reduced to the much easier one of finding the inverse of a matrix. This is an operation well suited to a digital computer. In a paper dealing with future trends in engineering analysis, Denis Gabor has pointed out that it is quite typical that the discrete problem should yield a solution more readily than the continuous problem (35). It is further stated in his article that algebraic methods can be expected to

replace analytic approaches to a great many problems in applied mathematics. This opinion, it seems, is supported by the fact that high speed computers are now available to most mathematicians and engineers. While these machines have great algebraic power, they are not naturally suited to limiting processes.

The point of view expressed by Levin appears to offer an attractive alternative to correlation methods. The continuation of the research discussed in this report involves the application of similar approaches to nonlinear systems.

3. FUNCTION SPACE DESCRIPTIONS FOR DYNAMIC SYSTEMS

3.1 Introduction

All of the methods employed for linear system identification appear to run aground when attempts are made to apply them to general nonlinear systems. This unfortunate situation results from the necessity of abandoning the principle of superposition in dealing with nonlinear systems. In order to gain a vantage point on nonlinear systems comparable to the conventional treatment of linear systems, it seems to be necessary to introduce the much more sophisticated idea of transformations defined over function spaces (36). That is, it is necessary to recognize that the output of a physical system in the most general case depends upon the entire past history of its input in some complicated nonlinear fashion. This section provides a brief summary of the function space point of view.

The application of function space methods is considerably facilitated by the expansion of input functions into orthogonal series. The particular

type of orthogonal functions used for this expansion will depend upon the nature of the input and the test being conducted. For example, if the input is periodic, then a Fourier series expansion is appropriate. On the other hand, if the input is not known to be periodic but is known to have finite bandwidth, B, it can be completely represented by time samples separated by intervals equal to 1/2B. This representation is called a "signal space" expansion in communication theory (37). If the input is a general random process unspecified except for the properties of continuity and finite variance, then the entire past history of its values may be summarized by the coefficients of an expansion in terms of Laguerre functions (38). Whatever the type of expansion chosen, the result is that the input may be thought of as either a point or a curve in a space of infinite dimension; i.e. in a Hilbert space. Representation as a point occurs when the entire input is known in advance while a space curve results when the input is a random process with only past values known and a "backward looking" expansion is used.

When the input to a system has been appropriately expanded, the output of the system may be regarded as the result of a mapping or transformation defined over the input space. Physically, this transformation is produced by the reaction of the system under test to a "probing" signal applied at an input (38). When the entire input can be represented as a stationary point in a Hilbert space, the output can be thought of as another point in a similar space. This is the approach used, for example, in steady state harmonic analysis of linear systems. When the input is represented by a moving point in a Hilbert space as in the case of a random input process,

then the output space is just the real line. That is, the entire past history of the input determines the present output; future inputs and outputs are not known.

When the function space point of view is adopted, the system identification problem becomes a matter of determining what transformation characterizes the system under test. Again, there is considerable freedom of choice in selecting a method for describing this transformation. For example, if the system is known to be linear and time invariant, then the frequency response is an appropriate characterization of the system input to output transformation. Specifically, if an input x, is periodic in an interval T and is square integrable within that interval, then

$$x(t) = \sum_{n=-\infty}^{+\infty} a_n \, e^{j \frac{2\pi n t}{T}} \tag{26}$$

The response of the system under test to this forcing function is given by

$$y(t) = \sum_{n=-\infty}^{+\infty} b_n \, e^{j \frac{2\pi n t}{T}} \tag{27}$$

where

$$\frac{b_n}{a_n} = G\left(\frac{2\pi n}{T}\right) \tag{28}$$

The complex function, $G(\omega)$, is just the frequency response defined by equation 2. Since x may be thought of as a point or vector in a space whose coordinates are the complex quantities, a_n, and y is likewise a

point in the b_n space, equation 28 is an example of a particularly simple transformation from a Hilbert space to a Hilbert space.

While equation 28 represents merely a rephrasing of the significance of frequency response, as indicated, the function space point of view is not limited to linear systems. It is possible to choose families of orthogonal functions defined over the input Hilbert space coordinates which are capable of representing arbitrary nonlinear operators [3]. Norbert Wiener, for example, chose to represent a Gaussian input process in terms of normalized Laguerre functions and then described the system output as an expansion involving products of normalized Hermite functions whose arguments are the Laguerre coefficients (3,38). In electing to use these particular series, Wiener restricted himself to the use of broadband Gaussian random processes as a source of probing or test signals and to time invariant operators having the property that inputs applied arbitrarily far in the past have an arbitrarily small influence on the present system output. This latter restriction means, for one thing, that the Wiener expansion cannot be used for the study of unstable or oscillating nonlinear systems.

It is certainly possible to choose expansions different from the one chosen by Wiener to characterize a nonlinear operator. For example, A. Bose has discussed an expansion which involves basically the idea of partitioning a Hilbert space into cells and associating a particular output with each cell (3). A. W. Balakrishnan has discussed the application of polynomials defined over Hilbert spaces to the problem of nonlinear operator representation (39). L. A. Zadeh has provided a tutorial exposition of the Hilbert

---

[3] The operator may be of an arbitrary nature providing only that it produces outputs which lie in the selected output function space.

space characterization of nonlinear operators wherein several more expansions are described (36) [4].

No matter what orthogonal functions are used as a basis for the representation of the input and the system transformation, with each such pair of expansions there will be a class of inputs and a class of transformations which can be represented. The choice of representation is thus determined by the problem to be solved.

3.2 A Function Space Definition of "Static" and "Dynamic" Systems

Function space terminology provides a basis for a precise definition of the terms "static" and "dynamic" as applied to physical systems. A static system is one in which the output is a function of the present input only. Thus, to represent a general static transformation, it is sufficient to expand the output variable in terms of orthogonal functions whose argument is simply the present value of the input; it is not necessary to utilize a Hilbert space description for the input. A static system, therefore, is one which effects a transformation from a real line to a real line. In contrast to static systems, in a dynamic system the output depends not only upon present values of the input but also upon past values. Consequently, a dynamic system performs a transformation from a Hilbert space to a line. Such systems are sometimes said to have "memory" or "energy storage" while static systems are often described as being "memoryless". By its nature, the determination of the output of a dynamic system requires a double orthogonal expansion. First, the past of the input must be expanded and

---

[4] Zadeh also points out that infinitely iterated integrals can be used with appropriate weighting functions to obtain an alternate representation for nonlinear operators which does not involve expansion of the input.

then the resulting coefficients used as independent variables for a multivariable expansion representing the transformation to the output variable. In the remainder of this report, the adjectives "static" and "dynamic" will always be used in accordance with the above definitions.

3.3 Experimental Evaluation of the Characteristics of Static Nonlinear Devices

While this report is concerned fundamentally with the characterization and identification of nonlinear dynamic systems, some interesting simplification of function space methods occur when the system under test is either static or linear. In the case of static nonlinear systems, the output depends only on the present value of the input so an expansion of the input in orthogonal functions is not necessary. The output may be represented simply as

$$y(t) = f[x(t)] = \sum_{i=1}^{\infty} a_i \, \phi_i \, [x(t)] \qquad (29)$$

which is a single rather than a double expansion. L. A. Zadeh has described several suitable orthonormal sets of functions, $\{\phi_i(x)\}$, in a paper dealing with static nonlinearities which are completely defined by their describing functions (2). H. J. Lory, et al. have discussed the application of harmonic analysis utilizing growing real exponentials to obtain the coefficients of a Taylor series expansion of $f(x)$ (40).

3.4 Simplification of Function Space Representations for Linear Time Invariant Dynamic Systems

When the system under test is linear and time invariant, only an input function space is required. The transformation from the input function space

to the output space need not be expanded in orthonormal functions since the linearity of the system demands that the output be expressable as a simple linear function of the input Hilbert space coordinates. For example, Y. W. Lee shows that

$$h(t) = \sum_{i=1}^{\infty} c_i \, A_i(t) \tag{30}$$

where $A_i(t)$ are the coefficients of an expansion of the part of the system input in terms of orthonormal Laguerre functions and the c_i are coefficients characterizing the system (41). The orthonormal Laguerre functions used by Lee are obtained from an orthogonalization of the family of functions

$$f_n = (\alpha t)^n \, e^{-\alpha t} , \qquad n = 0,1,2, \cdots \infty \tag{31}$$

Another example of linear system representation utilizing a single expansion is provided by T. P. Goodman and J. B. Reswick (29). As previously mentioned, Goodman and Reswick used a finite term approximation to the expression

$$y(t) = \lim_{\substack{T \to 0 \\ NT = t}} \sum_{n=0}^{N} T\{ x(t-nT) \, h(nT) \} \tag{32}$$

In their mechanization, the values of $x(t-nT)$ are obtained from a tapped delay line while $T \cdot h(nT)$ represents potentiometer settings weighting these delayed values. Other methods for describing linear systems in terms of a single expansion are given in Lee (41) and T. Kitamori (42).

3.5 The Wiener Theory of Nonlinear Systems

While the idea of expanding operators over Hilbert spaces is not new, it appears that only recently has an explicit experimental technique been proposed for the evaluation of the coefficients in such expansions. A specific method was described by Norbert Wiener in a summer lecture series in 1953-54. However, the first documentation of Wiener's approach apparently occurred with the publication of A. Bose's dissertation in 1956 (3). Since it is felt that the experimental approach taken by Wiener represents the only reasonably well explored alternative to the methods proposed in this report, the elements of the Wiener theory are summarized fairly completely in the following paragraphs.

As proposed by Wiener, the input to a system is expanded in Laguerre functions. These functions are derived from the conventional Laguerre polynomials by including the square root of the Laguerre weighting function, e^{-t}, as a multiplying factor on each polynomial. Thus, since the nth Laguerre polynomial is given by

$$L_n(t) = \frac{1}{(n-1)!}\, e^{t}\, \frac{d^{(n-1)}}{dt^{(n-1)}}\left(t^{(n-1)}e^{-t}\right), \quad n=1,2,3 \cdots \infty \tag{33}$$

the n-th Laguerre function may be written

$$h_n(t) = \begin{cases} e^{-\frac{t}{2}} L_n(t) & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{34}$$

The functions $h_n(t)$ are orthonormal over $[0, \infty]$ with unit weighting function. Using these functions, the past of the input may be expanded at any instant as

$$x(-t) = \sum_{n=1}^{\infty} u_n \, h_n(t) \qquad t \geq 0 \qquad (35)$$

where

$$u_n = \int_0^{\infty} x(-\tau) \, h_n(\tau) \, d\tau \qquad (36)$$

Now this equation has the form of a convolution integral. Furthermore, the functions, $h_n(t)$, are of the same form as the impulse response of a linear lumped constant network with a pole of order n. It is to be expected, therefore, that the desired Laguerre coefficients can be obtained continuously in time by feeding $x(t)$ into a bank of appropriate linear filters. Wiener pointed out that this is indeed the case. The appropriate filter transfer function may be determined by application of the Laplace transform to equation 33 yielding (41)

$$H_n(s) = \frac{1}{s + \frac{1}{2}} \left(\frac{s - \frac{1}{2}}{s + \frac{1}{2}} \right)^{n-1} \qquad (37)$$

This result is illustrated in Figure 2. It is interesting to note that the filter depicted is nothing more or less than a low pass filter followed by a lumped constant approximation to a tapped delay line [5]. This type of filter is easily constructed using standard analog computer elements.

---

[5] Since the Laguerre functions form a complete basis for bounded, continuous, square integrable functions, any nonlinear system of the class treated by Wiener can be represented (for such inputs) by a bank of linear filters (as in Figure 2) followed by a "zero memory" nonlinear function generator having the $u_i(t)$ as inputs and $y(t)$ as an output.

