TECHNIQUES IN TREATING NONLINEAR CLOSED-CYCLE SYSTEMS

Thesis by

Robert South Neiswander

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1954

ACKNOWLEDGEMENTS

The logarithm type function generator described in Section II should be credited to a number of people. To clarify a slight confusion in the field, it is noted that the device, which may be one of a large variety of components constituting an analog computer, does not produce arbitrary functions but those functions which it does produce are easily manipulated. Possibly the device was first suggested during World War II, for a complete block schematic including exponent control appeared in the MIT Radiation Laboratory series (Ref. 23). The Aerophysics division of North American Aviation (and quite probably others) constructed simple multipliers using the logarithm principle, and freely contributed to us results of their research. Research at the California Institute of Technology was originally pointed at developing an electrical analog of a hydraulic servo-valve; since the logarithm type function generator selected for this purpose was obviously useful in modeling other nonlinear problems, a combined project was initiated with R. C. Howard and C. J. Savant. With this arrangement, one of the important items of the device, the logarithm converter, was produced (Ref. 25). Subsequently, a somewhat successful generator was completed and used by R. C. Howard and C. J. Savant (Refs. 26 and 27). Section II describes a different approach to the same problem.

Experimental data and general information regarding aircraft hydraulic servomechanisms (Section III-B) were generously provided by S. Millstone, North American Aviation, Inglewood.

The technique of modeling the instability phenomenon of the supersonic ramjet (Section III-C) largely paralleled an analytic approach suggested by Dr. Stoolman. I am indebted to Dr. Stoolman for his constructive criticisms and comments.

Much of the groundwork leading up to the technique of treating nonlinear servomechanisms (Section IV) resulted from a study (Ref. 34) made under the supervision and with the kind assistance of Dr. MacNeal.

Such lucidity and accuracy as exist in the thesis, particularly in Section IV, are mainly attributable to Dr. Wilts.

For the kind assistance of Jeanne Shacklett, Betty Wood, and of the editorial staff of Hughes Aircraft in producing the finished thesis copy, I am most grateful. Recognizing that the success of an electronic device is always due to many factors beyond theoretical design, I mention R. Jones as being responsible for the final, well-operating construction of the function generator.

To my wife, Margo, I offer first my apologies for an antisocial preoccupation with certain scholastic duties, and second my appreciation of her indulgence, source of inspiration and sense of humor.

A large portion of this work was carried out while I was a member of the technical staff of the Hughes Research and Development Laboratories and while a holder of the Howard Hughes Fellowship. Without Dr. McCann's generous assignment of funds, this research project would not have been realized.

ABSTRACT

The objective of this research is to examine and to develop methods of treating nonlinear, closed-cycle, single loop systems. Analytical and graphical methods are immediately rejected for their failure to include the complexities of realistic systems. Experimental techniques as, for example, electrical analog modeling are flexible and can accommodate complicated system descriptions. Associated with such a technique are three important steps: 1) selection of suitable analog elements, 2) arraying or modeling these elements to properly simulate the actual system, and 3) devising an experimental test procedure that produces appropriate insight into the cause and the control of results.

Designed for the exploratory testing of nonlinear systems the electronic, logarithm type function generator is a nonlinear element that creates an easily modified function of two or more input variables. Its computing speed is sufficient for oscilloscope monitoring of solutions.

System modeling is the most critical step; as demonstrated by two system examples, the deceptively simple hydraulic servomechanism and the supersonic diffuser instability.

Based upon a limited Taylor series approximation of the input function the experimental procedure used to design a realistic servomechanism for best saturated performance is quite simple. It utilizes intentional shaping of the rate feedback function (or the static error). Performance within the system's customarily linear region can be improved by intentional saturation.

TABLE OF CONTENTS

I. INTRODUCTION

The objective of this research is the development of techniques useful in treating nonlinear, closed-cycle systems. In contrast to much of the existing work in nonlinear mechanics, the approach here is essentially nonmathematical. Arguments and results appeal to physical intuition.

Here, the formulation of the research program is reviewed. Since the aim of the research is to develop ideas and concepts that might be useful to an engineer faced with a nonlinear problem, methods rather than detailed solutions are emphasized. Most of the realistic systems are complicated. Response solutions of a real nonlinear system require detailed definitions of the system's characteristics and in general have meaning only in connection with the specific, defined system. On the other hand, a method concerns itself with a group or class of systems and thereby offers possibility of application to many detailed problems.

The justification of such a research project entails first, evidence that nonlinear systems which are in need of study exist, and second, evidence that methods already available are inadequate. The first point can be shown in many ways: for one, all natural (and for that matter, all manmade) occurrences are nonlinear if for no other reason than having limited energy available. In treating nonlinear systems, formal mathematical methods with their associated rigor are not generally applicable, for the infinitely diverse cannot be codified. In many instances, experimental techniques are available, but the task of interpretation is often difficult.

Closed-cycle phenomena form an important family of nonlinear occurrences. The basic elements are 1) a controller of energy (manifested by mechanical force, temperature, hydraulic pressure, etc.), 2) a reaction of part of the system, the load, to the controlled energy, and 3) one or more feedback or loop closures by which the reaction influences the controller. The reciprocating engine is a nonlinear feedback system, an oscillator, in which the energy source is combustion and the reaction is crankshaft rotation influencing combustion by valving, ignition, and piston position. External influences control steady state engine speed (as denoted in nonlinear mechanics, the outer, stable limit cycle); the inner (low speed) unstable limit cycle establishes minimum starting speed initiating oscillations. In modern aeronautics, aero-elastic effects such as flutter, supersonic diffuser "buzz", and rocket and ramjet flame instabilities are important, unsolved problems. Intentionally closed systems such as hydraulic and pneumatic servomechanisms exhibit prominent nonlinearities, and all servo controls display large nonlinearities in saturated regions of operations. The present research is concerned with the elemental form, the single loop nonlinear system. Although efforts are for the most part directed toward treating servomechanisms, one naturally closed-cycle system, the simplified supersonic diffuser model (Section III, Part C), is included.

One of the basic system characteristics of general interest is stability which has an intimate relation with all closed-cycle phenomena. The other characteristic is performance, which is associated with systems intentionally closed to perform certain operations. A nonlinear system may display both stable and unstable regions of operation; the requirement that the system be stable (or perhaps unstable, as in the case of an oscillator) must also precisely specify the region of operation. For example, stability near the null or singular point of a servomechanism is, despite conceptions extrapolated from linear servoanalysis, not of principal interest. If a servomechanism has some tolerated error, $|\delta|$, null instability is acceptable and may be desirable provided a stable limit cycle occurs within the allowable error. Any nonlinearity whose effect is felt in the null region may produce this situation, as, for example, back lash or dead space or coulomb friction. Alternatively, an acceptable oscillator may have a stable null but also have a small unstable limit cycle such that a small excitation pushes it into the unstable region. The other characteristic, dynamic performance, is as diversified as the servomechanism's applications. Obviously, no one universal performance requirement exists. Some of the commonly encountered dynamic requirements for control systems are;

1. Position jump (step function): Move from initial position 'a' to position '$b \pm |\delta|$' (δ being the tolerated error) within the shortest possible time.

2. Velocity jump (ramp function): Move most quickly from initial position 'a' to '$b \pm \delta$', b moving at a constant velocity.

3. Sinusoidal input: Follow a sinusoidal input within

tolerated amplitude error, $|\delta|$ (or tolerated phase error $\Delta\phi$) maximizing frequency.

4. Regulator: Match a fixed input minimizing error for changing loads.

5. Defined time varying input: Minimize the rms error of the output in following some defined input.

A complete performance requirement includes at least two specifications: 1) a definition of the input (and possibly other explicit time variables), and 2) a definition of the departure or error from ideal performance, i.e. a criterion for evaluation and improvement. Sinusoidal inputs and jump functions are most commonly used in analyses. The other specification, evaluation criterion, perhaps needs a little clarification because customary linear analysis fails to define a useful ideal system. The theoretically ideal linear system has no limitations, a point of academic interest but of no assistance to the design engineer. In practice, a real servomechanism must accept a set of physical limitations: e.g., actuator force, actuator speed, actuator inertia, total power, etc. The intelligence of the system, provided by intentionally added components, ideally makes the most of these physical handicaps★; that is, with ideal intelligence, a system is limited only by its physical restrictions (a state usually referred to as "saturated"). If the performance of such a system

★ It is assumed throughout this thesis that signals from the input and the feedback have been stripped of spurious information, thus isolating system response subjected to severe noise as a separate problem.

is known, it can serve as a criterion for evaluating actual system performance.

As previously mentioned, a scarcity of methods applicable to nonlinear systems exists at present. To substantiate this assertion, a review of currently available methods is made. For clarity, "internal" techniques are distinguished from "external" techniques. An internal technique involves the human mind intimately in all the details and quite possibly in the mathematical rigor of the solution. Included here are all of the analytical and most of the graphical methods. Although solutions by such processes are thorough, the many restrictions exclude most practical applications. In contrast, "external" techniques, essentially experimental methods, involve the important inputs and outputs of the system; the complex of internal workings of the system are often neither defined nor understood in detail. Necessarily this technique involves either testing the system itself or using some suitable analog. The assignment of the human mind to such method is as problem director, rather than problem solver.

Most of the current methods are "internal". From an internal point of view, the nonlinear system may be seen as an ordinary, differential equation:

$$L\left\{x,t,x_i\right\} = 0^{*}$$

where x_i is the input, x, the output, and t the independent variable,

* Throughout this report, the brackets $\{\}$ denote "a function of ... ."

time. For our purposes, x_1 is a defined function of time, allowing the equation to be written

$$L\left\{x,t\right\} = 0 \qquad \text{I-1}$$

If the equation is quasi-linear, autonomous, and its solution has certain continuity properties, the graphical phase space method developed in Appendix A can be employed. Unfortunately, this type of analysis applied to higher order systems is tedious, and insight into system action is obscured.

Admitting linearity restrictions, we obtain a more mathematically acceptable equation:

$$L_o\left\{x,t\right\} = f\left\{x,t\right\} \qquad \text{I-2}$$

where L_o denotes a linear, constant coefficient, ordinary differential operator. From here, two directions are open. By letting $f\left\{x,t\right\} = k \operatorname{sgn}\left[g\left\{x,t\right\}\right]$,* time appearing only in derivatives, we get the family of linear systems with discontinuous forcing functions, i.e. linear, relay-controlled servomechanisms. Mme. I. Flugge-Lotz (Ref. 1) rigorously treated the linear switching problem in which L_o is second order and $g\left\{x,t\right\} = g_1 x + g_2 \dot{x}$. Bushaw (Ref. 2) extended the solution to include an optimum $g\left\{x,\dot{x}\right\}$, introducing the criterion that optimum performance is defined as the quickest response to a step function input. McDonald (Refs. 3 and 4) somewhat empirically arrived at a similar result for the linear second order system with the "spring" (or x) term missing.

* Here, sgn [ ] means "the polarity, or sign, of ...".

More recently, Rose (Ref. 5) has studied the n dimensional linear system with an "optimmm" switching function. When $L_o\{x,t\}$ is second order, the switching solution can always be treated by phase plane analysis, as has been done for various systems by MacColl, Weiss, Hopkin, Uttley and Hammond (Refs. 6, 7, 8 and 9, respectively), and others. An alternative method, suggested originally by Hazen (Ref. 10), utilizes standard Laplace transforms or other linear techniques for the successive linear pieces of the response.

Another attack on Eq. I-2 was devised by Poincare (Ref. 11) in which the nonlinear portion, $f\{x,t\}$ was considered as a small perturbation. His method of small parameters applies to the equation:

$$\ddot{x} + \omega^2 x = \mu\{x, \dot{x}, t\} \qquad \text{I-3}$$

with μ a small quantity compared with x and $\ddot{x}$. Van der Pol (Ref. 12), the man responsible for inciting the modern wave of interest in nonlinear mechanics, intuitively treated Eq. I-3 with $\mu = k(1 - x^2)\dot{x}$, approximating a vacuum tube oscillator. In Russia, Kryloff and Bogoliuboff (Ref. 13) revised Poincare's method for engineering applications and introduced the concept of "equivalent linearization". Ignoring all output components from a slightly nonlinear element except the fundamental of the input, one can define an equivalent impedance for the element. It is interesting to note that this approach, suggested possibly fifteen years ago, is identical in concept to the recent popularized frequency method of Kochenburger (Ref. 14).

Linearization of a system to the extent that nonlinearities appear only as switching functions or small perturbations is not usually possible. The topological method (Ref. 15) suggested by Poincare is in theory applicable to nonlinear systems with many dimensions, but from an engineering point of view is applicable only to second order, quasi-linear, autonomous equations. Singular point stability concepts of Liapounoff and the limit cycle theorems of Bendixson provide information about specific operating regions. Lienard devised a simple geometric technique for the construction of phase plane (actually distorted and designated the Lienard plane) trajectories for equations of the forms:

$$\ddot{x} + f\{x\}\dot{x} + x = 0$$

and

$$\ddot{x} + f\{\dot{x}\} + x = 0$$

Levinson and Smith (Ref. 16) investigated the relaxation oscillator equation

$$\ddot{x} + g\{x,\dot{x}\}\dot{x} + h\{x\} = 0$$

and Rauch (Ref. 17) has treated one of the few third order systems mentioned in literature:

$$\dddot{x} + (k_1 + k_2\, g\{x\})\ddot{x} + k_3\, g'\{x\}\dot{x}^2$$

$$+ g\{x\}\dot{x} + x = 0$$

There are several methods not noted above which for the most part are versions of known techniques: to mention two, the variation of system parameter of Gypser (Ref. 18) and the "General Linearizing Process for Nonlinear Control Systems" of Loeb (Ref. 19) a frequency method notable for its lack of generality.

In summary, internal techniques at present are confined to systems which can be represented either by second order quasi-linear, autonomous equations or by higher order, piece-wise linear equations.

Although the internal methods are exceedingly attractive to the engineer with a nonlinear problem, most nonlinear real systems cannot be fitted into this minute group. Attempted pruning down of a complicated phenomenon to one of the simple internal forms usually produces misleading results. For example, a nonlinear servomechanism can, by ignoring a sufficient number of terms, be represented by a quasi-linear second order equation and treated by phase-plane analysis. However, we should be most skeptical of the results; experience with linear systems tells us that the solution of a second order system is trivial, and for proper analysis orders as high as fifth or tenth may be required. The effects of these higher order terms would not be expected to lose their importance when the system is nonlinear. For the majority of real, nonlinear, closed-loop phenomenon, experimental treatments must be used.

External techniques can accommodate complex systems with numerous nonlinearities. This admits a great many systems not tractable to internal techniques but also introduces difficulties associated with experimental techniques. Excepting cases in which the system itself is available for first hand observation, modeling is necessary. Modeling often takes the form of scaling, as is done in wind tunnels and towing tanks, or by analogs -- the former being generally more accurate and the latter more flexible. In particular, the electrical analog is adaptable to a large variety of modeling, and is utilized throughout this research. The ability of the analog to duplicate the original system depends upon both the elements of the analog and the problem director's translation of the phenomenon into analog terms. These are fundamental considerations of experimental solutions and consequently form the first portion of the research.

The other basic factor of the analog solution is procedure. Usually, the system is explored by varying one or more of its elements and observing and interpreting the resultant solutions. The admission of nonlinearities greatly confounds the problem. This is apparent even in the simplest example, a nonlinear servomechanism in which some parameter function is controllable. If the system were linear, the parameter would be a constant coefficient, and all probable values of this constant could be readily tested in the model. In the nonlinear system, this parameter is a variable coefficient, a function of one or more variables. Without plan, the problem director is confronted with the somewhat futile task of selecting samples from all possible curve shapes. The experimental technique must be capable of establishing a procedure, initially rejecting by simple means all but a small, logically organized family of functions.

Experimental tests not based upon some logical procedure have value only in connection with the specific system tested. Correlation is poor or does not exist on the level of results alone. This is due to the individualistic nature of the nonlinear system;

an element inserted to provide good performance in one system probably will not work in another, and system performance cannot be synthesized by a study of component performances. There is more generality associated with the experimental methods in that they apply to groups, whereas results apply to specific systems.

As previously mentioned, the aim of the logical experimental method is good performance, which requires a definition of the system input. Procedures which involve sinusoidal inputs are termed here "Fourier Methods", and procedures involving jump functions are termed "Taylor Methods."*

If a single sinusoidal input of frequency f is introduced into a system having the proper nonlinearities, the system output can be represented by a Fourier series with terms, f, 2f, 3f, ... When the harmonics, 2f, 3f ... are sufficiently small to be neglected, the system or a component of the system can be represented by an "equivalent impedance" (Ref. 13) which usually changes with both frequency and amplitude. Experimental techniques of treating these equivalent impedances have been developed by Kochenburger (Ref. 14), McCann and Wilts (Ref. 20), and others. The usefulness and popularity of this approach lies in its close correlation with existing linear techniques. Extended Nyquist diagrams, extended Bode plots, etc. (Refs. 21 and 22) are permissible. On the other hand, there are some serious limitations to the equivalent linearization techniques. First, there is the same lack of a

* "Power Series Methods," in which the inputs are expressed as polynomials of time, might also be used but at present have been concerned only with linear systems.

useful performance criterion as exists in linear system techniques. Second, the basis of the method consists in small perturbations from a linear system. A profoundly nonlinear system may produce many important effects excluded from the analysis, such as large higher harmonics, subharmonics, non-integer harmonics, and frequency entrainment. Superposition, assumed in the method, is not in general possible; the response of a nonlinear system to simultaneous input frequencies f_1 and f_2 is not the sum of the individual frequency responses.

The "Taylor Methods" can be grouped loosely as those which expand the input, $x_1(t)$ into a Taylor series,

$$x_1(t) = x_{1_o} + a_1 \dot{x}_{1_o} t + a_2 \ddot{x}_{1_o} t^2 + a_3 \dddot{x}_{1_o} t^3 \; . \; . \; . \; .$$

and consider a limited number of terms. An approximation of the general input can be built up as follows: at time t_1, the system (at rest) is instantly requested to move to some x_{1_o} , which is of course a simple step. After steady state is regained, the system is requested to move to another x_{1_o}. Thus, in stairstep fashion, a general input can be approximated by a series of step functions. Crude as this one term approximation may seem, it has the great advantage of allowing a simple interpretation of the "ideal system." Here the ideal system is one which with defined physical limitations most rapidly follows a step input. Simple servomechanisms have been treated in this fashion by McDonald (Ref. 4) and others, with interesting results. One might expect a system designed by such a technique to work best for step inputs and rather poorly for continuously varying inputs such as sinusoids. Curiously, the systems thus optimized not only were superior in step function response, but also had superior frequency response compared to identical systems optimized by standard linear techniques. (For example, see the responses Fig. IV-20.) Perhaps this isn't as surprising as it first appears. Of the infinitude of intentional variations of a given control system, the likelihood of the special variety, the perfectly linear system, best fulfilling a given performance requirement is negligible.

Intuitively, one would expect that the inclusion of more terms of the Taylor series in treating the system would better performance.

In summary, the exploration of real, nonlinear, closed-cycle phenomena usually requires experimental techniques, introducing the necessity for providing suitable analogs or models and also the necessity of logical procedures for testing the models. The latter point is particularly applicable to servomechanism performance tests.

Therefore, the objectives of the research program are:

1. To develop modeling techniques associated with the electrical analog, including both analog elements and problem "translations" from actual to analog terms.

2. To develop a logical procedure of performance improvement of nonlinear servomechanisms, based upon a) an "ideal" system which utilizes its maximum physical capabilities and b) approximation of performance responses to an input by a limited Taylor series.

II. MODELING TECHNIQUES

A SPECIAL NONLINEAR ANALOG ELEMENT

A. Introduction

The selection of the electrical analog for nonlinear exploration is not by chance; its erstwhile rival, the digital computer, is noncompetitive for applications where versatility both initially and in the process of the problem solutions is held more important than accuracy. Whereas the digital machine requires complete and precise (and laborious) problem programming prior to performing a relatively slow solution, the electrical analog can present a complete transient response of a complicated system in possibly 1/100th of a second and responds instantly to changes of system parameters. Spur-of-the-moment changes necessary in efficient exploration are simply effected.

Having experienced several years of development and use, the basic electrical analog designed for linear and simple arithmetic operations has reached an acceptable degree of perfection. Elements such as passive units (resistors, capacitors, inductances, and transformers), d.c. feedback amplifiers, and electronic multipliers are assumed throughout this research to be standard items. Nonlinear functions generators are in a less advanced stage. For the application at hand, the function generator should be rapid enough to allow oscilloscope monitoring of solutions, should produce functions of two or three independent variables, and should provide for easy variation of these functions.

Accuracy need not be comparable with that of non-exploratory, specific function generators. A review of currently available nonlinear function generators is presented to show the desirability for new nonlinear elements designed especially for exploratory work.

Diode switches: Diodes are arrayed to switch resistances in or out at various voltage (or current) levels. Complexity ranges from simple limiters to 20 or 30 segment units. Advantages: high speed operation, can approximate unusual curve shapes by multi-segments. Disadvantages: nonlinear function is of only one variable, a simple curve change may require 20 or 30 adjustments.

Photoformers: Beam of a cathode ray tube is slaved in one coordinate to follow a photographed mask as the other coordinate is varied either with time or with another independent variable. Advantages: high speed operation, good accuracy, duplicates any single valued mask curve. Disadvantages: nonlinear function is of only one variable, a simple curve alteration requires a new photographic mask.

Multipliers: By successive multiplication of the input by itself, a limited power series is available, $f(x) = a_o + a_1x + a_2x^2$ .. Advantages: multipliers are usually available as standard components. Disadvantages: electro-mechanical multipliers are slow, electronic multipliers perform only one multiplication per unit and are not efficiently utilized if the function requires several series terms.

Resolvers, etc.: Electro-mechanical devices can produce sine-cosine functions by resolvers and certain nonlinear functions by means of potentiometer combinations. Advantages: computers such as the REAC have these functions readily available. Disadvantages: limited number of curve shapes available, slow speed operation.

Electro-Mechanical Followers: The arbitrary function generator can be thought of as an X-Y recorder run in reverse manner, with the arbitrary function as a curve (often metallic) affixed to the chartboard. The independent variable moves one coordinate of the follower, and the corresponding curve value is measured either by resistance devices or by slaving a curve follower. Advantages: any single valued function can be duplicated. Disadvantages: nonlinear function is of one variable, slow speed operation, and a simple curve alteration requires a new chart.

In summary, the electro-mechanical devices are slow in operation and are either quite limited with regard to functions available or require new chart setups for each function alteration. The latter is also true of photoformers. The diode switching device, demanding 20 to 30 segment switches to reproduce general curves, also demands 20 to 30 controls for simple curve alterations. This observation is not intended to mark their inferiority, since all of these devices are extremely useful for the express purposes for which they were designed. None, however, adequately fills the new requirements for an exploratory nonlinear function generator.

From a survey of potentially useful schemes, as yet undeveloped, the logarithmic multiplier showed the most promise. The logarithmic multiplier is all electronic, allowing high speed operation; it can accept several independent variables; and although it is not an arbitrary function generator, a fairly large and useful set of functions are available by simple exponent controls. It was concluded that the successful development of such a device would constitute a useful contribution to nonlinear exploratory techniques; accordingly, this device constitutes the first portion of the research.

B. The Logarithmic Function Generator

The theory of the logarithmic function generator is most simple: inputs, appearing as d.c. voltages, are converted to logarithms; then added, subtracted, or scaled (multiplied by constants); and finally, converted to inverse logarithms. If the inputs are x, y, or z, the output is of the form $x^p y^q z^r$, where p, q, and r have continuous range both positive and negative. Since the inputs must always be positive when converted to logsrithms, polarity signs must be handled independently -- an additional requirement, but one which also provides some useful discontinuous functions. With polarity control, a three variable input type generator can produce functions of the form:

$$\text{Output} = \text{sgn}\ (x-a_1)\ \left|(x-a_2)\right|^p\ \text{sgn}\ (y-b_1)\left|(y-b_2)\right|^q\ \text{sgn}\ (z-c_1)\left|(z-c_2)\right|^r \qquad \text{II-1}$$

where again, p, q, and r have continuous range positive and negative. Some of the useful forms of Eq. II-1 will be shown later in this section.

The logarithmic multiplier as an electronic device has been proposed several times (for example, Refs. 23 and 24), but a number of practical difficulties lie in the way of obtaining a working device. One obvious difficulty is obtaining by electronic means, true logarithms and inverse logarithms; another difficulty is electronic polarity control.

Since so few published data exist concerning the logarithmic multiplier as a unit, historical notes here are brief and are concerned principally with research done at the California Institute of Technology. As evidenced by earlier experiences at North American Aviation, the heart of the computer is the logarithm converter, and a joint research effort by C. J. Savant, R. C. Howard, and the writer resulted in a triode converter (Ref. 25) having fairly stable conversion abilities over an input range of 3 to 300 volts. At the conclusion of this research, a simple multiplier was set up using a standard d.c. amplifier with a logarithm converter in its feedback for inverse logarithms. Operating speed was slow, and no polarity control was included. One version of the logarithmic function generator, completed by R. C. Howard and C. J. Savant (Refs. 26 and 27), used the logarithm converter mentioned above, a high loop gain inverse logarithm converter, and a limiter type polarity control. The generator satisfactorily performed several nonlinear problems. The invaluable contribution of this generator was that it proved the initial scheme was sound. Secondarily, it pointed out several of the troubles and sources of error expected of such a device, the most important of which were 1) time hysteresis effects, attributable to the high gain inverse logarithm converter; 2) excessively large number of preliminary balancing adjustments; 3) poor accuracy, directly due to the low level of input voltage to the logarithm converters; 4) instability tendencies of the inverse logarithm converter; 5) drift; 6) lack of simple control of scaling and exponents; and 7) large space requirement. Developments described below were aimed at eliminating as many of these objectionable features as possible.

The present version of the function generator is shown schematically in Fig. II-1 and physically in Fig. II-2. Except for a standard summing amplifier and the power supply, the complete function generator is assembled as plug-in components behind a relay rack panel 17 1/2 inches high. The face of the panel is arranged as a patchboard on which most of the jumpers are shorted double pin banana plugs. Any component can be isolated and checked individually from the patchboard. Each of the logarithmic converters has a helipot in its output for control of exponents between 0 and 1; exponents greater than 1 are obtained by using higher gains in the summing amplifier. Another helipot supplies 0 to ±100 volts as a bias term (the "b" factor noted later). The components of the generator are briefly discussed below in the sequence they are utilized. For a more detailed discussion, refer to Appendix B.

Three inputs are received at the jacks on the far left of the panel (front view, Fig. II-2). A plug-jumper introduces these inputs to simple germanium diode arrays (as seen at the right hand edge of the rear view, Fig. II-2) which split each input into its positive portion and its negative portion. The diode array is passive, has a gain of unity, and requires no adjustments. It is frequency insensitive within the frequency range of interest.

The positive and negative portions of each input are fed (by a pair of plug-jumpers) to differential inputs of small amplifier (shown in Fig. II-3) such that the positive part is amplified by +6 and the negative part by -6. As established by best logarithmic converter range, the output of the amplifier is linear between 0 and +350 volts. The amplifier has a relatively low loop gain (about 1500), which is easily checked by removal of a chassis plug-jumper opening the loop; it requires only one balancing adjustment, and it passes full wave rectified sine waves of 500 cps without undue phase distortion. Linearity distortion contributed by the amplifier is negligible compared to the overall computer errors.

On the patchboard, the amplified absolute value of each input is applied by a plug-jumper to a logarithmic converter (shown in Fig. II-4). This is the triode converter mentioned in the historical sketch; it produces - $\log_a$ of its input. Unfortunately, five adjustments are required to get acceptable conversions. Thus, the accuracy of this device is difficult to pin down and depends largely upon the patience of the experimenter. With reasonable diligence, an accuracy of ±5% referred to the input might be expected.