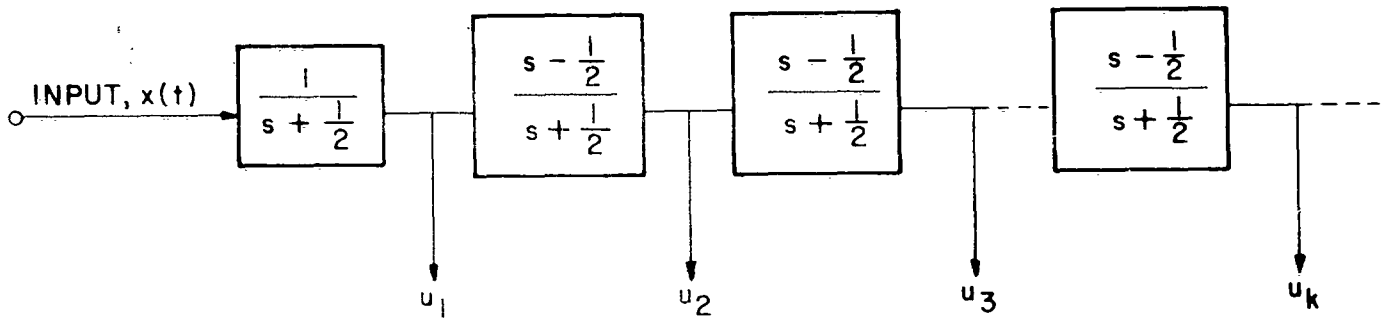

FIG. 2 CONTINUOUS GENERATION OF LAGUERRE COEFFICIENTS BY A LINEAR NETWORK.

In order to fully exploit the possibilities of a Laguerre function expansion of the input signal, the Wiener theory requires that the input used to probe the system response be a broadband Gaussian process such as shot noise. With this choice, it turns out that the Laguerre coefficients are themselves uncorrelated Gaussian random processes with equal variances. This being the case, it seems natural to expand the system operator in Hermite functions since the Hermite polynomials are orthonormal over $[-\infty, \infty]$ with a Gaussian weighting function. If $\eta_n(x)$ is the n-th Hermite polynomial (orthonormal with e^{-x^2} weighting function), then the n-th Hermite function as defined by Wiener is given by

$$\phi_n(x) = e^{-\frac{x^2}{2}} \eta_n(x) \tag{38}$$

These functions are orthonormal in $[-\infty, \infty]$ with unit weighting function. Weiner has shown that the transformation from the Laguerre coefficient input space to the system output can be written as an expansion in terms of Hermite functions. Specifically (3)

$$y(t) = \lim_{s \to \infty} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \cdots \sum_{h=1}^{\infty} a_{ij\cdots h} \; \phi_i(u_1) \, \phi_j(u_2) \cdots \phi_h(u_s) \tag{39}$$

The coefficients in this expansion, $a_{ij\cdots h}$, can be determined by multiplying both sides of the equation by the appropriate products of $\phi_n(x)$ functions and averaging over $[-\infty, \infty]$. However, due to Wiener's judicious choice of input signals and expansions, the necessary averaging can be accomplished by

simply cross-correlating the system output with the proper products of Hermite polynomials. Thus (3,38)

$$a_{ij\cdots h} = (2\pi)^{s/2} \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} y(t)\, \eta_i(u_1)\, \eta_j(u_2) \cdots \eta_h(u_s)\, dt \qquad (40)$$

This can be written more compactly by denoting the set of subscripts by α, the product of Hermite polynomials by $V_\alpha(\vec{u})$, and the time averaging by the conventional bar symbol of statistics. Thus

$$a_\alpha = (2\pi)^{s/2}\, \overline{y(t)\, V_\alpha(\vec{u})} \qquad (41)$$

This equation summarizes the experimental part of the Wiener theory. Figure 3 is a schematic representation of the equipment required to carry out the operations indicated (3).

3.6 Difficulties Associated with Implementation of the Wiener Theory

Despite the apparent simplicity of equation 41, about all that can be said in its favor is that only a countable infinity of limiting operations is involved in the evaluation of the required coefficients. This is of scant comfort to an investigator faced with the task of actually determining the characteristics of a real physical system. In order to actually make use of the Wiener theory, it is necessary to truncate all of these limiting operations both with regard to measurement time and the number of terms taken in the series for $y(t)$. So far as is known to the writer, there has been no analysis of the errors of such truncations.

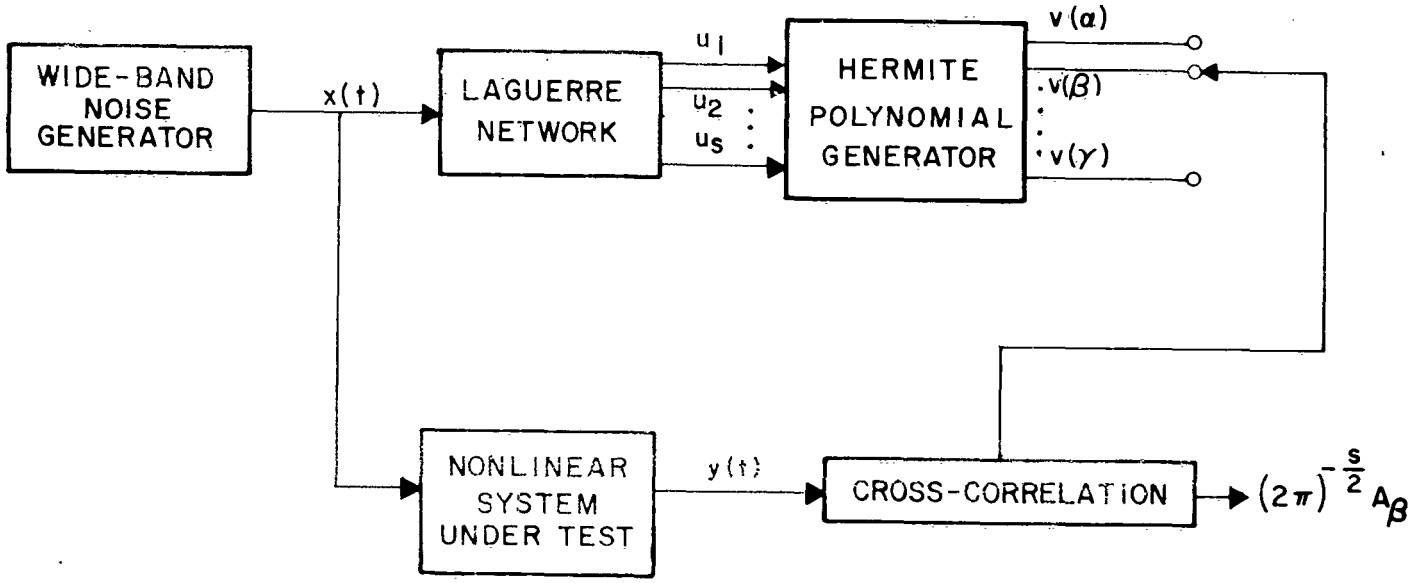

FIG. 3 EXPERIMENTAL ARRANGEMENT REQUIRED FOR THE EVALUATION OF WIENER COEFFICIENTS.

Such analysis would certainly be very difficult since nested limiting operations of high dimensionality are involved in the Wiener theory. In addition to this analytic difficulty, there are serious problems involved in attaining the computing speeds needed to implement the experimental arrangement shown in Figure 3. It appears that a hybrid computer possessing the best features of both analog and digital devices will be required. The extensive literature search undertaken in conjunction with the writing of this report has produced no reference relating an actual experiment of the type indicated by Figure 3.

The difficulties discussed above relate to dimensionality and time of observation. Even if these problems should be resolved, there are other fairly serious weaknesses inherent in the Wiener theory. First of all, there remains the problem of obtaining some relationship between the Wiener coefficients and meaningful system parameters. That is, from an engineering point of view, it would be very desirable (and in many circumstances essential) to invert the Wiener coefficients to obtain the parameters of the system differential equation. There is no evidence that this can be done. Secondly, the Wiener theory provides a basis only for a "bench test" type of experimental system identification somewhat analogous to frequency testing for linear systems. The experimental technique requires that the system under test be disconnected from its normal operating signals and subjected to a special test signal over a long period of time and under rather ideal conditions of observation. This is not possible in many identification problems of practical importance. Finally, in addition to all the other stumbling blocks, there

remains the fact that the Wiener theory is not completely general. It excludes the very important classes of time varying and unstable systems from consideration.

Regardless of these shortcomings, the Wiener theory appears to stand as the only general experimental technique for the identification of nonlinear systems which has been proposed up to the present time. Whether or not the theory is ever actually used in its present form, it serves a very useful purpose in providing a new conceptual basis for the experimental aspects of nonlinear theory. The practical problems associated with the theory are certainly worthy of further attention.

4. PARAMETER SPACE METHODS

4.1 Introduction

The Wiener theory of nonlinear systems is applicable in situations where a complete state of ignorance exists concerning the nature of the nonlinear system under test. This is rarely the case in practice. As a general rule, systems or devices subjected to experimental tests are governed by physical principles which are reasonably well understood. The uncertainty in such tests usually relates to the magnitude of various effects rather than to the basic mechanisms operating to produce the observed data. Even when this is not the case, the investigator is at least able to suggest several competing theories to explain the observed phenomena. Under these circumstances, it is possible to construct a <u>finite</u> parameter model for the system under test rather than an <u>infinite</u> parameter model as proposed by Wiener. This approach will be called a

"parameter space method" in this report to distinguish it from the function space methods used by Wiener and others. Parameter space may be viewed as a generalization of the familiar "phase" or "state" space employed in mechanics for the description of the state of a physical system.

When the object under test is a dynamic system, a very natural finite parameter characterization may be obtained by utilizing the system differential equation. The previously mentioned damped pendulum equation provides a simple example. Referring to equation 1, the pendulum angle, Θ, is governed by

$$I\ddot{\Theta} + B\dot{\Theta} + Mg\ell \sin \Theta = 0 \tag{42}$$

which can be normalized to

$$c_2\ddot{\Theta} + c_1\dot{\Theta} + \sin \Theta = 0 \tag{43}$$

The determination of c_2 and c_1 along with two initial conditions provides a complete characterization of the system. This report is basically concerned with the formulation of an approach to permit the inference of parameters of this type from unreliable records of the input and output of a system under test.

As mentioned at the beginning of this report, the idea of representing a system by a differential equation is scarcely a new concept. Indeed it would seem to be somewhat paradoxical to advocate a return to differential equation models in light of the remarks made previously. The explanation for this apparent regression lies in the emergence of electronic computers as a revolutionary force in scientific and engineering methodology.

While it is indeed difficult to solve a nonlinear differential equation such as equation 43 by analysis, it is quite simple to obtain a solution by electronic computation. In fact, electronic analog computers are specifically constructed for the purpose of solving high order sets of nonlinear differential equations and are poorly suited to any other task. Graphical and various approximate methods which have been found extremely valuable by human investigators are of little or no value to a computer. In recognition of this fact, the parameter space methods to be described in the remainder of this report are entirely computer based. Both the determination of the system parameters and the evaluation of the resulting response will be accomplished by completely automatic methods.

4.2 An Abstract Comparison of Function Space and Parameter Space Characterizations

In the Wiener theory, the input to a system is regarded as a trajectory in a Hilbert space. Time appears parametrically along this curve. The output at a particular instant is obtained by the application of a transformation from the appropriate point in the Hilbert space to the real line. This transformation is completely described by the infinite set of Wiener coefficients, $\{a_{ij\cdots h}\}$. Figure 4 shows this situation graphically. In this figure, $T_{\{A\}}$ stands for the transformation given by equation 39. The distinctive feature of the expansions chosen by Wiener is that when the trajectory in the input space is produced by a wide-band Gaussian random process, the transformation from the input space to the output space is uniquely determined and simply (conceptually) computed from simultaneous observation of the input and output trajectories.