Outputs from the logarithmic converters and the bias potentiometer are channeled to a standard d.c. summing amplifier having a negative gain for positive exponents and a positive gain for negative exponents. From the summing amplifier, which produces the logarithm of the absolute value of the output function, we transfer to the inverse logarithmic converter (see Fig. II-5). This component is a positive gain d.c. amplifier with the logarithmic type converter as its feedback. Generally a "forward" gain is selected in a functional (feedback) amplifier such that the feedback element essentially determines the overall closed-loop characteristics. Peculiarities of the amplifier supplying the gain are thus reduced below some specified minimum. Here, however, the feedback element itself is a peculiarly operating nonlinear triode. It makes no difference in the overall system whether the desired closed-loop characteristic is due entirely to the feedback or is in part due to the amplifier. With a nominal gain of 1000 or so, variations (with time) of amplifier characteristics are negligible compared to drifts in the feedback. In short, there is no particular reason for requiring high loop gain.

The present logarithmic converter has a forward gain of about 1500 and as a consequence is easily stabilized without numerous large capacitors and consequent frequency restrictions. By removing the two chassis plug-jumpers, forward and feedback sections can be isolated for independent check and adjustment; the forward amplifier has one balance control and the feedback has the usual five adjustments. Accuracy is comparable to that of the logarithmic converter: about 5% (including patience) referred to the output. The frequency response is considerably improved over a comparable high loop gain converter; frequencies below 200 cps are not appreciably distorted. This distortion is roughly equivalent to that contributed by the standard summing amplifier (remembering that here we are concerned with the full wave rectification of the base frequency).

From the inverse logarithmic converter, the signal is transferred by a plug-jumper to the polarity control (Fig. II-6). This unit is basically a logical circuit: an assembly of triode clamps all operating to ground that produces a positive output if both sign control voltages (provisions for two sign controls were incorporated in this particular generator) are of the same sign, and otherwise a negative output. Only one balancing adjustment is required. The unit has a gain of ± 1/4 with no major inaccuracies in linearity and will switch +300 volts input at frequencies up to 500 cps. The threshold of the sign control voltages is less than 1/4 volt.

Although the ultimate output function (Eq. II-1) of the generator may be quite complicated, it can be resolved into 1) the sign control and 2) factors of the form $|x|^r$. The final function is a product of the terms of 1) and 2). Control of signs is fairly obvious from Eq. II-1, and no difficulties or restrictions are involved in physically effecting this control in the computer. The general factor $|x|^r$ requires some special effort if the computer is to be used most efficiently, as established by the logarithmic converter requirement of approximately 300 volts maximum input.

Consider the production of a function $|x|^r$ having a defined range x_{min} to x_{max}. This requires one input channel of the computer. From the defined gains of the input amplifier and output polarity control, we can write (voltages are defined in

Fig. II-1):

$$e_1 = 6\,|e_o|$$

$$\pm\, e_7 = 1/4\; e_6$$

To get 300 volts maximum desired at the logarithmic converter, the maximum input voltage should be 50 volts,

$$e_o = \frac{50}{|x_{max}|}\,|x|$$

From the logarithmic converter, we obtain

$$e_2 = -\log_a e_1 = -\log_a \left|\frac{300}{x_{max}}\,x\right|$$

which is summed with a bias voltage we will define as

$$e_3 = \log_a \left|\frac{300}{x_{max}}\,b\right|$$

Giving

$$e_4 = e_2 + e_3 = -\log_a |x/b|$$

The exponent control multiplies e_4 by $-r$,

$$e_5 = +\log_a |x/b|^r$$

And the final output is

$$e_7 = 1/4\; e_6 = 1/4\,\log_a^{-1}(e_5) = \frac{1}{4b^r}|x|^r \qquad \text{II-2}$$

Now we can see why the bias, b, is necessary. For best accuracy, the inverse logarithmic converter must have a maximum input corresponding to 300 volts maximum output; that is,

$$e_{6_{max}} = 300 = \frac{1}{b^r} |x|^r_{max} \qquad \text{II-3}$$

and it follows that the bias voltage, e_3, is defined by

$$e_3 = \log_a(300b) = \log_a (300|x|_{max})(300)^{-1/r} = \frac{r-1}{r}\log_a(300) \qquad \text{II-4}$$

From this set of equations, the procedure for setting up functions can be established.

Procedure 1. Determination of the logarithm base, a.

If the input to a converter is scaled 1:1, i.e. number = volts, and the output is similarly considered, then the base of the conversion logarithm is defined by the measured input and output, $E_{out} = - \log_a E_{in}$. The use of one-to-one scaling to define a is done for convenience.

1. Set exponent $r = 2$.
2. Set $e_0 = 50$ v.
3. Adjust b (that is, e_3) until
   $e_6 = 300$ v.
4. From equation II-4, $a = (300)^{\frac{1}{2e_3}}$.

Procedure 2. Producing a single factor function $|x|^r$.

1. Maximum input voltage should be 50 volts.
2. Exponent, r, set by helipot for $0 < r < 1$, or by the summing amplifier gain for $r > 1$. Summing amplifier should produce reverse sign from sign of r.
3. Voltage bias, e_3, may be set by equation II-4, or by trial.

4. The output function, $|x|^r$, has its scale determined by equation 2; i.e.

$$|x|^r = (4b^r)\ e_7$$

If multiple factors are desired, the mathematical determination of bias, e_3, is less obvious; the bias can probably best be found by sweeping $|x|$ through its intended range and adjusting e_3 by observation of the maximum value of e_6 or e_7.

C. Typical Functions Produced by the Generator

An attempt to categorize and describe all of the possible functions available would be a sizable research program in itself. However, expressly for purposes of demonstration, a few typical functions available are presented here.

1. Single factor functions.

In general, the single factor functions are of the form $u = \operatorname{sgn}(x-a)\cdot\left[|x-b|^r + d\right]$.

a. $u = |x|^r$. Positive values of r are shown in Fig. II-7, and negative values of r are shown in Fig. II-9. All curves are symmetric about the u axis.

b. $u = \operatorname{sgn} x\cdot|x|^r$. The set of traces shown in Figs. II-8 and II-10 are identical to those of case a above, except that symmetry is now about the origin.

c. $u = \operatorname{sgn} x\left[|x|^r + d\right]$. The polarity control has been designed to accept only positive inputs, as would be normally expected from the inverse logarithmic converter. Consequently, when the term d, added to the output prior to switching, is negative, certain portions of the output may be zeroed. For example, Figs. II-11 and II-12 show this "dead space" effect. When d is positive, discontinuities such as shown in Figs. II-13 and II-14 are available.

d. Inclusion of the term b simply translates the picture left or right, while term a translates the switch point.

2. Double and Triple Factor Functions.

The double and triple factor functions are product combinations of the functions described above. Quadratics, cubics, and certain higher order polynomials are available, as well as more complicated switching effects. Three examples of double factor functions are shown in Figs. II-15, II-16, and II-17.

3. Special Techniques.

There are also special arrangements of the computer which will produce unusual functions. For example, if the output u is reintroduced as one of the inputs, barring instabilities, we get a function of the form

$$u = \mathrm{sgn}\,(x-a_1)\cdot|x-a_2|^p\ \mathrm{sgn}\,(y-b_1)|y-b_2|^q\ |u-c_1|^r$$

The phrasing of the nonlinear function always demands that accuracy be weighed against simple adjustments. Specific applications of the function generator to various investigations are made in the succeeding sections of this research.

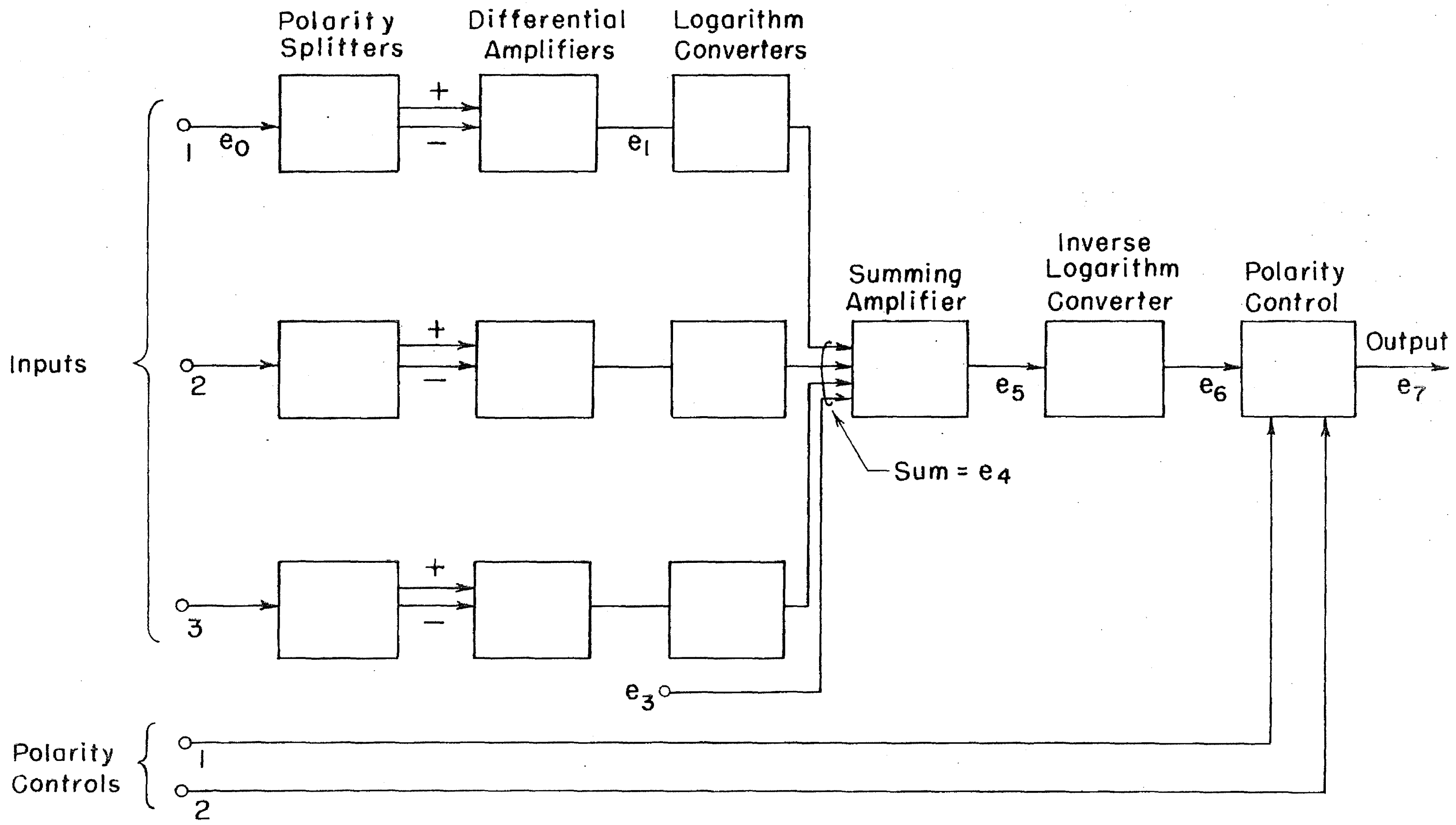
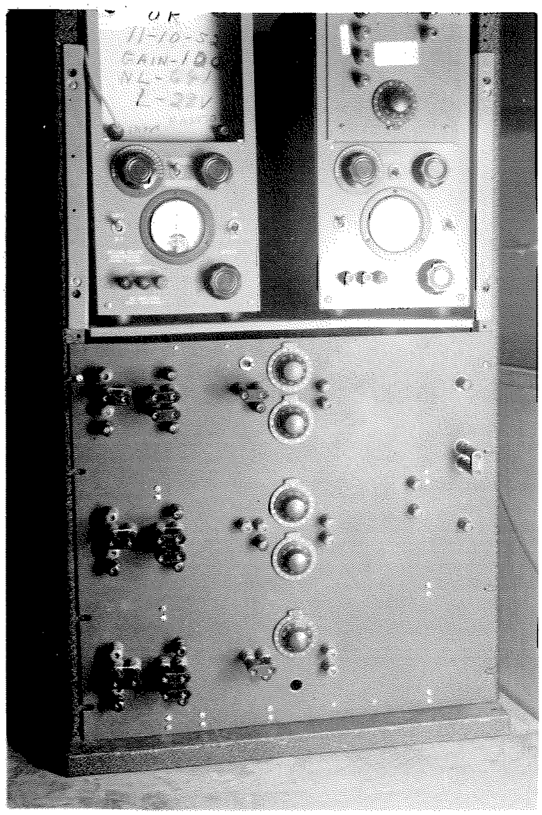
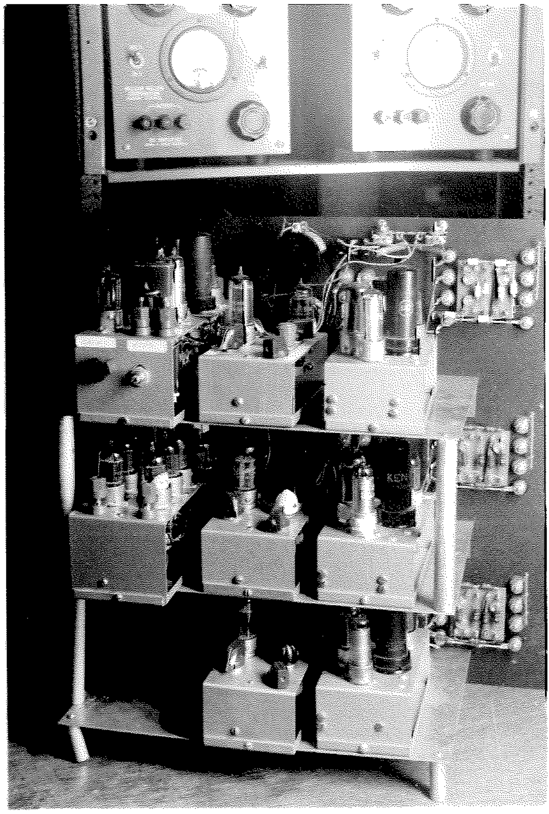