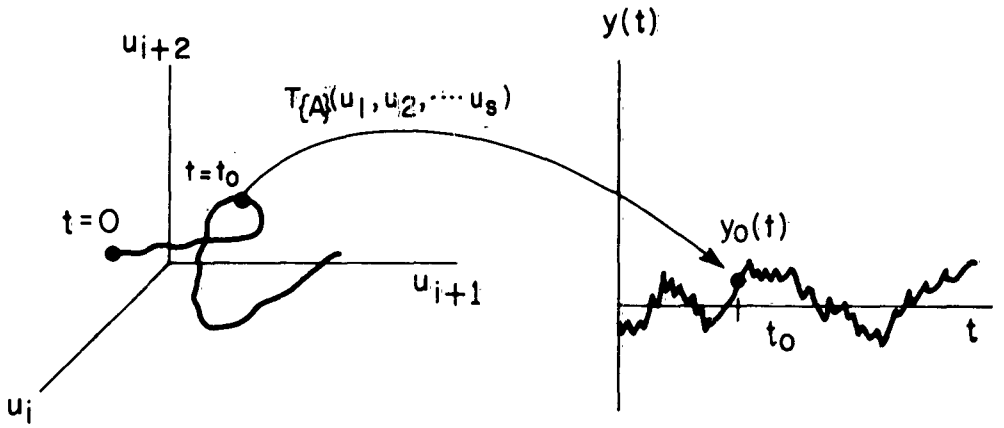

FIG.4 DESCRIPTION OF A TRANSFORMATION IN TERMS OF WIENER COEFFICIENTS.

In contrast to the Wiener theory, in the parameter space method uncertainties relating to both the input and the system (including initial conditions) are described by a finite number of parameters. Moreover, due to the uniqueness properties of solutions to specified differential equations, the output variable is completely determined for all time from the initial point in parameter space. Thus, solution of the differential equation is equivalent to performing a transformation, T_o, from a point, $\vec{c}$, in a finite dimentional vector space to a point in function space. Figure 5 illustrates this relationship [6]. Because the transformation, $T_o(\vec{c})$, operates in a space of finite dimensionality, deducing the system description is no longer a problem in functional analysis. Rather, in the formulation to be employed in this research program, the system differential equation is found by utilizing the techniques of nonlinear programming. This approach will be explained in detail in subsequent reports.

4.3 Choice of a Metric for the Output Function Space

In order to permit iterative techniques to be employed in the estimation of parameter vectors, it is necessary to define a distance function or metric which measures the distance between two functions. For the purpose of this investigation, the metric chosen is the conventional "Euclidian" or L_2 metric

$$\rho^2(y_1, y_2) = \int_a^b [y_1(t) - y_2(t)]^2 \, dt \qquad (44)$$

---

6 It is possible that the input signal, x(t), may be a random process. In that event, for identification processes taking place in real time, only past values of x(t) will be known so that the output, y(t), can be determined only up to the present time. In such circumstances, the output can be thought of as a trajectory in a space of coordinates associated with a semi-infinite "backward looking" expansion rather than as a single point derived from an expansion over $[-\infty, +\infty]$ or $[0, +\infty]$.

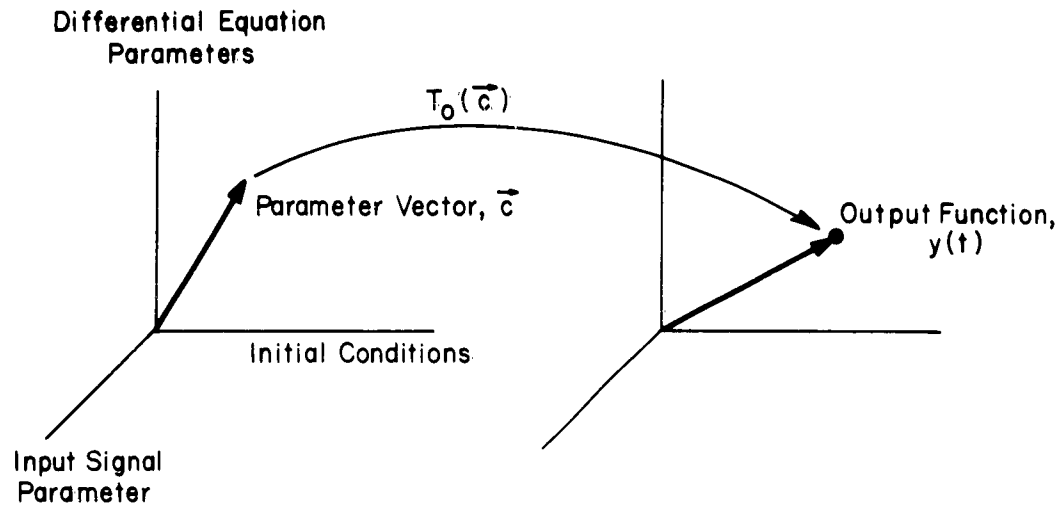

FIG. 5 PARAMETER SPACE DESCRIPTION OF A NONLINEAR SYSTEM.

With this metric, a complete linear function space is commonly called a "Hilbert space". Note that while ρ is a functional in the output space, it is an ordinary function of the parameter space coordinates. This is of the utmost importance so far as the computational aspects of system identification are concerned.

One reason for choosing a Euclidian metric is that a considerable body of knowledge exists relating to Hilbert spaces. However, an even more important reason from the point of view of this research program is that the L_2 metric yields a set of linear simultaneous equations in the process of iteratively deducing the system parameters from response data. It will be shown in subsequent reports that this is the only metric which has this property. Since nonlinear simultaneous equations are very difficult to solve even by computer methods, this feature of the L_2 metric makes its use almost mandatory in some of the computational procedures to be described in the sequel to this report.

4.4 Experimental Evaluation of Parameter Space Coordinates

The parameter space method begins with the specification of a finite dimensional parametric model for the system under test. For example, the behavior of the system described by equation 43 is completely determined by the specification of a four dimensional parameter vector, $\vec{c}$:

$$\vec{c} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ \theta(0) \\ \dot{\theta}(0) \end{pmatrix} \tag{45}$$

In the event that a forcing function with certain unknown parameters were present, the dimensionality of this vector would be increased. Having established a model, experimental observation of the system output may begin. When a sufficient length of record has been obtained, say $0 \leq t \leq T$, the identification or inference process described in the following paragraphs may be initiated.

For the most general sort of nonlinear system, the parameter space identification can be accomplished by an iteration procedure utilizing a metric in the output function space. The process begins with an initial guess for each of the coordinates of the parameter vector, $\vec{c}$. This guess may be generated either by a computer as part of the inference process or it may represent the best estimate of a human investigator. Let this vector be denoted by $\vec{c}_1$. Furthermore, let the true parameter vector be denoted $\vec{c}_o$ and the observed system response be $y_o(t)$. Associated with the vector $\vec{c}_1$ will be another response, $y_1(t)$, which can be determined by a computer. When this has been accomplished, the distance between the two functions can be computed by evaluating the integral:

$$\rho^2(y_o, y_1) = \int_o^T [y_o(t) - y_1(t)]^2 \, dt = \phi(\vec{c}_1; \vec{y}_o) \tag{46}$$

This notation emphasizes the fact that while ρ is a functional in the output space, it can be equated to a simple function, ϕ, defined over the parameter space for a given set of data, y_o. Such a function is usually called a "criterion" or "objective" function in the terminology of mathematical programming (43,44).

Since ϕ can be evaluated for every $\vec{c}$, partial derivatives can be determined in the parameter space yielding a gradient:

$$\vec{\nabla}\phi = \frac{\partial \phi}{\partial \vec{c}} = \begin{pmatrix} \frac{\partial \phi}{\partial c_1} \\ \frac{\partial \phi}{\partial c_2} \\ \vdots \\ \frac{\partial \phi}{\partial c_n} \end{pmatrix}_{c = \vec{c}_1} \tag{47}$$

This gradient can serve as a guide in choosing a new set of parameters

$$\vec{c}_2 = \vec{c}_1 + \vec{\Delta c} \tag{48}$$

such that

$$\phi(\vec{c}_2 \; ; \; y_o) < \phi(\vec{c}_1 \; ; \; y_o) \tag{49}$$

Since for an arbitrary parameter vector, $\vec{c}$

$$\phi(\vec{c} \; ; \; y_o) = \rho^2(y, \; y_o) \geq 0 \tag{50}$$

it follows that when equation 49 is satisfied at every stage of an iteration, the resulting sequence, $\{\phi_1, \phi_2, \cdots \phi_k \cdots\}$ must converge to some limiting value, say ϕ_e. Under ordinary circumstances, the corresponding sequence in parameter space will also converge to a limiting value, say $\vec{c} = \vec{c}_e$. When this occurs, the limit vector, $\vec{c}_e$, provides an "estimate" of the true parameter vector, $\vec{c}_o$. Specific algorithms for obtaining the convergence specified by equation 49 are being developed as part of the continuation of this research program.

Figure 6 summarizes the steps involved in the experimental determination of the parameter space coordinates of an unknown physical system. In the actual implementation of such an experiment, all operations could be performed by a general purpose digital computer with analog inputs. However, in many circumstances it would be preferable to construct the block labeled "computer model" on an analog computer under the control of the digital computer.

4.5 Local Minima and Non-uniqueness

The conceptual and experimental simplicity of parameter space methods is not achieved without penalty. Aside from the fact that the method requires considerable apriori knowledge concerning the system under test, there are certain computational and mathematical difficulties involved. First of all, equations 49 and 50 do not guarantee that $\phi(\vec{c}_e ; y_o) = 0$. It may turn out that $\vec{c} = \vec{c}_e$ represents merely a stationary point for the function ϕ in the parameter space. That is, the computed response, $y_e(t)$, may represent the data very poorly even though a small change in any parameter makes the fit even poorer. When this is the case, correct identification of the system under test requires that a more sophisticated search be carried out in the parameter space to locate the absolute minimum of $\phi(c_1 ; y_o)$ rather than a simple stationary point. This problem will also be explored as part of this research program.