FIG. II-1 BLOCK DIAGRAM OF FUNCTION GENERATOR



Front View of

Function Generator



Rear View of

Function Generator

Figure II-2

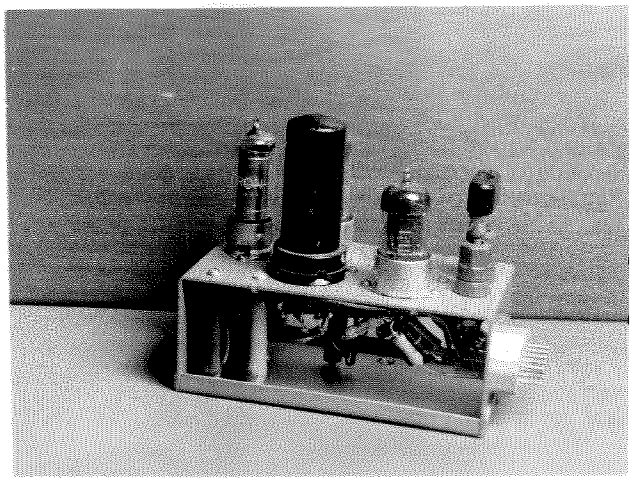
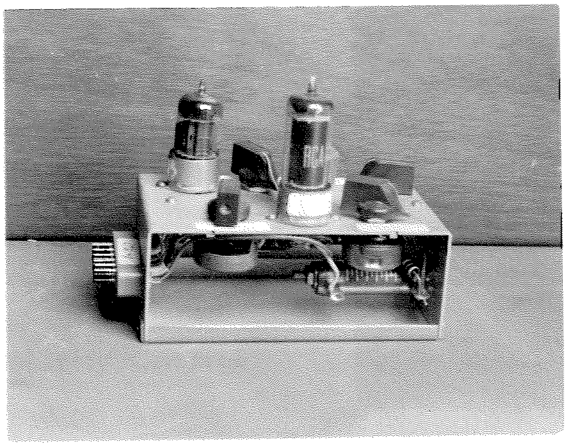


Figure II-3

Differential Amplifier

Figure II-4

Logarithmic Converter



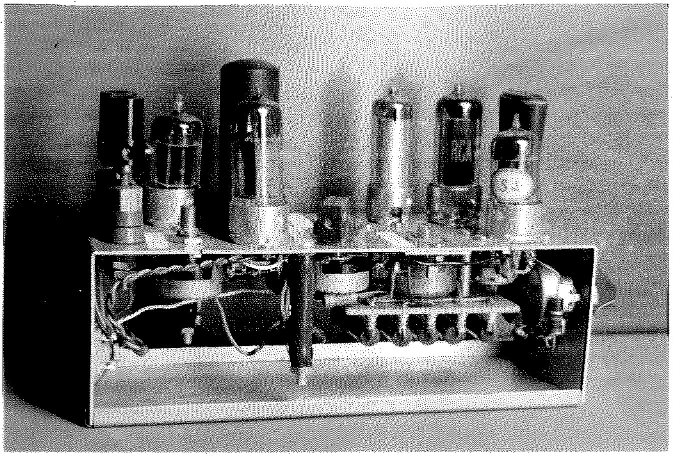


Figure II-5

Inverse Logarithmic Converter

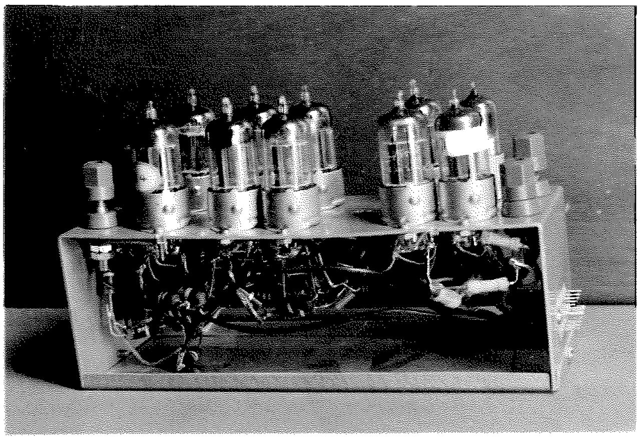


Figure II-6

Polarity Control

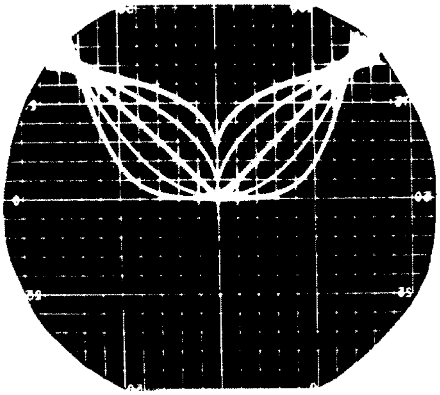


FIGURE II-7

Functions of tho form

$u = |x|^r, r > 0$

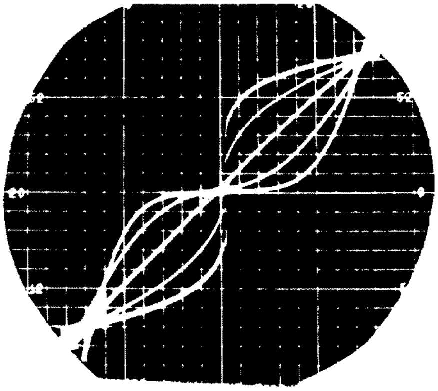


FIGURE II-8

Functions of the form

$u = \operatorname{sgn} x |x|^r, r > 0$

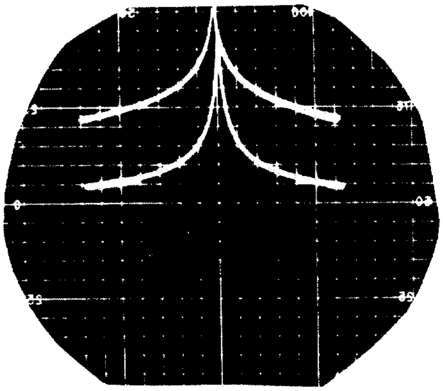


FIGURE II-9

Functions of the form

$u = |x|^r, r < 0$

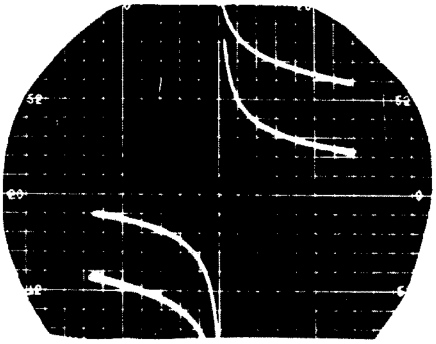


FIGURE II-10

Functions of the form

$u = \operatorname{sgn} x |x|^r, r < 0$

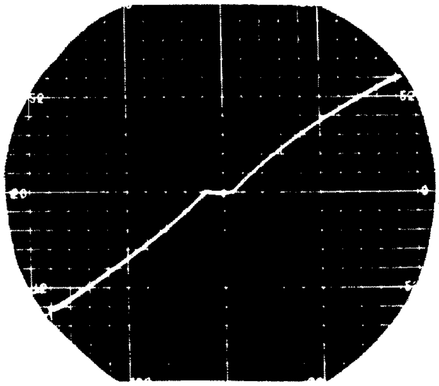


FIGURE II-11

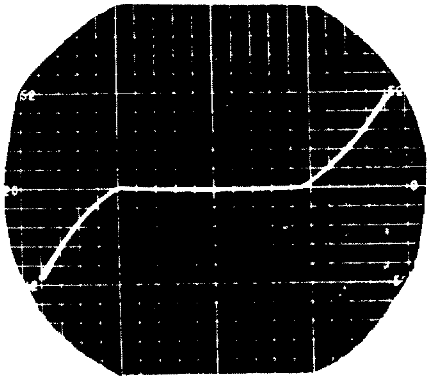


FIGURE II-12

Functions with dead space

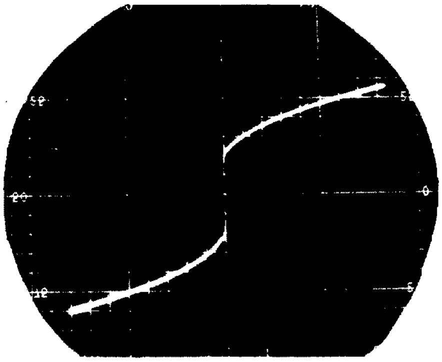


FIGURE II-13

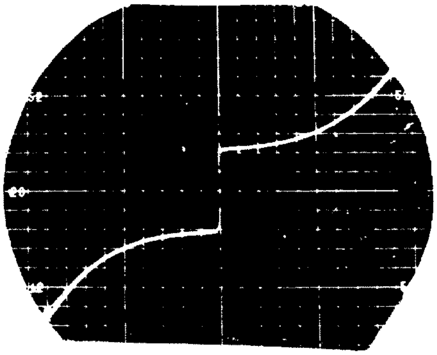


FIGURE II-14

Discontinuous Functions

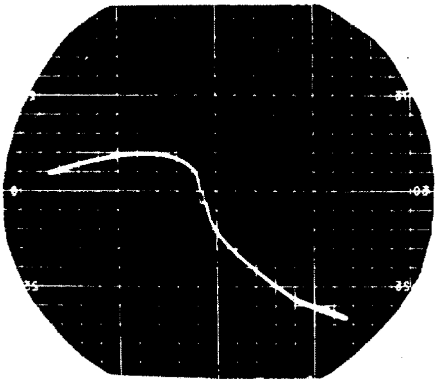


FIGURE II-15

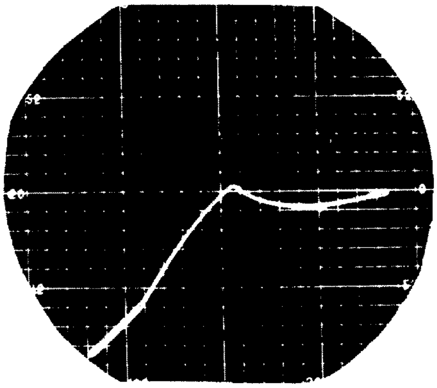
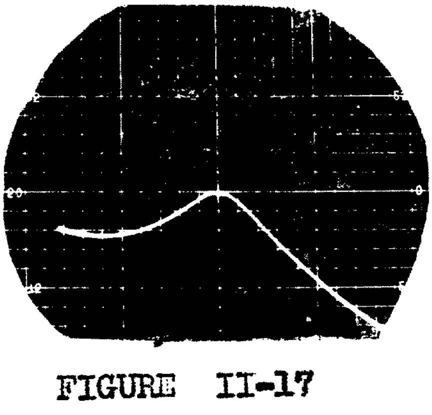


FIGURE II-16



Some double factor functions

FIGURE II-17

III. TWO EXAMPLES OF MODELING: THE HYDRAULIC SERVOMECHANISM AND THE SUPERSONIC DIFFUSER

A. Introduction

The purpose of modeling is to simulate the system under study in an environment such that testing is more controllable and more simply instrumented than the natural environment. By modeling, one can sometimes extract from a complex affair a fundamental factor and determine its independent characteristics. Occasionally the model is quite simple and entertains conclusions by direct physical intuition. Based on the means of modeling, a rough categorization of modeling is 1) environmental methods, 2) scaling, 3) physical analogs, and 4) mathematical analogs.

Environmental techniques make no changes in the model (usually some type of operating device) but provide controlled environment. Centrifuge testing of human beings and various temperature, vibration, and shock testing of equipment fall in this category.

Scaling techniques maintain the same basic phenomenon but alter physical dimensions or characteristics to obtain more convenient test requirements. These techniques may utilize actual environments, as for example captive flight tests of an aerodynamic model, or may use controlled environments such as in wind tunnels, towing tanks, and centrifuge tests of monkeys.

Inventing physical analogs of natural phenomena is an innate ability of man. With varying degrees of exactness, techniques have been used by poets, painters, and scientists since their respective historical beginnings. Although scientific history in particular appears to be a heap of discarded models, one cannot assume the general approach has fallen into disuse; it is an engineering necessity. A useful example is the electrical analog, where current and voltage across an electrical element are likened to forces and velocities of a mechanical element, or to pressure and flow of a hydraulic or acoustic element, or to temperature and heat flow of a thermal element. Once the coupling between the elements is defined, elaborate systems can be assembled by simple correspondence. Boundary conditions between elements or blocks of elements are automatically fulfilled. Many other ingenious forms of analogs have been used for engineering problems. A ball in a salad bowl has analogous motion to the pitch-yaw coordinates of a homopolar missile, and by relating the local slopes of the bowl to the stability derivates, nonlinear dynamic responses are directly obtained (Ref. 28).

Mathematical analogs imply that elements at least of the phenomenon can be described by differential equations. Various computing devices may then be employed to perform these mathematical operations.

If the analog is selected as the means of modeling, the problem director first musttranslate the important features of the real system to equivalent analog terms. This appears to be a step so obvious it warrants no further discussion, but appearances are deceptive. It is the most critical step of the entire experimental procedure; it must be handled logically and carefully. Overlooking one innocently small factor can invalidate the entire analog solution. The inclination to concentrate upon the end result, a convenient analog, arbitrarily simplifying features of the system to fit, is attractive but dangerous. The important and difficult task is not evaluating what remains as the simplified model, but evaluating what is excluded from the model.

The present discussion makes no pretense of supplying a universal translation procedure; of course, no such procedure exists. Two systems are treated, purely as examples, to demonstrate the pitfalls and the potentialities of electrical analogs. The first example, the hydraulic servomechanism, is selected because it often has been subject to abuse by modelers. Most of the discussion is directed at pointing up some of the important features of the real system and the modeling compromises which have to be evaluated.