In addition to the local minimum problem, there is a more fundamental difficulty associated with parameter space methods. Since it is not assumed that a special test signal can be applied to the system to be identified,

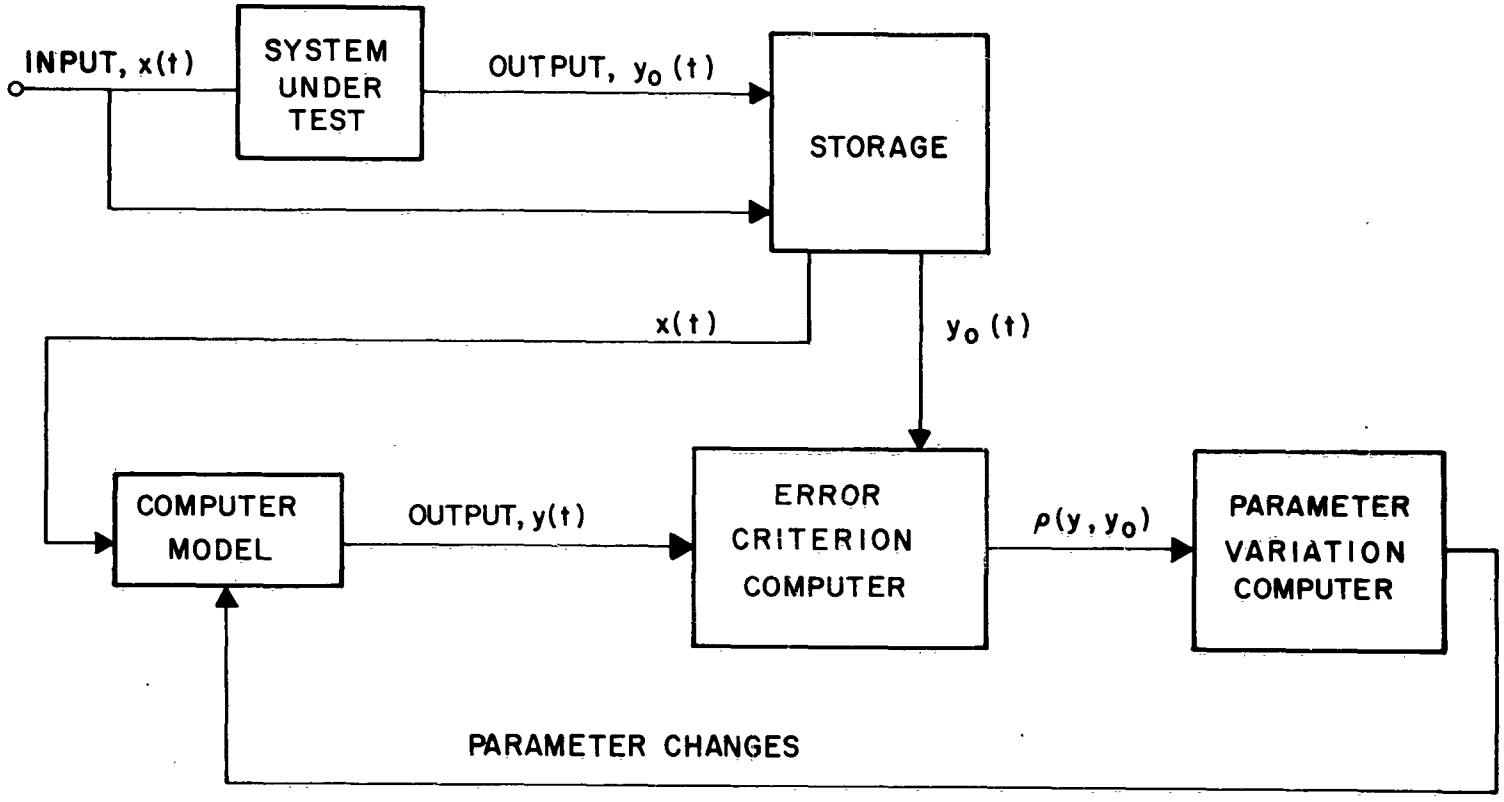

FIG. 6 EXPERIMENTAL DETERMINATION OF THE PARAMETERS OF A DYNAMIC SYSTEM.

there is no guarantee that the mapping shown in Figure 5 is one-to-one. It may, in fact, be many-to-one. A single example suffices to show this. Suppose that the system under test is an unforced linear time invariant system with an observed time response resulting entirely from unknown initial conditions. In such a situation, it is possible to select initial conditions which excite only one normal mode of the system. This being the case, all systems possessing this particular mode are indistinguishable from one another when so excited.

Abstractly, the non-uniqueness property of parameter space methods may be said to result from the fact that the metric used for iteration is in the wrong space. Let the ordinary Euclidian distance associated with the parameter vector space be denoted by

$$\rho_p^2 (\vec{c}_o , \vec{c}_k) = \sum_{i=1}^{n} (c_{oi} - c_{ki})^2 \tag{51}$$

Then what is really desired is a computational algorithm which will guarantee

$$\rho_p (\vec{c}_o , \vec{c}_{k+1}) < \rho_p (\vec{c}_o , c_k) \tag{52}$$

Unfortunately, $\rho_p (\vec{c}_o , \vec{c}_k)$ cannot be computed from an observation of the response functions $y_o(t)$ and $y_k(t)$ associated with $\vec{c}_o$ and $\vec{c}_k$ respectively. Instead, it is necessary to base iteration upon the computable function space distance, $\rho(y_k , y_o)$, given by equation 46. Now in the absence of measurement error, if equation 43, for example, is indeed an exact description of the system under test, it does follow that $\rho(y_o, y_k) \to 0$

is a <u>necessary</u> condition for $\rho_p(c_o , c_k) \longrightarrow 0$; i.e., for $\vec{c}_k \longrightarrow \vec{c}_o$. However, as the example discussed previously shows, $\rho(y_o , y_k) \longrightarrow 0$ is not <u>sufficient</u> to ensure that $\vec{c}_k \longrightarrow \vec{c}_o$. In order to make $\rho(y_o , y_k) \longrightarrow 0$ a sufficient condition, it is necessary to restrict the region of parameter space to be explored to a subspace S in which the mapping is one-to-one. The determination of suitable subspaces by purely mathematical techniques is apt to be extremely difficult in the majority of situations. It seems more likely that such regions would be determined by a preliminary computer investigation of the properties of the particular parametric model under investigation.

4.6 Effects of Noise, Measurement Errors and Imprecise Models

For any one of a number of reasons, the computer solution for $y(t)$ may fail to correspond exactly to the measured response data even when the correct values of the system parameters $(\vec{c} = \vec{c}_o)$ are used in the calculation. Among the major contributors to this situation will be measurement errors, random noise internally generated by the system under test, and the use of a computer model which ignores some of the more subtle effects influencing the actual physical system. In such circumstances, the minimum value of the distance function $\phi(\vec{c} ; y_o)$ over the permitted subspace S will be greater than zero and the parameter vector, $\vec{c}_e$, associated with the minimum will represent a "least squares estimate" of the true parameter value, $\vec{c}_o$. The optimal quality of such estimates may be shown in a varity of circumstances (34,45).

4.7 An Alternate Criterion Function for the Evaluation of Parameter Space Coordinates

The distance function $\phi(\vec{c}\ ;\ y_o)$ is computed by an implicit procedure requiring solution of the assumed system differential equation. It is possible to use another distance which can be defined explicitly in a parameter space. Suppose that the experimental observation of a system to be identified is carried out in such a way that values are obtained for all derivatives up to the n^{th}, the order of the systems. Furthermore, suppose that the system description can be written in the form

$$F(\vec{\alpha}\ ,\ \vec{y}) = 0 \tag{53}$$

where $\vec{y}$ is the conventional phase space vector, $\vec{y} = (y,\ \dot{y},\ \ddot{y},\ \cdots\ y^{(n-1)})$ and $\vec{\alpha}$ represents the m remaining parameter space coordinates (to be determined). Thus

$$\vec{c} = \begin{pmatrix} \vec{\alpha} \\ \vec{0} \end{pmatrix} + \begin{pmatrix} \vec{0} \\ \vec{y} \end{pmatrix} \tag{54}$$

where $\vec{y}$ is an n dimensional vector and $\vec{\alpha}$ is a vector with dimensionality m. If $\vec{y}$ can indeed be measured without error, then equation 53 must hold for the observed data; i.e.

$$F(\vec{\alpha}_o\ ,\ \vec{y}_o) = 0 \tag{55}$$

This equation must be satisfied for all values of time in $(0,\ T)$. Utilizing this fact, an objective or criterion function, $\phi(\vec{\alpha}\ ;\ \vec{y}_o)$ may be defined over the $\vec{\alpha}$ space for a given experimental record:

$$\phi(\vec{\alpha} ; \vec{y}_o) = \int_o^T F^2 \left(\vec{\alpha} , \vec{y}_o(t)\right) dt \geq 0 \qquad (56)$$

In light of equation 53,

$$\phi(\vec{\alpha}_o ; \vec{y}_o) = 0 \qquad (57)$$

This relationship may be used to calculate $\vec{\alpha}_o$ by iterative methods in exactly the same way that the function $\phi(\vec{c} ; y_o)$ is used.

In an actual physical experiment, it will usually be extremely difficult to directly measure all of the system derivatives. On the other hand, attempts to differentiate experimental data several times are likely to fail due to the inevitable presence of small amounts of system noise and measurement error. As is well known, these effects are accentuated by differentiation and may even lead to derivatives which are unbounded in a mean square sense. Most likely, statistical estimation of the necessary derivatives will be required to obtain values for substitution in equation 56. Whatever method is used to obtain derivatives, in the practical situation the various sources of error will prevent equation 57 from holding true. Rather, as with the function ϕ , there will be an estimator, $\vec{\alpha}_e$, which possesses the property

$$\min_{\vec{\alpha}} \phi(\vec{\alpha} , \vec{y}_o) = \phi(\vec{\alpha}_e , \vec{y}_o) \qquad (58)$$

The vector $\vec{\alpha}_e$ is a least squares estimate of $\vec{\alpha}_o$ in a sense somewhat different from the estimate obtained using the criterion function ϕ . An analog computer mechanization of equation 58 has been described by Y. Kaya and S. Yamamura in a paper dealing with the identification of linear

systems (46). Kaya and Yamamura used identical filtering on all derivatives to obtain a relation derived from equation 53 which would be satisfied exactly in the absence of noise. The statistical optimality of their method was not discussed.

The choice between ϕ or $\emptyset$ as a criterion function appears to depend to a large extent upon the relative difficulty of obtaining a sufficient number of derivatives of the system output as compared to solving the assumed system differential equation with a computer. Clearly, attempts to minimize ϕ by iterative nonlinear programming methods will encounter the same difficulties as occur in the minimization of $\emptyset$. There is one special situation however, in which the function ϕ appears to offer significant advantages. When the differential equation expressed by equation 53 is linear <u>in the parameters</u> $\vec{\alpha}$, then according to equation 56, ϕ is a quadratic form in the parameter space. Moreover, ϕ is positive definite and therefore has a unique stationary point which is the minimum existing at $\vec{\alpha} = \vec{\alpha}_e$, the least squares estimate of $\vec{\alpha}_o$. As a consequence, the non-uniqueness and local minimum problems associated with the general nonlinear system identification problem do not occur in this case. The nonlinear system described by equation 43 is linear in the parameters c_1 and c_2 and so furnishes an instance in which the above statements apply.

It is worth noting at this point that the conventional way of employing equation 53 when $y_o(t)$ is an analytic function (rather than an experimental function) involves substituting y_o and its derivatives into

equation 53 at m different values of time. This results in m simultaneous equations, generally nonlinear, which can be solved for the m unknown parameters. While this approach has also been suggested as an experimental procedure, it does not appear to be appropriate due to the difficulties associated with differentiation of experimental response data (47,48,49).

4.8 Parameter Tracking Servomechanisms

If the minimization of a criterion function is undertaken in real time concurrently with the unfolding of a real physical process, then it is possible to devise computational procedures which permit parameter tracking for time variable processes (49,50). This approach permits a considerably simpler model to be used for the short time description of complicated systems. In most of the work carried out thus far in this connection, analytic methods, (i.e., analog computer methods) rather than algebraic methods of tracking have been used. As a consequence of this somewhat unnatural restriction, severe stability problems have been encountered in attempting to carry out actual experiments. It has been necessary to use functions of error rather than functionals in most cases to achieve stable parameter determining loops. For example, M. Margolis found in the investigation of a simple linear first order system that the parameter determining loop was necessarily unstable when an integral squared error criterion was used (50). In contrast to this situation, it is not difficult to devise algebraic methods for parameter estimation which are always stable. The development and experimental evaluation of such techniques is being pursued in the continuation of this research effort.

The parameter space methods described in this report are of course, entirely applicable to time varying systems. The variation of parameters may be accounted for either explicitly by increasing the dimensionality of the parameter space or implicitly by evaluating the selected criterion function over a sliding time interval.

5. SUMMARY AND CONCLUSIONS

The parameter space approach is believed to represent a new point of view regarding the general problem of system identification. The method is restricted only by the requirement that a parametric model of finite dimensionality must be provided for the system under test. In contrast to the Wiener approach, the parameter space method involves the determination of transformations defined over a finite dimensional vector space rather than over a Hilbert space. The computational consequences of this difference are of fundamental importance; by adopting the parameter space point of view, a problem in functional analysis is reduced to a nonlinear programming problem. In addition to the computational advantages gained by a parameter space approach, there is the further advantage that no special test signals are required as in the Wiener theory. The model for a nonlinear system can be inferred from normal operating records.

The current phase of this research is directed toward the development of explicit computational procedures for carrying out the proposed minimization of error functionals by parameter space methods. This work will provide the basis for future reports.

REFERENCES

1. Stout, T. M., "Mathematical Models for Computer Control Systems", Proceedings of the First International Congress of the International Federation of Automatic Control, Moscow, 1960, Butterworths, London, 1961, pp. 991-997.

2. Zadeh, L. A., "On the Identification Problem", IRE Transactions on Circuit Theory, Vol. CT-3 No. 4, pp. 277-281 (December, 1956).

3. Bose, A. G., A Theory of Nonlinear Systems, TR 309, Massachusetts Institute of Technology, Cambridge, Mass. (May 15, 1956).

4. Eykhoff, P., Process Parameter Estimation, Technological University, Electronics Laboratory, Delft, Netherlands.

5. Nyquist, H., "Regeneration Theory", Bell System Tech. Jour., XI, pp. 126-147 (1932).

6. James, H. M., Nichols, N. B., and Phillips, R. S., Theory of Servomechanisms, Rad. Lab. Series, Vol. 25, McGraw-Hill Book Co., Inc., New York, 1947

7. Fuchs, A. M., "A Bibliography of the Frequency Response Method as Applied to Automatic Feedback Control Systems", Trans. of the ASME, Vol. 76, pp. 1185-1194 (November, 1954).

8. Ehret, R. J., et al., "An Automatic Transfer Function Measuring and Recording System", AIEE Transactions, Vol. 72, Part I, pp. 664-669 (November, 1953).

9. Leonov, Yu P. and Lipatov, L. N., "The Use of Statistical Methods for Determining the Characteristics of Objects", Automation and Remote Control (English Translation), Vol. 20, No. 9, pp. 1254-1258 (September, 1959).

10. Kochenburger, R. J., "A Frequency Response Method for Analyzing and Synthesizing Contactor Servomechanisms", AIEE Transactions, Vol. 69, Part I, pp. 270-284 (February, 1950).

11. Truxal, J. G., Control System Synthesis, McGraw-Hill Book Co., Inc., New York, 1955.

12. Dudnikov, E. E., "Determination of Transfer Function Coefficients of a Linear System from the Initial Portion of an Experimentally Obtained Amplitude Phase Characteristic", Automation and Remote Control (English Translation), Vol. 20, No. 9, pp. 552-558 (May, 1959).

13. Levy, E. C., "Complex Curve Fitting", IRE Transactions on Automatic Control, Vol. AC-4, No. 1, pp. 37-43 (May, 1959).

14. Kardashov, A. A., and Karniushin, L. V., "Determination of System Parameters from Experimental Frequency Characteristics", Automation and Remote Control (English Translation), Vol 19, No. 4, pp. 327-338 (April, 1958).

15. Brussolo, J. A., "Pole Determinations with Complex Zero Inputs", IRE Transactions on Automatic Control, Vol. AC-4, No. 2, pp. 150-166 (November, 1959).

16. Darovskikh, L. N., "Experimental Determination of Automatic Control Systems Links' Transfer Functions by Means of Standard Electronic Models", Automation and Remote Control (English Translation), Vol 20, No. 9, pp. 1180-1187 (September, 1959).

17. Lendaris, G. G., and Smith, O.J.M., "Complex Zero Signal Generator for Rapid System Testing", AIEE Transactions, Vol. 77, Part II, pp. 534-539 (January, 1959).

18. Huber, E. A., "A Technique for the Adaptive Control of High Order Systems", IRE Transactions on Automatic Control, Vol. AC-7, No. 3, pp. 22-29 (April, 1962).

19. Laning, J. H., and Battin, R. H., Random Processes in Automatic Control, McGraw-Hill Book Co., Inc., New York, 1956.

20. Mishkin, E., and Haddad, R. A., "Identification and Command Problems in Adaptive Systems", IRE Transactions on Automatic Control, Vol. AC-4, No. 2, pp. 121-131 (November, 1959).

21. Wiesner, J. B., and Lee, Y. W., "Experimental Determination of System Functions by the Method of Correlation" presented at the IRE National Convention, New York, N. Y., March 1950.

22. Aseltine, J. A., et al., "A Self-Adjusting System for Optimum Dynamic Performance", IRE National Convention Record, 1958, Part 4, pp. 182-190.

23. Turin, G. L., "On the Estimation in the Presence of Noise of the Impulse Response of a Random, Linear Filter", IRE Transactions on Information Theory, Vol. IT-3, No. 1, pp. 5-10 (March, 1957).

24. Lichtenberger, W. W., "A Technique of Linear System Identification Using Correlating Filters", IRE Transactions on Automatic Control, Vol. AC-6, No. 2, pp. 183-199 (May, 1961).

25. McGrath, R. J. and Rideout, V. C., "A Simulator Study of a Two Parameter Adaptive System", IRE Transactions on Automatic Control, Vol. AC-6, No. 1, pp. 35-42 (February, 1961).

26. Smith, K. C., "Adaptive Control Through Sinusoidal Response", IRE Transactions on Automatic Control, Vol. AC-7, No. 2, pp. 129-139 (March, 1962).

27. Weygandt, C. N., and Puri, N. N., "Transfer-Function Tracking and Adaptive Control Systems", IRE Transactions on Automatic Control, Vol. AC-6, No. 2, pp. 162-166 (May, 1961).

28. Eykhoff, P., and Smith, O.J.M., "Optimalizing Control with Process Dynamics Identification", IRE Transactions on Automatic Control, Vol. AC-7, No. 2, pp. 140-155 (March, 1962).

29. Goodman, T. P., and Reswick, J. B., "Determination of System Characteristics from Normal Operating Records", Trans. of the ASME, Vol. 78, pp. 259-271 (February, 1956).

30. Kalman, R. E., "Design of a Self Optimizing Control System", Trans. of the ASME, Vol. 80, pp. 468-478 (February, 1958).

31. Joseph, P., Lewis, J., and Tou, J., "Plant Identification in the Presence of Disturbances and Application to Digital Adaptive Systems", AIEE Trans. Vol. 80, Part II, pp. 18-24 (March 1961).

32. Kerr, R. B., and Surber, W. H., "Precision of Impulse-Response Identification based on Short Normal Operating Records", IRE Transactions on Automatic Control, Vol. AC-6, No. 2, pp. 173-182 (May, 1961).

33. Solodovnikov, V. V., and Uskov, A. S., "A Frequency Method for Determining the Dynamic Characteristics of Objects of Automatic Control from Data on their Normal Usage", Automation and Remote Control (English Translation), Vol. 20, No. 12, pp. 1533-1542 (December, 1959).

34. Levin, M. J., "Optimum Estimation of Impulse Response in the Presence of Noise", IRE Transactions on Circuit Theory, Vol. CT-7, No. 1, pp. 50-56 March, 1960.

35. Gabor, D. "Communication Theory and Cybernetics", IRE Transactions on Circuit Theory, Vol. CT-1, No. 4, pp. 19-31 (December, 1954).

36. Zadeh, L. A., "On the Representation of Nonlinear Operators", IRE Convention Record, 1957, Part II, pp. 105-113.

37. Resa, F. M., An Introduction to Information Theory, McGraw-Hill Book Co., New York, N. Y., 1961, pp. 315-317.

38. Wiener, N., Nonlinear Problems in Random Theory, John Wiley and Sons, Inc., New York, N. Y., 1958.

39. Balakrishnan, A. V., A General Theory of Nonlinear Estimation Problems in Control Systems, Department of Engineering, University of California, Los Angeles, California (November, 1961).

40. Lory, H. J., Lai, D. C., and Huggins, W. H., "On the Use of Growing Harmonic Exponentials to Identify Static Nonlinear Operators", IRE Transactions on Automatic Control, Vol. AC-4, No. 2, pp. 91-100 (November, 1959).

41. Lee, Y. W., Statistical Theory of Communication, John Wiley and Sons, Inc., New York, N. Y., 1960, pp. 473-476 and 487-489.

42. Kitamori, T., "Applications of Orthogonal Functions to the Determination of Process Dynamic Characteristics and to the Construction of Self Optimizing Control Systems", Proceedings of the First International Congress of the International Federation of Automatic Control, Vol. 2, Moscow, 1960, Butterworths, London, 1961, pp. 613-618.

43. Vadja, S., Mathematical Programming, Addison-Wesley Publishing Co., Inc., Reading, Massachusetts, 1961.

44. Bellman, R., Dynamic Programming, Princeton University Press, Princeton, N. J., 1957.

45. Graybill, F. A., An Introduction to Linear Statistical Models, McGraw-Hill Book Co., Inc., New York, N. Y., 1961.

46. Kaya, Y. and Yamamura, S., "A Self Adaptive System with a Variable Parameter PID Controller", AIEE Transactions, Vol. 80, Part II, pp. 378-386 (January, 1962).

47. Staffin, R., "Executive-Controlled Adaptive Systems", AIEE Transactions, Vol. 78, Part II, pp. 523-530 (January, 1960).

48. Braun, L., "On Adaptive Control Processes", IRE Transactions on Automatic Control, Vol. AC-4, No. 2, pp. 30-43 (November, 1959).

49. Potts, T. F., Ornstein, G. N., and Clymer, A. B., "The Automatic Determination of Human and Other System Parameters", Proceedings of the Western Joint Computer Conference, Vol. 19, pp. 645-660 (May, 1961).

50. Margolis, M., On the Theory of Process Adaptive Control Systems, the Learning Model Approach, Report No. 60-32, University of California, Department of Engineering, Los Angeles, California (May, 1960).

PART II: RESEARCH IN OPTICAL COHERENCE

PART II.

RESEARCH IN OPTICAL COHERENCE

A detailed study of certain aspects of the theory of optical coherence has been carried out. A major objective of the study is to obtain criteria for optimizing optical systems employing coherent sources and amplifiers such as lasers. This is the first in a series of studies to investigate the effect of various optical system devices such as antennas, filters, modulators and demodulators, etc., on the coherence properties of optical signals. The present report concerns itself primarily with spatial and temporal filters.

After defining the weighting function and aperture field distribution for a spatial frequency filter, the general transfer function for the spatial filter is defined, with a discussion of the approximations involved. The weighting function or transfer function concept is then extended to include both spatial and temporal variations. The weighting function of the system in this sense is equal to the response of the filter to a point source in space and a unit impulse in time. An image equation is then defined which relates the image field strength (as a function of the spatial coordinates and time) to the object field strength (as a function of the same coordinates) and the weighting function of the filter. A generalized coherence function is then defined as a correlation function involving the spatial separation of two sources in space and the time separation of two points on a signal as well as the space

and time variables themselves. Ensemble, time, and space averages are then defined and it is shown that the dependence of the coherence function on the origin in space and time can be removed by appropriate time and space averaging. This operation yields a coherence function which is a function of the separation in space and time but not the origin in these variables. If the process is stationary in these variables, no space and/or time averaging is needed. A normalized correlation coefficient or correlation function can then be defined. After these general definitions are made, several special cases of interest are considered. In particular, the case where the object field strength is separable into the product of an object field strength in space and an object field strength in time is of interest. The special cases of combinations of perfect time coherence and incoherence and perfect space coherence and incoherence are considered and the coherence function evaluated for these cases. The image equation is written for each of these cases and the image function evaluated. It is pointed out that the problem of temporal partial coherence or incoherence is particularly difficult to treat because the spatial frequency filter transfer function is a function of the wavelength and hence of the temporal incoherence. This means that the transfer function of the filter depends upon the input. Consequently, the general problem is nonlinear and very difficult to treat. Basically, what is needed to handle this problem are the general solutions of Maxwell's equations without the simplifying assumption of a sinusoidal time variation of the driving function. A first approximation to the solution to this problem was

made by assuming that the modulation bandwidth is small compared to the carrier frequency, which is almost always the case in optics. This assumption yields a quasi-stationary approach in which the frequency variations are sufficiently slow so that the response for any given input frequency will be the steady-state response of the filter to that frequency. The overall response is found by summing the responses to each individual frequency.