The second example, the ramjet diffuser, is included because it demonstrates that if one accepts (and justifies) a set of compromises leading to a simplified model, the electrical analog can be a powerful means of analysis. Elements of the system can be modeled independently. The mathematically difficult problem of matching boundary conditions between elements is automatically performed by analog couplings. An element whose influence on the overall system is unknown can be readily evaluated. Furthermore, unlike analytical treatments, the analog accepts nonlinearities as readily as linearities.

B. Modeling a Hydraulic Servomechanism

1. General

The hydraulic servomechanism is selected as an example of intentionally closed loop systems primarily for its deceptive simplicity in physical appearance coupled with its extreme complexity in operation. It is true that some analyses have been made viewing the hydraulic system as linear (Refs. 29 and 30), but the assumptions included in such simplifications put these analyses in the academic class with mathematical analyses of second order nonlinear servomechanisms. Two approaches have been made toward modeling hydraulic systems. The first involves overall system response tests and subsequent synthesis of the responses by a defined electrical network. There is rarely a one-to-one correspondence between an electrical component and an isolated hydraulic component. The second method involves defining an analog for each of the hydraulic components and then assembling these analog components by direct correspondence to simulate a given system (Ref. 31). The second method is obviously more versatile and more useful in associating improvements noted on the analog with changes in physical components. It is also less reliable unless careful overall system tests are made to insure that the assembly corresponds to the hydraulic assembly.

There is no generalized hydraulic system. Each system is a specialized design intended for a specialized application, and its performance and peculiarities principally have meaning only for this system. Furthermore, subclassifying hydraulic systems is of little assistance, for there is no generalized aircraft hydraulic autopilot, nor missile autopilot, nor hydraulic industrial control. Unless the investigator undertook to test (for example) a number of missile autopilot systems, each operating under a number of typical conditions, little generality could be claimed for the results. Even here, extension of conclusions to a new system is risky. Fortunately, the components of the hydraulic system are more nearly standard. Hydraulic lines, double acting spool valves, diaphram or piston-cylinder accumulators, constant pressure pumps, piston-cylinder actuators, and couplings such as tees, crosses, elbows, and expansion and contraction junctions are components employed by a large number of servomechanisms. Therefore, the present brief investigation is aimed at a critical examination of some of the components and their analogs, rather than a detailed study of a specialized assemblage of components.

At least two external systems couple into the hydraulic servomechanism and often in such a manner as to have a profound effect on its characteristics. The first is the input system. Due to friction and unbalanced pressures, the hydraulic servo valve is not an irreversible device. The dynamic characteristics of the mechanical, electrical, or secondary hydraulic system inserting an input into the control system should usually be included in modeling. One type of aircraft hydraulic control is a hydraulic servomechanism which moves a control surface to match a mechanical input inserted by the control stick (by means of a cable or push rod system). It has been shown in tests that the difference between fixing or freeing the supposedly isolated stick control in the hydraulic system can be the difference between a stable and an unstable loop. The second external influence upon the servomechanism is its load. The single degree of freedom linear mass-spring-dashpot load attached to the hydraulic actuator avoids reality. Consider for example the loading on the aforementioned aircraft hydraulic control system. First, there is a nonrigid coupling between the actuator and the control surface. Second, there is the mechanics of the control surface, usually involving at least two dynamic modes and hinge friction. Third, there is the aerodynamics of the control surface, which is profoundly nonlinear for any surface which is efficiently balanced. Fourth, there is a nonrigid system supporting both the actuator and the hinge of the control surface which is subjected to reactions from the actuator as well as to warpages generated by aeroelastic effects. Fifth are disturbances (gust loading effects) influencing most of the above factors. Elimination of any of these factors in the model requires careful analysis and experimental justifications.

Inside the hydraulic servo system, the components discussed below appear reasonably frequently.

2. Hydraulic Lines

The hydraulic line is a pipe commonly of circular cross-section used to transport hydraulic energy. Static hydraulic pressures are often high (1000 to 3000 psi). With certain assumptions, the electric transmission line can be considered analogous to the hydraulic line, thereby providing a convenient model of this component. Correspondence between hydraulic pressure (p) and mass flows (velocity • area • density, ρuA) and electric voltages (v) and currents (i) are examined here to determine the limitations of this model.

The inertial term of the flow is isolated by assuming, for the moment, that the fluid is frictionless and the tube walls are rigid. Thus, for a pipe running in the X direction, by Newton,

$$A\frac{\partial \rho u}{\partial t} = - A\frac{\partial p}{\partial x} = \frac{\partial \rho u A}{\partial t}$$

which corresponds favorably with the transmission line analog, where

$$L\frac{\partial i}{\partial t} = -\frac{\partial v}{\partial x}$$

L being the inductance per unit length.

The density, ρ, of the hydraulic fluid varies slightly due to temperature variations and entrapped vapors, but these effects are insignificant compared to the other compromises of the analog.

Now assume the fluid is viscous and maintains a steady state flow rate. A drag force will exist at the wall and for laminar flow is proportional to the gradient of the velocity profile at the wall. For low Reynolds numbers, the flow introduced into a typical hydraulic tube in a short distance downstream assumes a velocity profile which (as long as the flow is laminar) has an invariant shape, but its amplitude varies with mean velocity. Thus the wall slope and consequently the drag is directly proportional to velocity:

$$Ku \doteq A\rho\ Du = -\frac{\partial p}{\partial x}$$

a situation analogous to the transmission line term

$$Ri = -\frac{\partial v}{\partial x}$$

where R is the electrical resistance per unit length, and D is the corresponding fluid friction constant dependent upon viscosity and tube sectional geometry. This portion of the analog is pleasantly simple, but unfortunately it is not always accurate. First, the laminar flow assumed above occurs only at low flow rates. As flow rate increases, first the flow experiences transition (typical fluid resistances per unit pipe length are shown in Fig. III-1), where resistance characteristics are fairly unpredictable. Further increase in flow rate sponsors turbulent flow, which exhibits a predictable but highly nonlinear resistance. Second, viscosity, the mechanism of fluid resistance, is markedly affected by temperature. An accurate analog would include in the transmission line distributed resistances which were constant for small currents and properly nonlinear for high currents. If experimental tests show appreciable temperature fluctuations during, say, transient tests of a system, these may also be important in the analog.

Now assume the pipe full of fluid is being tested for various static pressures. Compressibility of the fluid and elasticity of the wall allow fluid to be stored in the pipe as pressure is increased. If the pressure is increased at some slow fixed rate, a flow rate is observed into the pipe; that is,

$$E\frac{\partial p}{\partial t} = -\frac{\partial(\rho u A)}{\partial x}$$

which is analogous to electrical capacity to ground in a transmission line,

$$C\frac{\partial v}{\partial t} = -\frac{\partial i}{\partial x}$$

Here, E is the fluid capacitance per unit length, and C the electrical capacitance of the analog.

Fluid compressibility and wall elasticity are reasonably linear. Entrapped vapors cause significant (and unpredictable) variations in fluid capacitance. Furthermore, although fluids compress, they do not stretch. It is possible in a low pressure return line to experience imposed negative pressure pulses of the same magnitude as the static pressure. The result is a form of cavitation; separation of the fluid in the tube. If we assume cavitation is distributed uniformly, this can be interpreted in the fluid equations, below some low total pressure, as nonlinear variations of inertial, friction, and capacitance terms. In the analog, this effect might be simulated by insertion of coupling elements along the transmission line which are inactive above some small total pressure and obey appropriate gas laws below this pressure.

Since no distributed leakage occurs along the hydraulic line, the corresponding electrical leakage term of the transmission line is omitted.

Nonlinear coupling terms must be ignored in the simple analog. For example, the effect of fluid acceleration upon velocity profile is ignored. Intuitively, one would expect that a step jump in velocity would not result in an instantaneous establishment of a new steady state velocity profile. Radial flows sponsored by tube stretch similarly affect profiles. The importance of these nonlinear, dynamic couplings has not been evaluated experimentally.

Combination of the constant coefficient equations results in the wave equation, showing that under certain assumptions pressure and flow disturbances in a hydraulic tube propagate much in the manner of electrical disturbances along a transmission line. The customary model of the distributed constant transmission line is a lumped constant line in which short sections are simulated by lumped elements (T or π sections). If these lumped constant sections are of the order of eight wavelength sections, the approximation introduces less than 10% error. The effects of introducing pronounced nonlinearities into the lumped section analog (compared with introducing the same nonlinearities as distributed constants) probably must be determined experimentally for each variety of nonlinearity introduced.

3. Double-acting, spool valves

Fluid flow into the actuator is often controlled by a double acting spool type hydraulic valve (Fig. III-2). In simple systems, the housing is attached to the actuator output (position feedback) and the spool is moved by the input. As mentioned before, the valve is not unilateral as seen by the input.

The static performance of a valve is the relation of three variables: pressure drop across the valve, flow rate through the valve, and spool position relative to its housing. One manner of displaying this performance is flow rate vs. pressure drop characteristics for various spool positions (z variable). An examination of valve characteristics for a fairly wide assortment of spool shapes (a typical characteristic is shown in Fig. III-3) indicates that pressure drop as a function of flow rate, spool position being constant, can usually be written as

$$\Delta p = K_{z_1} (w)^r$$

where K_{z_1} is a constant associated with a particular spool position z_1, w is the mass flow rate, and the exponent, r, lies between 1.5 and 3.0. Thus the static characteristic for a constant z_1 cannot be intentionally molded to any great extent, since only small variations in the exponent are available.

Rewriting the characteristic as

$$\Delta p = K_{w_1} \cdot f \{z\}$$

where K_{w_1} is a constant associated with a particular mass flow, w_1, we get a set of curves such as those shown in Fig. III-4. The function, $f\{z\}$ can be approximately predicted from the manner in which the spool, moving with x, unports the valve orifices. A bevel (or a radius) on the spool edge (shown dotted in Fig. III-2) alters the function $f\{z\}$. If an analog study indicated that some $f\{z\}$ gives desirable responses, this function might be physically realized by correlation with a spool design consisting of multiple bevels (smooth functions would not be practicable to fabricate), and perhaps orifice shaping.

The electrical analog of the hydraulic spool valve is not simple. As an example of extreme effort to maintain a linear model, references 29, 30 and others utilize the relation

$$w = C_1 \cdot z$$

with the corresponding electrical analog as a current generator whose output is proportional to z. This involves the unlikely assumptions that 1) Δp is constant and 2) $f\{z\} = C_2(z)^{-r}$. The former assumption is invalid for any well designed servomechanism having a real load, and the latter rigidly constrains spool design. A much improved relation is

$$w = g\{\Delta p\} \cdot z$$

where $g\{\Delta p\}$ is an arbitrary function of the approximate form $(\Delta p)^{1/r}$. This function can be modeled by a current generator controlled by the product of z and the output of a function generator forming $g\{\Delta p\}$. Spool design is defined by the selected function $g\{\Delta p\}$, which itself is only slightly formable.

The heart of the hydraulic servomechanism is its control valve; thus one would suspect spool design is a critical factor of the system. The electrical analog should provide the opportunity for testing, easily, a large number of $f\{z\}$ functions. To do this, the static valve characteristic is modeled, in the simplest form, as

$$\Delta p = f\{z\} \cdot (w)^r$$

The corresponding electrical analog might consist of a voltage amplifier whose input is the product of two arbitrary function generators. It can also be produced directly by the special function generator described in Section II with an easily altered function, $f\{z\}$ of the form

$$f\{z\} = \text{sgn}\,(z-a)\,|z-b|^r\,\text{sgn}\,(z-c)\,|z-d|^q$$

Utilizing this function generator and noting that the inflow and the outflow characteristics may all be different, we can represent the electrical analog of the valve as shown in Fig. III-5. The voltages v_a and v_b correspond to pressures in the cylinder chambers (at the valve outlets). When the (±) z input, to the function generators are negative, the generator blocks flow.

The generalized valve requires four specialized function generators (or the equivalent assembly of multipliers and arbitrary function generators) which often are not available. Consequently certain simplifications of the model may be useful. If the two inflow characteristics are identical and the two outflow characteristics are also identical, the valve can be reduced to two function generators and a switch reversing with the polarity of z (as shown in Fig. III-6). In this model, the pressure drop across the function generator varies as the absolute value of z and increases as z decreases. It resembles the negative exponent functions such as shown in Fig. II-9. Families of pressure drops vs. flow rates produced by the function generator are shown for various values of r in Fig. III-7.

For models in which a symmetry about the center of the load (the dotted line of Fig. III-6) can be devised, a further simplification can be made. This symmetry implies 1) that the supply and the exhaust lines are identical and cavitation does not occur, 2) that inflow and outflow valve characteristics are all identical, 3) that the cylinder and load can be split into symmetrical sections, and 4) no leakages to ambient pressure occur. If these features can be justified experimentally, the system reduces to the single function generator and switch, as shown in Fig. III-8 by this argument.

If the system is symmetrical and the power supply pressure is p_s, the center point pressure is $p_s/2$. Reducing pressures throughout the model by $p_s/2$ puts the center point at ground (zero pressure), and pressures at symmetrical points differ only in polarity. Thus switch point d has the negative of the pressure at point a.

Hydraulic time lags in the valve are probably small compared to other system time effects. Valve characteristics vary with viscosity, and thus with temperature. Since the valve is not unilateral, its backcoupling into the input must be determined, and even without the benefit of experimental data it is a reasonable guess that this backcoupling is not linear.

4. Hydraulic accumulator

The hydraulic accumulator is a chamber a portion of which is spring loaded. For zero'th order analyses, the quantity of liquid in the accumulator is proportional to the fluid pressure, and thus its electrical analog (utilizing voltages and currents as related to pressures and flow rates respectively) is a large capacitor between the line and ground. The purpose of the accumulator is identical to that of a large capacitor in an electrical power supply: it smooths supply pressure fluctuations. If substantial entrance and internal flow rates occur, inertial and resistance terms should also be included in the dynamic model. Furthermore, kinetic energy error discussed later (Part 6, Couplings) is always present and cannot be modeled by any simple electrical analog.