This initial study on the coherence properties of optical filters will be extended not only to other devices, but will provide a basic statistical description of the processes to be analyzed in the future. In particular, the effect of coherence on information content and rate will be considered. Ultimately, the optimum system will be synthesized using the criterion of maximization of information rate at the desired output of the system. The determination of information rates will be strongly affected by the coherence properties of the signals involved.

The principal effort during this period has concerned itself with the effects of coherence on filtering. In particular, the effect of linear spatial and/or temporal frequency filters on optical waves has been investigated. Optical filters differ from ordinary low frequency filters in that both spatial and temporal frequency effects must be taken into account. Let the weighting function G of a spatial frequency filter, such as an optical antenna, be defined as the response of the filter to a point target in the far field. It is a function of the image coordinates x' and the y', the object coordinates x and y, and of time t. Let

us temporarily assume that the time variation is sinusoidal and determine what the spatial weighting function must be for this case. The antenna spatial weighting function or "spread" function may be related to the aperture field distribution across the spatial filter. It can be shown* that the far field pattern arising from an aperture distribution $F(\xi,\eta)$ with nearly uniform phase across the aperture is given by

$$G(x,y) = \pi jk \frac{e^{-jkR}}{R} g(x,y) \tag{1}$$

where

$$g(x,y) = \frac{1}{(2\pi)^2} \int_A F(\xi,\eta)\, e^{jk(x\xi + y\eta)}\, d\xi\, d\eta \tag{2}$$

and R = distance between the aperture and a far-field point

$k = 2\pi/\lambda$

The object coordinates x and y are

$$x = \sin\Theta \cos\phi$$
$$y = \sin\Theta \sin\phi \tag{2a}$$

All the object dimensions are given in terms of spherical angular coordinates whose origin is coincident with the phase center of the spatial frequency filter. The object field strength is $O(x,y,t)$. Note that the object field strength in general depends not only on the position of the source in space but also on time. The time dependence

---

* Silver, S., "Microwave Antenna Theory and Design", McGraw-Hill, New York, 1949, p. 173, equation 9.

may rise from several sources. In the first place, the energy illuminating the object or the radiation from a passive object will in general be time-varying so that the object field strength will also be time-varying. Or if the object is a source, the source field strength will be time-varying. In addition, the object may be moving so that it will be located at a different position at a later instant of time yielding a time dependence. This dependence is implicit rather than explicit. The image field strength I depends upon the coordinates x' and y' along the image surface as well as on time, where

$$x' = \sin \Theta' \cos \phi'$$
$$y' = \sin \Theta' \sin \phi' \qquad (2b)$$

Note that equation (2) is a Fourier transform, that is, the far field pattern $g(x,y)$ is the Fourier transform of the aperture field distribution $F(\xi,\eta)$. Since x and y are spatial coordinates, the corresponding transform variables $k\xi$ and $k\eta$ have the dimension of spatial frequency. Let us therefore relabel both ξ and η as

$$\xi = \omega_x = 2\pi f_x$$
$$\eta = \omega_y = 2\pi f_y \qquad (3)$$

where f_x and f_y are spatial "frequencies" which are the Fourier transform variables corresponding to x and y, respectively. For the special case of nearly uniform phase across the aperture, the Fourier transform pair relating the far field pattern $G(x,y)$ and the aperture field distribution $F(\xi,\eta)$ is (except for a constant)

$$g(x,y) = \frac{1}{(2\pi)^2} \int_A F(\omega_x, \omega_y)\, e^{jk(x\omega_x + y\omega_y)}\, d\omega_x\, d\omega_y$$

$$F(\omega_x, \omega_y) = \iint_{-\infty}^{\infty} g(x,y)\, e^{-jk(x\omega_x + y\omega_y)}\, dx\, dy \qquad (4)$$

When the simplifying assumption of uniform phase cannot be made, the aperture field distribution $F(\omega_x, \omega_y)$ corresponding to the given far field pattern must be found by solving the integral equation*

$$G(x,y) = \frac{jk}{4\pi} \frac{e^{-jkR}}{R} \int_A F(\omega_x, \omega_y)\, (\cos\Theta + \vec{i}_z \cdot \vec{s})\, \cdot$$

$$e^{j\,k(x\omega_x + y\omega_y)}\, d\omega_x\, d\omega_y \qquad (5)$$

where $\vec{i}_z$ is a unit vector normal to the aperture and the $\vec{s}$ is a unit vector normal to the phase front. Since $G(x,y)$ is the response of the spatial filter to a point source, the filter has an impulse response $G(x,y)$. The transfer function of this filter is then the Fourier transform of $G(x,y)$ or

$$T(u_x,u_y) = \iint_{-\infty}^{\infty} \frac{jk}{4\pi} \frac{e^{-jkR}}{R} \left[\int_A F(\omega_x,\omega_y) \cdot \right.$$

$$\left.(\cos\Theta + \vec{i}_z \cdot \vec{s}) e^{jk(x\omega_x + j\omega_y)}\, d\omega_x\, d\omega_y\right] e^{-j(xu_x + yu_y)}\, dx\, dy \qquad (6)$$

In the special case of nearly uniform phase across the aperture, equation (6) becomes

---

* Ibid, p. 173, equation 8.

$$T(u_x,u_y) = \pi jk \frac{e^{-jkR}}{R} \iint_{-\infty}^{\infty} g(x,y)\, e^{-j(xu_x + yu_y)}\, dx\, dy$$

$$= \pi jk \frac{e^{-jkR}}{R} F(u_x,u_y) \qquad (7)$$

where

$$u_x = k\omega_x$$
$$u_y = k\omega_y \qquad (8)$$

Equation (7) shows that to within a constant the spatial frequency transfer function of the filter is equal to the aperture field distribution over the aperture of the filter, providing the phase is reasonably uniform over the aperture. If the phase differs markedly from being uniform over the aperture, then the aperture field and transfer function are related by the more general equation (6).

In the more general case where the temporal as well as the spatial filtering effect of the filter is taken into account, the overall weighting function of the filter is $G(x,y,t)$. The function $G(x,y,t)$ is equal to the response of the filter to a point source in space and a unit impulse in time, or in other words, the response of the filter is an object specified by

$$O(x,y,t) = \delta(x,y,t) = \delta(x,y)\, \delta(t) \qquad (9)$$

The corresponding spatial and temporal frequency transfer function is given by

$$T(u_x,u_y,\omega) = \iint_{-\infty}^{\infty} dx\, dy \int_{-\infty}^{\infty} dt\, G(x,y,t)\, e^{j(xu_x + yu_y + \omega t)} \tag{10}$$

where

$$\omega = 2\pi f = \frac{2\pi c}{\lambda} = kc \tag{11}$$

This more general weighting function G takes into account both the spatial and temporal filtering properties of the filter. In order to find the response of the filter, or the image field strength, from an extended source or object (designated by $O(x,y,t)$), the object field strength O is convolved with the weighting function, G. Thus, when both spatial and temporal frequency filtering effects are taken into account, the image field strength at the output of a two dimensional linear additive spatial filter is given by

$$I(x',y',t) = \iint_{-l}^{l} dx\, dy \int_{o}^{t} d\alpha\, O(x,y,\alpha)\, G[x'-x,y'-y,t-\alpha] \tag{12}$$

If a 3 dimensional Fourier transform is taken of both sides of equation (12), the transform relationship is given by

$$\mathcal{I}[u_x, u_y, \omega] = \mathcal{O}(u_x, u_y, \omega)\, T(u_x, u_y, \omega) \tag{13}$$

where $\mathcal{I}$ is the Fourier transform of the image field strength and $\mathcal{O}$ is the Fourier transform of the object field strength.

For convenience, let us now restrict outselves to one spatial variable $x(t)$. In general, the object field strength will be random and

will be a function of two independent variables, x and t. The random process $O(x,t)$ will be completely defined by the following $2N$ dimensional probability distribution:

$$p(\underline{O}) = p[O(x_1,t_1),\ O(x_2,t_2),\cdots,O(x_1,t_N);\ O(x_2,t_1),O(x_2,t_2),\cdots,O(x_2,t_N);$$
$$\cdots\cdots;\ O(x_N,t_1),\ O(x_N,t_2),\cdots,\ O(x_N,t_N)] \tag{14}$$

where $\underline{O}$ represents the $2N$ dimensional row matrix. For convenience let

$$O(x_i,t_i) = O_{ij} \tag{15}$$

A complete statistical description of the process is usually obtained in much fewer dimensions than $2N$ as indicated in equation (14), depending on the nature of the process. The general "coherence" function will herein be defined as the correlation function

$$R_o(\Delta x,\ \Delta t,\ x,t) = \overline{O(x,t)\ O(x+\Delta x,\ t+\Delta t)} = \overline{O(x_i,t_j)\ O(x_{i+1},t_{j+1})}$$
$$= \int_{-\infty}^{\infty}\int O(x_i,t_j)\ O(x_{i+1},t_{j+1})\ p(O_{ij},O_{i+1,j+1})\ dO_{ij}\ dO_{i+1,j+1} \tag{16}$$

where

$$t_j = t \qquad\qquad x_i = x$$
$$t_{j+1} = t + \Delta t \qquad\qquad x_{i+1} = x + \Delta x \tag{17}$$

The overscore indicates an ensemble average as indicated in equation (16). Note that the object correlation function defined in equation (16) is a

function not only of the space and time difference Δx and Δt respectively, but also of the origin in space and in time x and t. This is the most general case. If R_o is explicitly a function of x it is said to be space nonstationary, if it is explicitly a function of t it is said to be time nonstationary, if it is explicitly a function of both x and t it is said to be both space and time nonstationary. If the spatial bounds on the object field strength O are $\pm X$, and if the temporal bounds on the object fields time strength are $\pm T$, let us define the ensemble average of a function g as

$$\overline{g[y(x,t),\ x,t)\ ]} = \int_{-\infty}^{\infty} g[y(x,t),\ x,t]\ p\ [y(x,t),\ x,t]\ dy \tag{18}$$

Let us define the time average as

$$\overset{\sim\sim\sim\sim\sim\sim\sim}{g[y(x,t),\ x,t]} = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} g[y(x,t),\ x,t]\ dt \tag{19}$$

Let us define the space average as

$$\overline{g[y(x,t),\ x,t]} = \lim_{X\to\infty} \frac{1}{2X} \int_{-X}^{X} g[y(x,t),\ x,t]\ dx \tag{20}$$

If the process is time stationary, equations (18) and (19) are equal. If the process is space stationary equations (18) and (20) are equal. If the process is both time and space stationary, equations (18), (19) and (20) are all equal to each other. The space-and-time-averaged correlation function, or coherence function, is then

$$R_o(\Delta x,\Delta t) = \overline{R_o(\Delta x,\Delta t,x,t)} \tag{21}$$

If the process is both time-and space-stationary, then the time and space averaging indicated in equation (21) are superfluous, since R_o does not depend explicitly on x and t in this case. The normalized coherence function, or correlation coefficient, is defined as

$$\rho_o(\Delta x,\Delta t,x,t) = \frac{R_o(\Delta x,\Delta t,x,t) - \overline{O}(x_i,t_j)\,\overline{O}(x_{i+1},t_{j+1})}{\sigma_o(x_i,t_j)\,\sigma_o(x_{i+1},t_{j+1})} \tag{22}$$

where

$$\overline{O}(x,t) = \int_{-\infty}^{\infty} O(x,t)\; p\,[\overline{O}(x,t)]\; dO \tag{23}$$

$$\overline{O^2}(x,t) = \int_{-\infty}^{\infty} O^2(x,t)\; p\,[(x,t)]\; dO \tag{24}$$

$$= R_o(\Delta x = 0,\ \Delta t = 0,\ x,\ t)$$

$$\sigma_o^2(x,t) = \overline{O^2}(x,t) - \overline{O^2}(x,t) \tag{25}$$

If the process is time stationary but not space stationary R_o and ρ_o are not explicitly dependent upon t; similarly, they are space stationary if they are not explicitly a function of x.