5. Hydraulic piston-cylinder actuators

The hydraulic piston-cylinder actuator is perhaps the simplest and most common form of the hydraulic actuator. The piston, sealed by O-rings, chevron rings, or other means, rides in a cylinder generally sealed at both ends. Driving force is extracted from the piston by a rod brought out of one cylinder end through sealed bushings. One method of operation maintains a chamber on one side of the piston at constant pressure and produces controlled pressures above and below this constant value in the other chamber. Another method obtains differential pressure across the piston by filling one chamber and simultaneously exhausting (to ambient pressure) the other chamber. The first method has simpler valving; the second requires half the operating pressure to get the same force.

For modeling purposes we can assume the load complex has been highly and accurately subminiaturized and fitted between the piston and the cylinder. The forces acting on the load are produced by differential pressure across the piston. These pressures are determined by fluid dynamics in the chambers and the inlet conditions are established by the valve; similarly to the accumulator, the variable volume chambers can be represented for low mass flows by variable condensers and for high mass flows, by including variable inertial and viscous effects. As with the accumulator, kinetic energy errors must be tolerated. Leakage across the piston and around the piston rod bushing are both quite nonlinear effects and if included at all in the model should probably be put in as nonlinearities.

6. Couplings

Couplings such as elbows and flexible couplings can be treated in a manner similar to hydraulic lines. However, couplings involving expansion or contraction of cross-sectional area introduce a feature of hydraulics which is not included in the electrical analog. Change in cross-sectional area occurs where tubes vent into chambers, where tubes change size, and where flows split or combine (T's, crosses, etc.).

Consider a steady state flow rate in a tube of initial area A_1 and final area A_2. The expansion is here assumed to be lossless. By extended one dimensional theory, continuity tells us

$$(\rho uA)_1 = (\rho uA)_2$$

which is analogous to

$$i_1 = i_2$$

From energy,

$$\frac{1}{2} u_1^2 + \frac{p_1}{\rho_1} = \frac{1}{2} u_2^2 + \frac{p_2}{\rho_2}$$

which is obviously not a simple relation to model. To use the electrical analog, we must assume that $\rho_1 = \rho_2$ and that the stagnation pressure $(\frac{1}{2} \rho_1 u_1^2)$ is negligible compared to the pressures, p_1; therefore, $p_1 = p_2$. These are rather broad assumptions. For flow velocities of 100 fps, the impact head is of the order of 500 psi. Couplings such as lines dumping into a large chamber almost fully realize this head. This is the most fundamental and perhaps the most serious disadvantage of the electrical analog.

7. Analysis of approximations

A slightly more rigorous approach shows the curious manner in which the analog must be justified. We start with the usual hydrodynamic equations of motion:

From continuity - - $\dfrac{\partial \rho}{\partial t} + \dfrac{\partial \rho u_i}{\partial x_i} = 0$

From momentum - - - $\dfrac{\partial \rho u_i}{\partial t} + \dfrac{\partial \rho u_i u_j}{\partial x_j} = -\dfrac{\partial p}{\partial x_i} + \dfrac{\partial \sigma_{ik}}{\partial x_k} + \rho\, g_i$

From energy - - - $\dfrac{\partial \rho J}{\partial t} + \dfrac{\partial \rho u_i J}{\partial x_i} = \dfrac{\partial p}{\partial t} + \dfrac{\partial (u_i \sigma_{ik} + q_k)}{\partial x_k}$

where σ_{i_k} is the shear tensor, $\sigma_{ii} = 0$, J is the total energy term $\frac{1}{2} u^2 + h$ (h being the enthalpy), and g_i is the gravity term.

The model to which these equations are applied is a two dimensional configuration (Fig. III-9) with flexible walls, and consequently the linearized characteristic,

	$A(x,t) = A_o(1 + \alpha\, p)$	wall elasticity
and	$\rho(x,t) = \rho_o(1 + \beta\, p)$	fluid compressibility.

A_o and ρ_o are functions of x, and α and β are small.

First, we assume that temperature variations (q terms), gravity, and cross flows are all negligible. Second, we simplify the viscous term of the momentum equation:

$$\frac{\partial \sigma_{ik}}{\partial x_k} = \frac{\partial \sigma_{xy}}{\partial y} = \frac{\partial}{\partial y}\left(\mu \frac{\partial u}{\partial y}\right) = \frac{\partial \mu}{\partial y} \cdot \frac{\partial u}{\partial y} + \mu \frac{\partial^2 u}{\partial y^2}$$

For low Reynolds numbers with reasonably constant A, the flow will have a nearly parabolic velocity profile, $u = \bar{u}(c - dy^2)$, with constants such that $u = 0$ at the boundary. Therefore, $\frac{\partial^2 u}{\partial y^2} = -2\, d\bar{u}$, a function of the mean velocity, $\bar{u}$. Again imposing the requirement that temperature and thus viscosity variations are small,

$$\frac{\partial \sigma_{ik}}{\partial x_k} = -2\, d\, \mu\, \bar{u} = -K_1 \bar{u}$$

From here on, we deal only with variables averaged over y for any station x, and, for convenience dropping the bar notation, we can write the first two hydrodynamic equations as

Continuity - - - $$\frac{\partial \rho A}{\partial t} + \frac{\partial \rho u A}{\partial x} = 0$$

Momentum - - - $$\frac{\partial \rho u A}{\partial t} + \frac{\partial \rho u^2 A}{\partial x} = -\frac{\partial p A}{\partial x} - K_1 u$$

For the defined qualities of A and ρ, neglecting the second order term,

$$\frac{\partial \rho A}{\partial t} = A_o \rho_o (\alpha + \beta) \frac{\partial p}{\partial t} = K_2 \frac{\partial p}{\partial t}$$

Reducing continuity to the form

$$K_2 \frac{\partial p}{\partial t} = -\frac{\partial \rho u A}{\partial x}$$

This is equivalent to the assumed fluid capacitance relation.

The momentum equation should reduce to the desired resistance-inertial relation

$$\frac{\partial (\rho u A)}{\partial t} + D(\rho u A) = -A \frac{\partial p}{\partial x}$$

To accomplish this, first we must assume that A and ρ are essentially constant with x, and second, we must accept the unorthodox logic that although $\frac{\partial \rho u A}{\partial x}$ is worthy of consideration in continuity, $\frac{\partial \rho u^2 A}{\partial x}$ is negligible in momentum. In other words, kinetic energy in the analog is everywhere constant. It follows that the energy equation is disregarded. For hydraulic systems, the latter is the most serious and unjustified assumption.

C. Modeling a Supersonic Ramjet Diffuser

The supersonic ramjet diffuser is a typical closed loop, highly complex phenomenon, for which modeling is essentially a process of determining and extracting from the complete system some component. The particular final model of this demonstration (for which no experimental wind tunnel tests were specifically conducted to corroborate assumptions) should not be viewed critically as to its authenticity, the procedural technique being the point of the discussion.

The supersonic ramjet is an air-breathing, continuous flow engine which undergoes the usual combustion sequence of intake, compression, addition of heat, and expansion. Fig. III-10 shows a typical cross section of this mechanically simple device in operation. Observed in body coordinates, the air approaches the ramjet (region 0) undisturbed and at some Mach number in excess of one (very likely greater than 1.5). The air entering the diffuser is supersonically compressed through a series of oblique shock waves sponsored by the cone and inner body and then at some point 2 becomes subsonic, expanding into region 3. This enclosed region up to the flame holder is sometimes termed a "plenum chamber", but one should not be misled into visualizing this region as a placid reservoir of compressed air. Imposed upon the average velocity of air through this region are considerable eddies, cross-flows, and fluctuations cuased by boundary layer peeling off the inner body, asymmetric inlet conditions (engine at angle of attack), and other factors. At station 4, fuel is injected and aereated until station 5 where, conditions being proper, flames are held in the stream by flame holders. Between 5 and 6 the fuel is burned and the hot expanding gas mixture is vented to the ambient air through the customary expansion nozzle.

One of the least understood classes of phenomena associated with the ramjet can be loosely grouped as instabilities affecting burning. These burner instabilities are extremely important, since they often incite flameout, i.e., engine disability. At least three somewhat distinct fields are involved in the ramjet; aerodynamics, aero-thermodynamics, and hydrodynamics (pertaining to the fuel system up to injection). Thus reasons for all sorts of possible instabilities can be invented or deduced. The pressure boundary at the injector allows air pressure external to the injector to affect fuel flow, and fuel flow later affects external pressure. Angle of attack of the ramjet influences mass flow, mass flow relates to thrust, and thrust, at least in an altitude stabilized vehicle, influences angle of attack through velocity. The instabilities of flames burning in a steady state draught and subject to solid boundaries are so numerous and complex they form a complete branch of scientific research. Reactions of flames to nonsteady flows such as those initiated by the aerodynamic compressors have scarcely been touched on. Several kinds of aerodynamic instabilities can occur in and forward of the diffuser. Furthermore, all of the instability phenomena are nonlinear, pretty well invalidating the accurate synthesis of an operating ramjet instability by superposition of various component instabilities. This by no means minimizes basic investigations. A knowledge of the physical mechanisms provoking various component instabilities can sometimes lead to appropriate designs reducing or eliminating these particular instability tendencies.

The present modeling is concerned with a specific aerodynamic instability. Its extraction from the general problem follows this argument (Ref. 32). Several experiments have shown that diffuser instabilities can occur without burning (and thus, are aerodynamic) and that these instabilities can be influential enough in an engine to blow out the flame. One mechanism sponsoring instabilities has been attributed to high mass flow charging and discharging of the "plenum chamber." The instability is evidenced by fairly infrequent, somewhat random pulsing of the flow through the diffuser. Another lower amplitude instability is evidenced by periodic (generally above 100 cps) oscillations. These occur both at high mass flow and at low mass flow with choked diffusers. Either of the instabilities requires subcritical diffuser conditions: i.e., the flow becomes subsonic some place ahead of the diffuser inlet, allowing internal disturbances to propagate forward, thus influencing the shocks and flow into and around the diffuser.

The high frequency instability appears to be a condition independent of the large amplitude charging phenomenon, and it is periodic. Therefore, it is here extracted for modeling.

One further subdivision of the high frequency instability is possible if certain arguments are accepted. Dailey's cold flow tests (Ref. 33) of a complete diffuser (including nose cone and inner body) revealed a tendency of the aerodynamic instability to resonate at some "organ pipe" mode of the diffuser tube, and the particular mode was always one which lay within a definite frequency band (470 to 940 cps). Furthermore, the frequency band seemed to be dependent upon Reynolds number rather than upon diffuser length. Dailey deduced from this, not without experimental justification, that the instability was due to viscous forces occurring in and around the diffuser entrance, and through coupling with a tuned aft tube, it developed approximately the frequency of the organ pipe mode nearest that of the forcing function. Tests performed at the California Institute of Technology by Stoolman have shown that such a viscous type of instability exists, even in a completely blocked diffuser (zero length pipe) and that again the instability is periodic (400 to 600 cps). However, an entirely different type of aerodynamic instability can be readily produced with the pipe alone (without cone or inner body) which is periodic over a large range of model dimensions at the fundamental mode of the pipe. It seems reasonable to assume that this instability is not due to viscous effects about the tube lip, for two reasons: 1) frequency is dependent upon tube length rather than Reynolds number, and 2) lip and diffuser shapes varying from a wall normal to the tube axis to a thin, sharpened lip such as the one shown in Fig. III-11 all seem to produce the same instability at approximately the fundamental longitudinal organ pipe mode (Ref. 32). It is entirely possible that this basic type of instability could also exist in the complete diffuser. An alternative explanation of Dailey's results is that both high frequency instabilities can occur with subcritical conditions, but both are nonlinear phenomena and the nonviscous instability is obviously more versatile in frequency range. A combination instability results, and the single resultant frequency is an entrainment of the two nonlinear instabilities. In practice, such deductions would be specifically checked by experiments. Here, without ado, we extract the nonviscous phenomenon from the more general problem and proceed to the modeling of this specific instability.

By eliminating viscous effects at the diffuser entrance as mechanizing the instability, we place responsibility for the affair upon the shock wave ahead of the diffuser. Consider Fig. III-11. A disturbance starting at, say, the diffuser lip (as evidenced by pressure, density and velocity changes) and aimed downstream propagates aft at a wave speed associated with the sonic velocity in the tube and the inner air velocity. At the exit and it is reflected, its new magnitude depending upon the mismatch of the exit boundary conditions. The reflected forward travelling wave is partially reflected at the lip and partially propagated forward externally. This latter portion bounces off the back of the shock wave and re-enters the tube. For instability to occur, the shock and the flow region ahead of the tube must have the properties of an aerodynamic amplifier, reinforcing in some manner the impinging disturbance.

We can reasonably well define the dynamic characteristics of the pipe and also the conditions across a nonsteady shock wave. The key to the instability (if it is a shock phenomenon) lies in the flow region between the shock and the diffuser entrance. This is the region of interest.

There is no obvious reason for the instability, which is closely associated with the longitudinal mode of the pipe, to be peculiar to three dimensional axially symmetric configurations within the diffuser, so here we assume (following an approach used by Ref. 32) that the internal model is one dimensional. (Fig. III-12.) Furthermore, wind tunnel tests show that the portion of the shock directly ahead of the tube (the portion one would expect to have significant influence on the instability) is almost normal; consequently, we model the shock wave as a normal shock.

A convenient break in the system occurs at the diffuser lip, separating the model into an axially symmetric forward section containing the shock and the external diffusion region and a one dimensional aft section representing the internal region of the tube. The aft section is, by itself, similar to an organ pipe with one end open and the other partially blocked. At its fundamental mode, the tube's length is approximately a quarter wavelength; that is, a complete cycle is described, for example, by a compression wave starting at the lip and making a round trip, being reflected at the forward open end as a rarefaction wave which then makes a second round trip and is reflected again at the lip as a compression wave beginning the second cycle. The reflections at the nozzle always include losses. Therefore, if an instability is to occur, the forward section must not only properly reflect disturbances but also reinforce them.