Let us consider some special cases of interest. In certain cases, the function $O(x,t)$ is separable as follows

$$O(x,t) = O(x)\; O(t) \tag{26}$$

The most common examples of situations satisfying equation (26) are when the object radiation or reradiation is monochromatic or the source is a point source, or both. Thus, for a monochromatic source, equation (26) becomes

$$O(x,t) = O(x) \cos \omega_o t \tag{27}$$

where ω_o is the frequency of the monochromatic radiation. In this case equation (16) becomes

$$R_o(\Delta x, \Delta t, x, t) = O(x_i)\, O(x_{i+1}) \cos \omega_o t_j \cos \omega_o t_{j+1} \tag{28}$$

Note that the cosine function is not affected by the ensemble average since it is nonrandom. Furthermore, $O(x,t)$ is not stationary in time, although it has been assumed that the object is stationary in the space variable. Therefore, a further average must take place in time yielding

$$R_o(\Delta x, \Delta t) = \overline{O(x_i)\, O(x_{i+1})} \cos \omega_o t_j \cos \omega_o t_{j+1}$$

$$= \frac{\overline{O(x_i)\, O(x_{i+1})}}{2} \cos \omega_o \Delta t \tag{29}$$

Note that the coherence function in equation (29) is periodic in τ. Coherence or correlation functions which are periodic in τ are said to be "perfectly time coherent" since a time function is perfectly predictable from a knowledge of its value at any given time. This is true of periodic functions in general, since they may be represented as a Fourier series, therefore their correlation functions are a sum of cosine terms of the form

occurring in equation (29). Any periodic or predictable waveform has zero bandwidth occupancy, in the sense that although there are many frequencies present, the set of spectral frequencies is a countable set of measure zero. In other words, the spectrum consists of delta functions at the various discrete frequencies involved with a net zero measure for the bandwidth occupancy. From henceforth, then, a time function is said to be perfectly time coherent if its coherence function is a periodic function in τ.

Now let us consider the case where the object is a point target in space. Equation (26) then becomes

$$O(x,t) = A\delta(x-x_o)\, O(t) \tag{30}$$

where A is a constant. For convenience assume the process to be time stationary; then the coherence function becomes

$$R_o(\Delta x,\Delta t,x,t) = A^2\delta(x_i-x_o)\,\delta(x_{i+1}-x_o)\,\overline{O(t_j)\,O(t_{j+1})} \tag{31}$$

The portion of equation (31) involving space variables is not random and therefore a space average is needed to remove the space dependence. Thus

$$R_o(\Delta x,\Delta t) = A^2\,\overline{O(t_j)O(t_{j+1})}\lim_{X\to\infty}\frac{1}{2X}\int_{-X}^{X}\delta(x-x_o)\delta(x-x_o+\Delta x)\,dx$$

$$= A^2\,\overline{O(t_j)O(t_{j+1})}\lim_{X\to\infty}\frac{\delta(\Delta x)}{2X} = KA^2\,\overline{O(t_j)O(t_{o+1})} \quad \text{for } \Delta x = 0$$

$$= 0 \quad \text{for } \Delta x \neq 0 \tag{32}$$

where

$$\lim_{X\to\infty}\frac{\delta(\Delta x)}{2X} = K \quad \text{for } \Delta x = 0$$

$$= 0 \quad \text{for } \Delta x \neq 0 \tag{32a}$$

The limiting operation in equation (32a) is indeterminate in that both the numerator and the denominator approach infinity. Assuming that in the limit this ratio approaches a constant K for $\Delta x = 0$, the space averaged coherence function is a constant for $\Delta x = 0$ and is zero for $\Delta x \neq 0$. This is the limiting case of a periodic function of a space variable where the period approaches infinity or the repetition frequency approaches 0. The function is then said to be "perfectly space coherent". In general, if the coherence function is a periodic function of the space separation Δx, it is said to be perfectly coherent in that variable. An example of a case where the process is both space and time coherent is

$$O(x,t) = A\bar{o}(x) \cos \omega_o t \tag{33}$$

In this case, the coherence function is:

$$R_o(\Delta x = 0, \Delta t) = \frac{KA^2}{2} \cos \omega_o \Delta t \tag{34}$$

Note that equation (34) is periodic in both the space and the time variable and therefore is both space-and time-coherent.

We have now considered a case where the variables may be perfectly space coherent, perfectly time coherent or both. Let us now consider the opposite extreme where the functions may be perfectly space incoherent, perfectly time incoherent or both. Let us again assume that the object is separable in the sense of equation (26). If the object is perfectly

space incoherent, the correlation between x_1 and x_2 is zero except when x_1 equals x_2. A similar argument holds for perfect time incoherence. Thus, the coherence function for a space-incoherent object is

$$R_o(\Delta x, \Delta t) = K_1 \delta(\Delta x) \overline{O(t_1)O(t_2)} \tag{35}$$

In equation (35), K_1 is proportional to the spectral density of $O(x)$. Similarly, for a temporally incoherent source, the coherence function becomes

$$R_o(\Delta x, \Delta t) = \overline{O(x_1)O(x_2)}\, K_2 \delta(\Delta t) \tag{36}$$

where K_2 is the temporal spectral density. If the source, or object, is both spatially and temporally incoherent, the coherence function becomes

$$R_o(\Delta x, \Delta t) = K_1 K_2\, \delta(\Delta x)\delta(\Delta t) \tag{37}$$

Thus, if the function is space incoherent it will involve a delta function with argument Δx; if it is temporally incoherent, it will involve a delta function with argument Δt and if it is both spatially and temporally incoherent it will involve the product of the two delta functions. In the most general case defined by equation (16), the coherence function will be neither periodic nor involve a delta function in either of the variables; under these circumstances the object is said to be partially coherent.

It is important to consider the coherence properties of the image of a spatial and/or temporal frequency filter under these circumstances. First let us consider the case where the object is perfectly space coherent

and time coherent. Then, substituting equation (33) in equation (12) and limiting the equation to one space variable only,

$$I(x',t) = \int_{-1}^{1} dx \int_{o}^{t} d\alpha\, O(x,\alpha)\, G(x'-x,t-\alpha)$$

$$= \int_{-1}^{1} dx \int_{o}^{t} d\alpha \delta(x) \cos \omega_o \alpha\, G(x'-x,t-\alpha)$$

$$= \int_{o}^{t} G(x',t-\alpha) \cos \omega_o \alpha\, d\alpha \qquad (38)$$

In a device which is intended to be a spatial frequency filter only, such as an antenna, it is common to try to make the device as independent of temporal variations as possible. If this is the case, the weighting function of $G(x,t)$ is approximately independent of t ; and consequently equation (38) becomes

$$I(x',t) = \int_{-1}^{1} dx\, \delta(x)\, G(x'-x) \cos \omega_o t\, dx = G(x') \cos \omega_o t \qquad (39)$$

The image coherence function for equation (39) is

$$R_I(\Delta x',\Delta t,x',t) = \overline{I(x',t)\, I(x'+\Delta x',t+\Delta t)}$$

$$= G(x')\, G(x'+\Delta x') \cos \omega_o t \cos \omega_o (t+\Delta t) \qquad (40)$$

Averaging with respect to both space and time,

$$R_I(\Delta x',\Delta t) = \frac{\overline{G(x')\, G(x'+\Delta x')}}{2} \cos \omega_o \Delta t \qquad (41)$$

Since the time-averaged coherence function in equation (41) is periodic in Δt ; the image is perfectly time coherent. The

coherence function will, in general, be space coherent in spite of the bandwidth limiting effect of the spatial-frequency filter, since the spatial signal response is known for all x' and is not random. In the case of the image represented by equation (38), the temporal response will be perfectly time coherent but its coherence function will not be periodic with finite period until the steady state is reached. When the transient terms have died out, then the steady-state response of a linear temporal filter to a cosine wave is another cosine wave changed in amplitude and shifted in phase. The steady-state is, of course, temporally coherent. Spatial coherence should not be expected in the sense that the spatial coherence function will be periodic with finite period or constant because there is, in effect, no equivalent steady-state phenomenon in the spatial domain. The reason for this is that the bounds of integration on the spatial domain are from -1 to $+1$ and the concept of steady-state loses its significance. Let us assume the special case where the weighting function is separable into the product of a space weighting function and a time-weighting function, or

$$G(x,t) = G_x(x)\, G_t(t) \tag{42}$$

For a spatial filter, every effort is made to make $G_t(t)$ equal to one and likewise in a temporal frequency filter every effort is made to make $G_x(x)$ one. The latter problem seldom comes up because temporal filters are usually used on voltages which are pure time functions. Now let us consider the image function for the case of pure spatial coherence only. Then,

$$\begin{aligned} I(x',t) &= \int_{-1}^{1} \delta(x)\, G_x(x'-x)\, dx \int_{o}^{t} d\alpha\, O(\alpha)\, G_t(t-\alpha) \\ &= G_x(x') \int_{o}^{t} d\alpha\, O(\alpha)\, G_t(t-\alpha) \end{aligned} \tag{43}$$

The coherence function for this case is

$$R_I(\Delta x', \Delta t, x', t) = G_x(x')\, G_x(x'+\Delta x')$$

$$\int_o^{t_1} d\delta \int_o^{t_2} d\beta\, \overline{O(t_1-\alpha)\, O(t_2-\beta)}\, G_t(\alpha)\, G_t(\beta)$$

$$= G_x(x')\, G_x(x'+\Delta x) \int_o^{t_1} d\alpha \int_o^{t_2} d\beta\, R_o(\Delta t+\alpha-\beta, t)\, G_t(\alpha)\, G_t(\omega) \qquad (44)$$

In general, the image in this case will be coherent in neither space nor time. Now let us consider the coherence properties of the image when the object is coherent in time but not in space.

$$I(x', t) = \int_{-1}^{1} O(x)\, G_x(x'-x)\, dx \int_o^t \cos\, \omega_o(t-\alpha)\, G_t(\alpha)\, d\alpha \qquad (45)$$

The image coherence function in this case becomes

$$R_I(\Delta x', \Delta t, x', t) = \iint_{-1}^{1} \overline{O(x)O(y)}\, G_x(x'-x)\, G_x(x'+\Delta x'-y)\, dx\, dy \cdot$$

$$\int_o^{t_1} d\alpha \int_o^{t_2} d\beta \cos\, \omega_o(t_1-\alpha) \cos\, \omega_o(t_2-\beta)\, G_t(\alpha)\, G_t(\beta)$$

$$= \iint_{-1}^{1} R_o(x-y)\, G_x(x'+\Delta x-y)\, G_x(x'-x)\, dx\, dy \cdot$$

$$\frac{1}{2} \int_o^{t_1} d\alpha \int_o^{t_2} d\beta [\cos\, \omega_o(t_1+t_2-\alpha-\beta) + \cos\, \omega_o(\Delta t+\alpha-\beta)]\, G_t(\alpha)\, G_t(\beta) \qquad (46)$$

Thus, even though the object is perfectly correlated in time, the image coherence function will not be periodic with finite period except in the steady state. Notice that in all of these integrations

the bounds on the space dimensions of the object are -1 to 1. The corresponding analogous bounds on t should be zero to infinity. It is interesting that this definition of the possibility of a transient coherence function, that is, one in which coherence is measured after the signal has been applied for only a finite period of time. It is clear that if the object covers a finite solid angle instead of the 4π radians required to take into account all of the volume of space, the bounds on the space dimension would be less than from -1 to 1.

Now let us consider the opposite extreme, namely pure space incoherence, pure-time incoherence and both time and space incoherence jointly. Assume that the object is separable in the sense of equation (26) and that the weighting function is separable in the sense of equation (42). The image field can then be written as

$$I(x',t) = \int_{-1}^{1} dx\, O(x)\, G_x(x'-x) \int_{o}^{t} d\alpha\, O(\alpha)\, G_t(t-\alpha) \tag{47}$$

If the object field strength is purely incoherent in both space and time, then

$$\overline{O(x_1)\, O(x_2)} = K_1\, \delta(x_2-x_1) = K_1\, \delta(\Delta x) \tag{48}$$

$$\overline{O(t_1)\, O(t_2)} = K_2\, \delta(t_2-t_1) = K_2\, \delta(\Delta t) \tag{49}$$

The corresponding coherence function for I is then

the bounds on the space dimensions of the object are -1 to 1. The corresponding analogous bounds on t should be zero to infinity. It is interesting that this definition of the possibility of a transient coherence function, that is, one in which coherence is measured after the signal has been applied for only a finite period of time. It is clear that if the object covers a finite solid angle instead of the 4π radians required to take into account all of the volume of space, the bounds on the space dimension would be less than from -1 to 1.

Now let us consider the opposite extreme, namely pure space incoherence, pure-time incoherence and both time and space incoherence jointly. Assume that the object is separable in the sense of equation (26) and that the weighting function is separable in the sense of equation (42). The image field can then be written as

$$I(x',t) = \int_{-1}^{1} dx\, O(x)\, G_x(x'-x) \int_{o}^{t} d\alpha\, O(\alpha)\, G_t(t-\alpha) \tag{47}$$

If the object field strength is purely incoherent in both space and time, then

$$\overline{O(x_1)\, O(x_2)} = K_1\, \delta(x_2-x_1) = K_1\, \delta(\Delta x) \tag{48}$$

$$\overline{O(t_1)\, O(t_2)} = K_2\, \delta(t_2-t_1) = K_2\, \delta(\Delta t) \tag{49}$$

The corresponding coherence function for I is then

$$R_I(\Delta x', \Delta t, x', t) = \iint_{-1}^{1} dx\, dy\, \overline{O(x)O(y)}\, G_x(x'-x)\, G_x(x'-y) \cdot$$

$$\int_o^{t_1} d\alpha \int_o^{t_2} d\beta\, \overline{O(t-\alpha)\, O(t-\beta)}\, G_t(\alpha)\, G_t(\beta)$$

$$= \int_{-1}^{1} dx\, K_1\, G_x^2(x'-x) \int_o^t d\alpha\, K_2\, G_t^2(\alpha) \qquad (50)$$

It is interesting to consider the case of a spatially-incoherent or diffuse source which emits monochromatic radiation. In this case the image field strength becomes

$$I(x', t) = \int_{-1}^{1} dx\, O(x)\, G_x(x'-x) \int_o^t d\alpha\, G_t(t-\alpha) \cos \omega_o \alpha \qquad (51)$$

If the temporal characteristic of the spatial frequency filter is a very broad compared to the spatial attenuation characteristic,

$$G_t(t) = \delta(t)$$

The image field strength in this case is

$$I(x', t) = \cos \omega_o t \int_{-1}^{1} dx\, O(x)\, G_x(x'-x) \qquad (52)$$

and the coherence function is

$$R_I(\Delta x', \Delta t, t, x') = \cos \omega_o t \cos \omega_o (t+\Delta t) \iint_{-1}^{1} dx\, dy \cdot$$

$$\overline{O(x)O(y)}\, G_x(x'-x)\, G_x(x'+\Delta x'-y) =$$

$$K_1 \cos \omega_o t \cos \omega_o (t+\Delta t) \int_{-1}^{1} G_x(x'-x)\, G_x(x'+\Delta x'-x)\, dx \qquad (53)$$

The corresponding mean square value for the image, sometimes called the image intensity, is

$$\overline{I^2(x',t)} = K \cos^2 \omega_o t \int_{-1}^{1} G_x^2(x'-x)\, dx \tag{54}$$

Similarly, when the object field strength is spatially coherent and temporally incoherent, for example, when the object is a point source emitting white noise, the image field strength, coherence function and intensity or mean square value are, respectively,

$$I(x',t) = G_x(x') \int_o^t O(\alpha)\, G_t(t-\alpha)\, d\alpha \tag{55}$$

$$R_I(\Delta x', \Delta t, x', t) = G_x(x')\, G_x(x'+\Delta x') \int_o^{t_1} d\alpha \int_o^{t_2} d\beta$$

$$\overline{O(t_1-\alpha)\, O(t_2-\beta)}\, G_t(\alpha)\, G_t(\beta) =$$

$$G_x(x')\, G_x(x'+\Delta x')\, K_2 \int_o^t G_t(\alpha)\, G_t(\Delta t+\alpha)\, d\alpha \tag{56}$$

$$\overline{I^2} = G_x^2(x')\, K_2 \int_o^t G_t^2(\alpha)\, d\alpha \tag{57}$$

Equations (54) and (57) illustrate the often-mentioned fact that image intensity is proportional to the integral of the square of the weighting function of the spatial filter for incoherent sources.

Before the advent of the laser, most optical problems made use of intensity relationships such as those in Equations (54) and (57). With the possibility of nearly coherent or at least partially coherent light, however, coherent optics must be considered, and their coherence properties can be measured and evaluated as has just been illustrated.

It will be observed that this weighting function $G(x,t)$ which is the transform of the spatial frequency filter transfer function in equation (7) is a function of the wavelength $(k = 2\pi/\lambda)$. Thus, the spatial weighting function is dependent on the temporal frequency of the input. This means that G in equation (1) is a function of λ and, consequently, so is the image field strength I. This frequency dependence must be taken to account when the input or object field strength is temporally partially coherent, that is, has a finite temporal bandwidth. In such a case, the time-varying portion of the object field strength can also be written as a function of wavelength or frequency by means of the Fourier integral theorem. Thus,

$$O_t(t) = \frac{1}{2\pi} \int_{-2\pi B}^{2\pi B} O_t(\omega)\, e^{j\omega t}\, d\omega$$

$$= \frac{1}{2\pi} \int_{-2\pi B}^{2\pi B} d\omega \int_{-\infty}^{\infty} O_t(z)\, e^{j\omega(t-z)} dz \qquad (58)$$

where

$$\omega = \frac{2\pi c}{\lambda} = kc \qquad (59)$$

$O_t(t)$ is then a function of the spectral bandwidth B , and if the spectral bandwidth consists only of discrete lines, and the integral in equation (58) reduces to a Fourier series which involves the various frequency components in the series. Strictly speaking, the basic Fourier transform pair, equations (7) and (8) were derived from Maxwell's equations on the assumption that the time variation of the driving function in the wave equation was sinusoidal and is only a steady state solution. To be strictly rigorous, these equations should be solved for the case when the input is not sinusoidal, but rather as represented in equation (58), that is, when the driving function for Maxwell's equations is not sinusoidally time dependent, but rather a spectrum of frequencies. The image time function in the steady state can be found as the sum of the solutions to the various frequency components emitted by the object by employing equations (7) and (8) for every frequency component and then summing the outputs, providing the filter is linear. This procedure does not in general lead to a closed form solution and furthermore does not yield the transient response. If the modulation bandwidth is very small compared to the carrier frequency, which is almost always the case in optics, the weighting function of a spatial frequency filter can be thought of as being the response to an input sinusoid which is frequency variant. This is analogous to the so-called quasi-stationary approach used in f-m. It is assumed that the frequency variations are sufficiently slow so that the response for any given input frequency will be the steady state response of the filter to that frequency from which the over-all response is found by summing the responses to each individual frequency. This sum becomes an integral in

the limit of a continuous frequency variation. With these approximations in mind, employing equations (7) and (8), the response of the filter to a point source (for example, at the origin in space) and a sine wave of frequency ω in time is

$$H(x,y,t,\omega) = H(x,y,\omega)\, e^{j\omega t}$$

$$= \frac{j\omega e^{j\omega t}}{4\pi c} \frac{e^{-j\frac{\omega R}{c}}}{R} \int_A F(\xi,\eta)\, e^{j\frac{\omega}{c}(x\xi+\eta y)}\, d\xi\, d\eta$$

$$= \frac{\pi j \omega e^{j\omega t}}{c} \frac{e^{-j\frac{\omega R}{c}}}{R}\, g(x,y,\omega) \qquad (60)$$

In order to apply equation (12), however, the function $G(x,y,t)$ which is the response to a unit impulse in <u>both</u> time and space must be found. Since the temporal weighting function is simply the integral of the steady state response to a sinusoid in time over all frequencies, the function G can be found from $H(x,y,t,\omega)$ in equation (60) as follows:

$$G(x,y,t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H(x,y,\omega)\, e^{j\omega t}\, d\omega \qquad (61)$$

Equation (12) then becomes

$$I(x',y',t) = \iint_{-1}^{1} dx\, dy \int_{o}^{t} O(x,y,\alpha)\, d\alpha \cdot$$

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} H(x'-x,y'-y,\omega)\, e^{j\omega(t-\alpha)} d\omega$$

$$= \iint_{-1}^{1} dx\, dy \int_{o}^{t} O(x,y,\alpha) \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{j\omega(t-\alpha)} d\omega \cdot$$

$$\frac{\pi k \omega}{c} \frac{e^{-j\frac{\omega R}{c}}}{R} g(x'-x,y'-y,\omega) \qquad (62)$$

where

$$g(x,y,\omega) = \frac{1}{4\pi^2} \int_{A} F(\xi,\eta)\, e^{j\frac{\omega}{c}(x\xi+y\eta)}\, d\xi\, d\eta \qquad (63)$$

If the modulation bandwidth is extremely small compared to the carrier frequency, to a first approximation the weighting function in equation (60) becomes independent of frequency ω in the sense that ω is equal approximately to a constant. In this case, the spatial weighting function is independent of time, that is, there is no temporal filtering by such a filter. The evaluation of the integral (62) is clearly difficult. Even equation (62) does not apply if the modulation bandwidth of the object spectrum is an appreciable fraction of the carrier frequency. When these expressions are invalid, the general expression for the image is even more complicated. Because of the difficulties associated with time incoherent signals, the principal effort for this report has been with temporally coherent sources but with any degree of spatial coherence.

Work is currently under progress to evaluate information rates under these conditions. Once the information rate has been evaluated, it can be maximized with respect to the variable parameters of the filter, thereby optimizing the filter.

Work for the forthcoming period will include an investigation of the difficulties associated with temporal incoherence and an attempt to evaluate information rates for signals which are partially time coherent as well as partially space coherent. Investigation will continue into the behavior of devices other than filters in an optical system, and particularly the effect of these devices on coherence and information rate. Included in this study will be the effect of receiver noise added after spatial frequency filtering, that is, in addition to the background noise present in the image. The effect of nonlinear operations will also be considered, especially simple nonlinear algebraic operations such as encountered in modulators and demodulators.