First we look at the aft, diffuser tube section. During natural oscillation, the shock oscillates at the same frequency about some position, ξ_o. Noting that the shock is an entropy generator, we deduce that nonsteady motions of the shock cause entropy fluctuations which drift downstream as properties of the air particles. A portion of the air is captured by the tube, and the included entropy fluctuations continue through the internal part of the tube. If a is the velocity of sound in the tube and U is the air internal velocity, the pressure-velocity disturbances propagate downstream with velocity a + U and upstream with velocity a - U. Since the tube length is roughly $\lambda/4$, the approximate oscillation frequency is $f = a/\lambda = a/4\ell$ (assuming U is small compared to a). Entropy waves propagate downstream only and at velocity U (which is of the order of 0.1 a). Two or three entropy cycles exist in the tube length, ℓ.

If it is assumed that the disturbances are propagated without losses, that their magnitudes are small compared to the average conditions in the tube (the waves are sound waves), that the flow is one dimensional, and that the gas is perfect, certain relations can be deduced (Ref. 32). We use the notation:

Total velocity = U + u

Total pressure = P + p

Total density $= \varrho + \rho$

Total entropy variable = H + h

where the capitalized symbols are the steady state values and the small symbols, the disturbances. h is an entropy variable, $Re^{\Delta s/c_v}$. (R is the gas constant, Δs the change in entropy, and c_v the specific heat at constant volume.) Velocity disturbances satisfy the wave equation and can thus be expressed as

$$u/U = f_1 \left\{x-(a+U)t\right\} + f_2 \left\{x+(a-U)t\right\} = \frac{u_r}{U} + \frac{u}{U} \qquad \text{III-1}$$

Pressure (actually, dp/dt) is coupled to velocity, giving

$$p/P = \gamma M(f_1 - f_2) \qquad \text{III-2}$$

where M = U/a. Entropy propagates downstream only:

$$h/H = f_3 \left\{ x-Ut \right\} \qquad \text{III-3}$$

Density couples both to the sonic disturbance and to entropy,

$$\rho/\varrho = M(f_1 - f_2) - \frac{1}{\gamma} f_3 \qquad \text{III-4}$$

Therefore, a given right running wave, f_1, propagates unchanged until it meets the exit boundary. Its transit time from lip to exit is $\Delta t_1 = \ell/(a+U)$. At the boundary it reflects as a new right running function, f_2, which propagates unchanged back to the lip. Its transit time is $\Delta t_2 = \ell/(a-U)$. The boundary condition at the exit, or nozzle (Ref. 32), is that the exit Mach number is constant, i.e.

$$u/U = a'/a$$

where a' is the disturbance effect on the speed of sound and can be defined from the properties of the fluid,

$$\begin{aligned} a'/a &= \frac{1}{2}\left[(\gamma - 1)\, p/P + h/H\right] \\ &= \frac{1}{2}\left[(\gamma - 1)\, M(f_1 - f_2) + \frac{1}{\gamma} f_3\right] \end{aligned}$$

Consequently,

$$u/U = f_1 + f_2 = \frac{1}{2}\left[(\gamma - 1)M(f_1 - f_2) + \frac{1}{\gamma}f_3\right] \qquad \text{III-5}$$

If f_3 and f_1 are defined at the exit, we can immediately determine the reflected disturbance, f_2. The model of the diffuser tube can thus be devised (Fig. III-13), with inputs f_1 and f_3 and an output f_2.

Before attempting to model the forward external section of the system, it is instructive to consider the fictitious case where the steady state shock is at the lip ($\xi = 0$). All the flow in the infinitesimal forward region is thus one dimensional. We wish to determine the reflection conditions looking from the lip section forward to the backside of the shock.

A nonsteady shock having some velocity, $\dot{\xi}$, produces changes compared to a steady state shock in pressure, velocity, and entropy immediately behind the shock. These changes can be determined by the Rankine-Hugoniot relations (as done in Ref. 32) and will simply be designated here as known functions:

$$u = f_4 \{\dot{\xi}\}$$
$$p = f_5 \{\dot{\xi}\}$$
$$h = f_6 \{\dot{\xi}\}$$

Now from equations III-1,2

$$u = u_r + u_\ell = f_4\{\dot{\xi}\}$$

$$p = \frac{p\gamma M}{U}(u_r - u_\ell) = f_5\{\dot{\xi}\}$$

The input (the known variable) here is u_ℓ, which leaves us with two equations and two unknowns. The corresponding model solution is shown in Fig. III-14. If the relation between f'_ℓ and $\dot{\xi}$ is monotonic, the infinite gain amplifier guarantees that the error between f'_ℓ and f_ℓ will be zero by controlling $\dot{\xi}$.

Now let us attempt a similar approach to the situation where the shock is ahead of the diffuser lip, and axially symmetric flow exists. The transit time of a particle across the external diffusion region is quite small compared to the period of oscillation, so we introduce the assumption that the flow field between the shock and the lip is established instantaneously by the position of the shock, ξ, and the shock velocity, $\dot{\xi}$. This quasi-stationary assumption allows a direct means of determining p and u at the lip section as functions of ξ and $\dot{\xi}$. If the freestream velocity (ahead of the shock) is U_o, a nonsteady shock at instantaneous position ξ and velocity $\dot{\xi}$ produces the same instantaneous flow conditions as a steady shock at ξ produces with corrected freestream velocity, $U + \dot{\xi}$. Thus static wind tunnel measurements[*] of p, u, and ρ at the diffuser lip for various freestream velocities and shock positions (as varied by nozzle choking) provide the functions

$$\begin{aligned} u &= f_7 \{\xi, \dot{\xi}\} \\ p &= f_8 \{\xi, \dot{\xi}\} \\ h &= f_9 \{\xi, \dot{\xi}\} \end{aligned}$$

The resultant model of the system is shown in Figure III-15. As before, if u'_ℓ is a monotonic function of the error signal, the amplifier guarantees that $u'_\ell = u_\ell$ by adjusting $\dot{\xi}$.

If such a model is successful, it becomes a powerful supplement to actual wind tunnel testing. Here, all important variables and functions are directly available and controllable. For example, the isolated influence of each of the assumed functions f_7, f_8, and f_9 can be readily explored in the model.

[*] These measurements can be made up to the point where shock instability occurs. One would expect the functions to be reasonably smooth, and thus a certain amount of extrapolation into the unstable region is permissible.

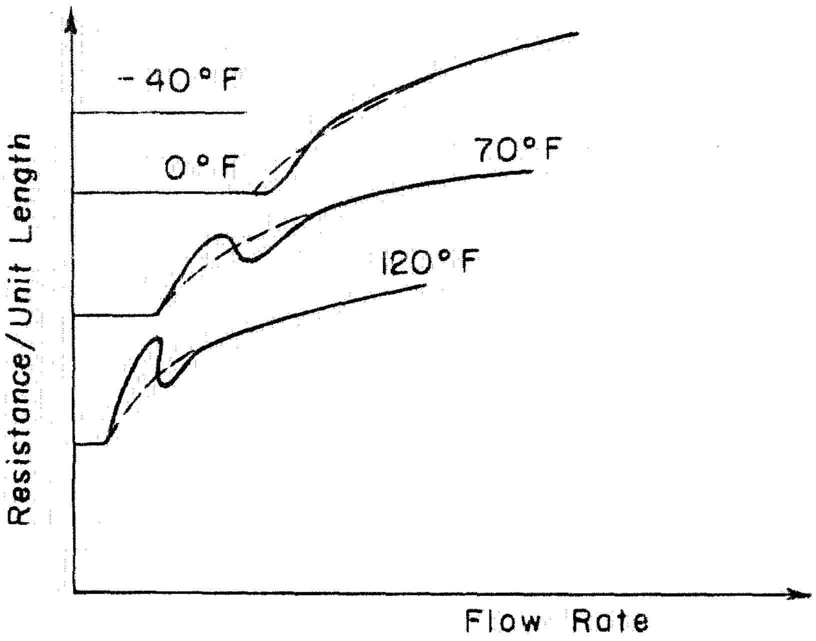


FIG. III - 1 RESISTANCE OF A HYDRAULIC LINE

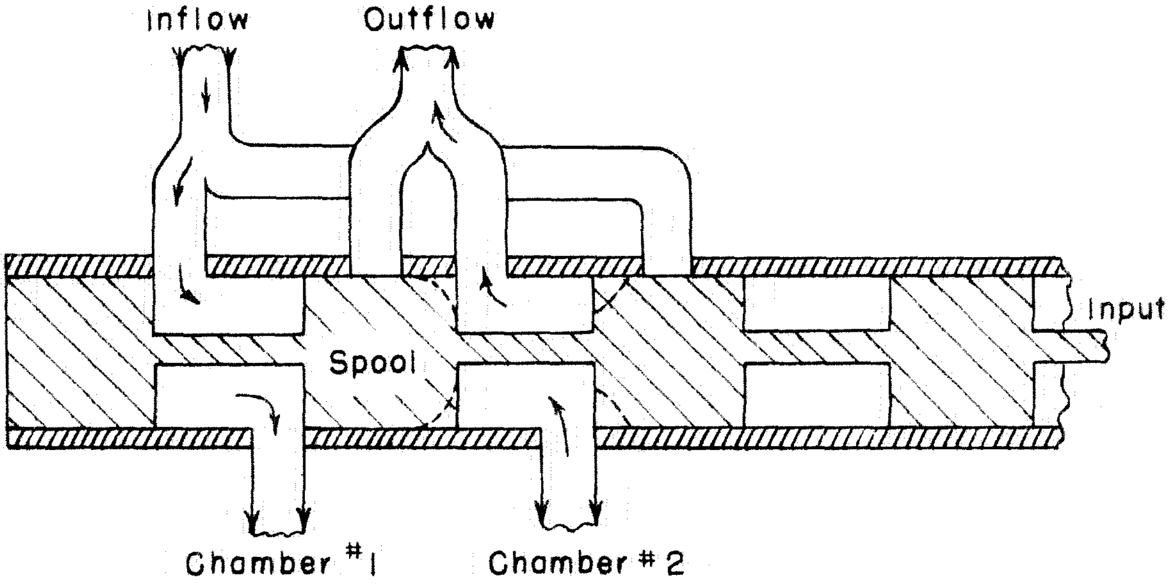


FIG. III - 2 DOUBLE ACTING SPOOL VALVE

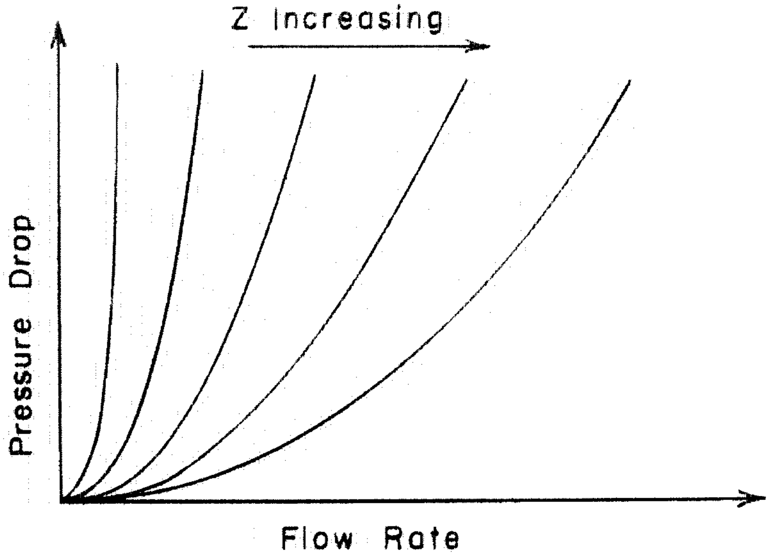


FIG. III-3 HYDRAULIC VALVE CHARACTERISTIC

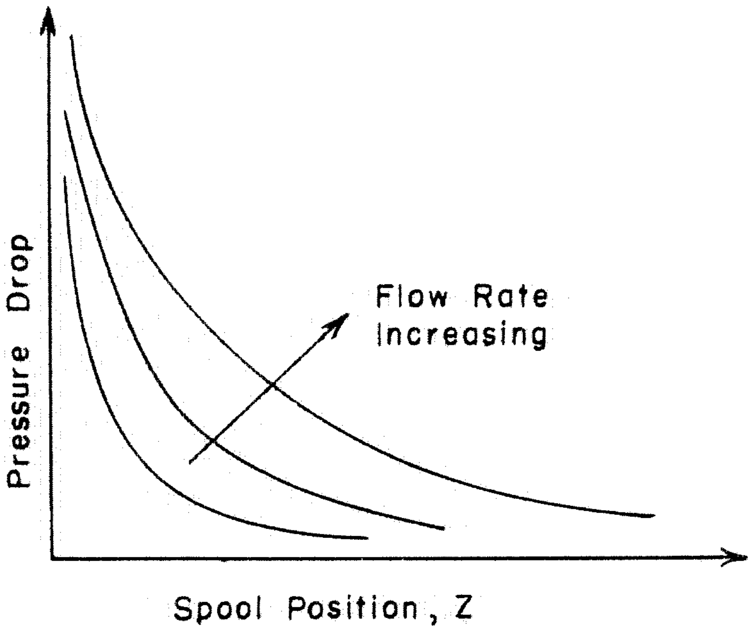


FIG. III-4 HYDRAULIC VALVE CHARACTERISTIC

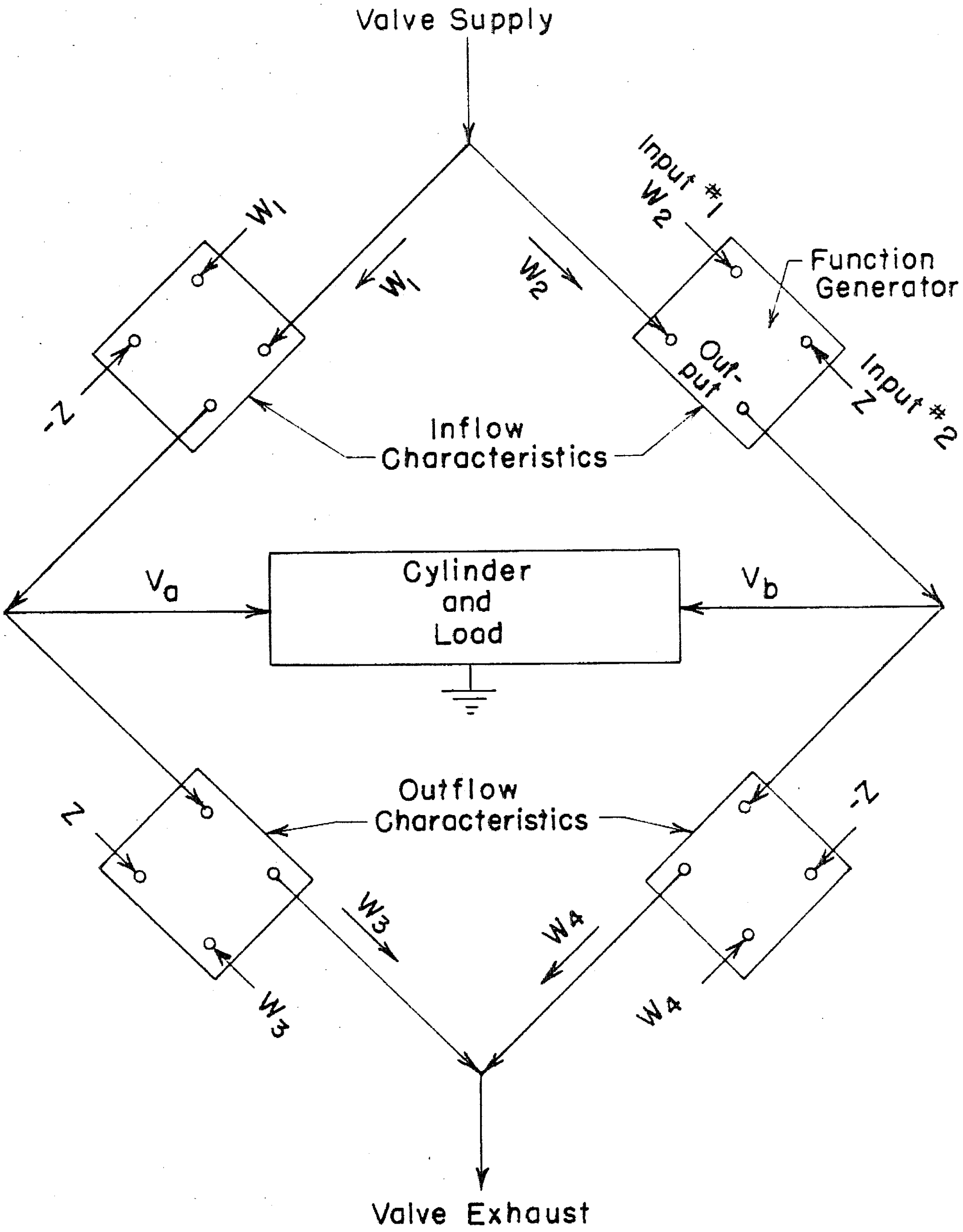
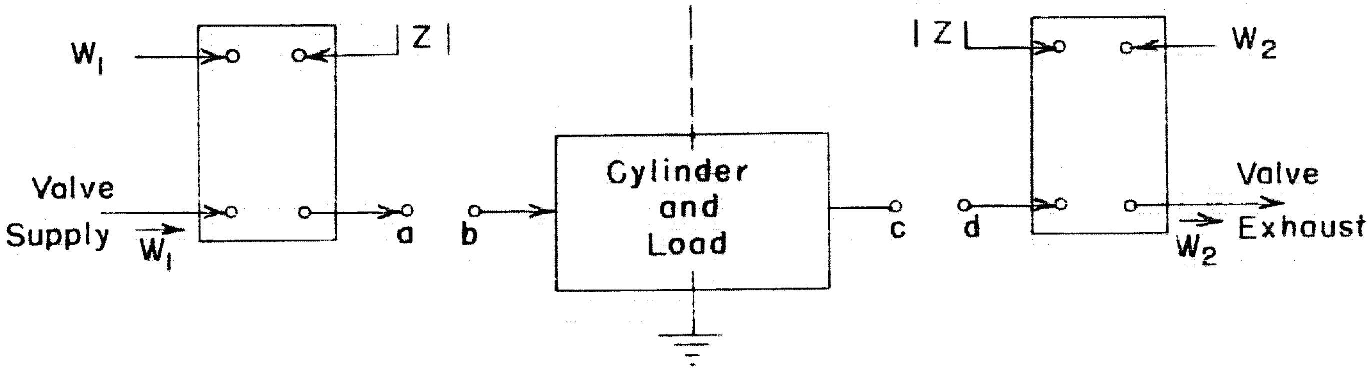
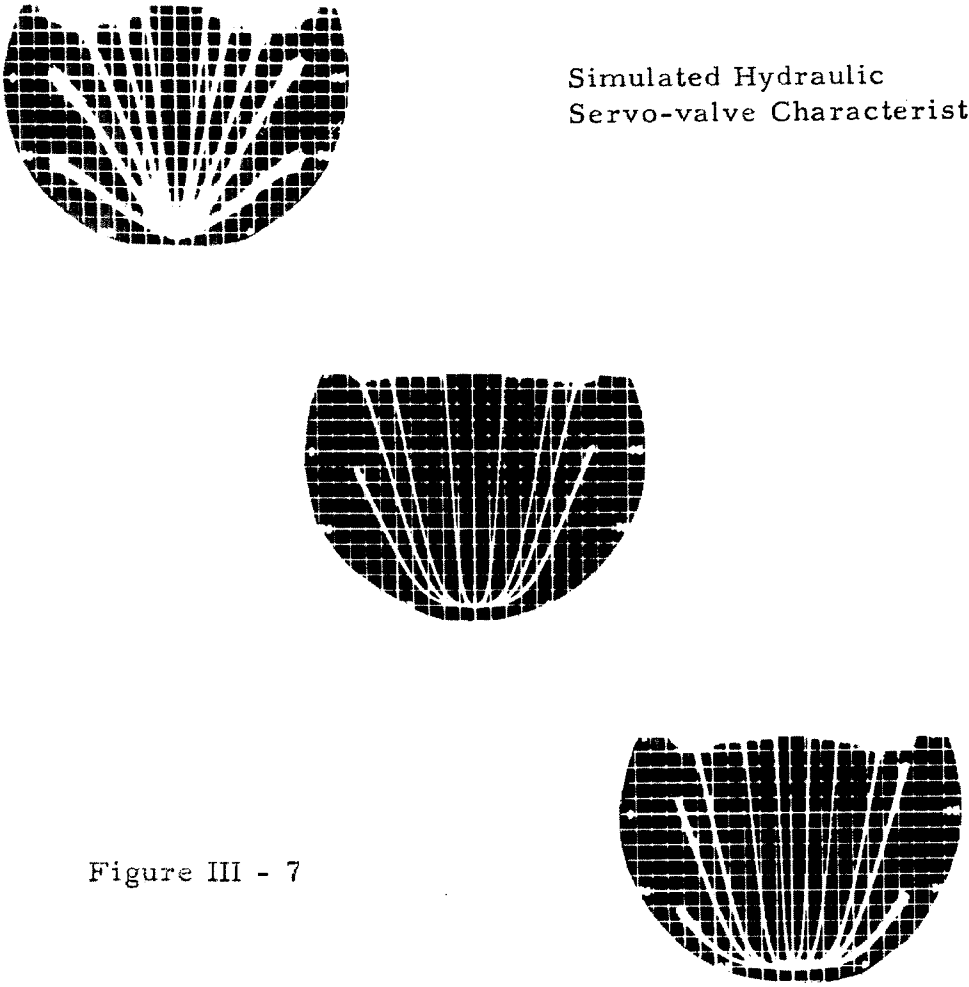


FIG. III - 5 HYDRAULIC VALVE ANALOG



Z > O Connects a-b, c-d : Z > O Connects a-c, b-d

FIG. III - 6 SIMPLIFIED VALVE ANALOG



Simulated Hydraulic Servo-valve Characteristics

Figure III - 7

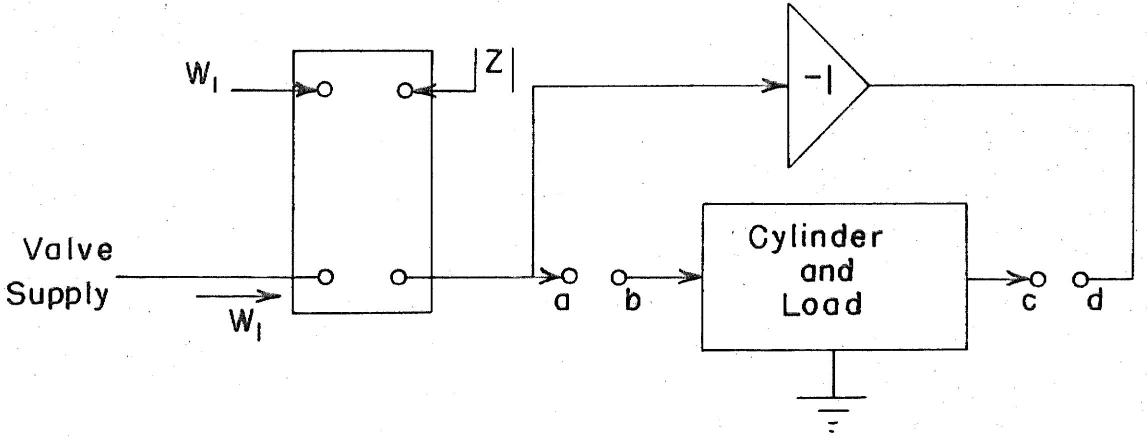


FIG. III - 8 FURTHER SIMPLIFIED ANALOG

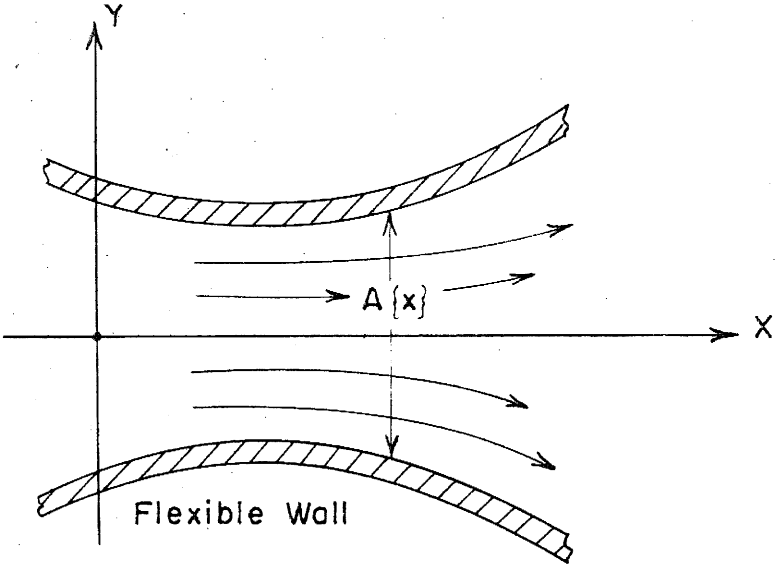


FIG. III - 9 EXTENDED ONE-DIMENSIONAL HYDRAULIC MODEL

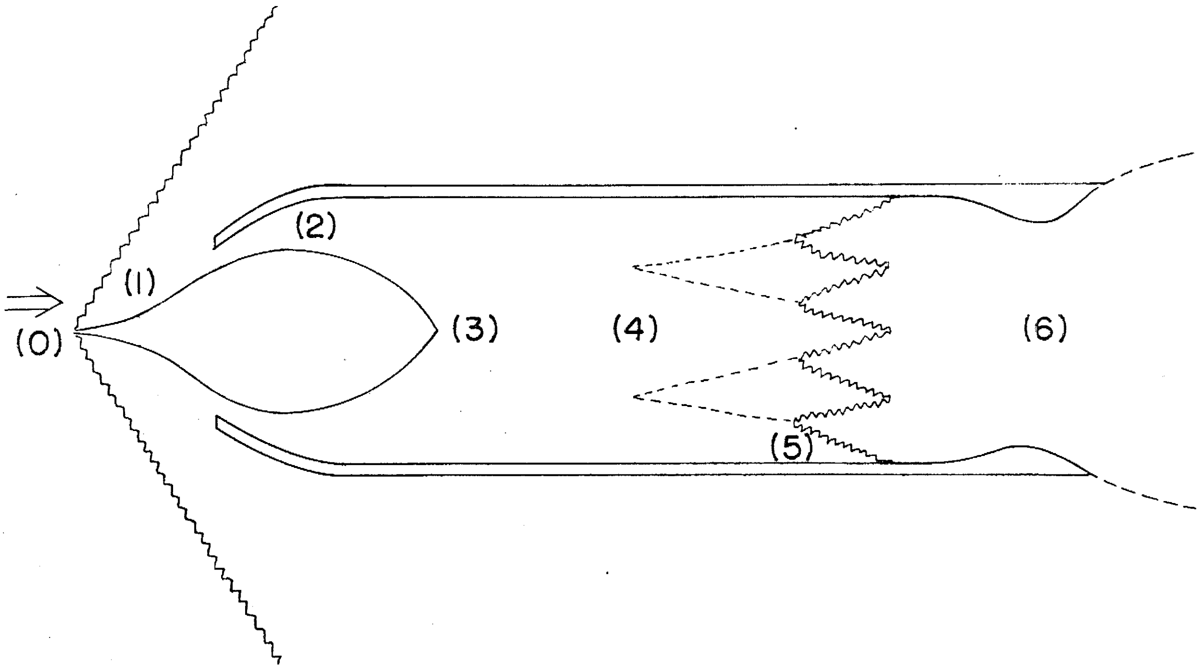


FIG. III - 10 RAMJET SCHEMATIC

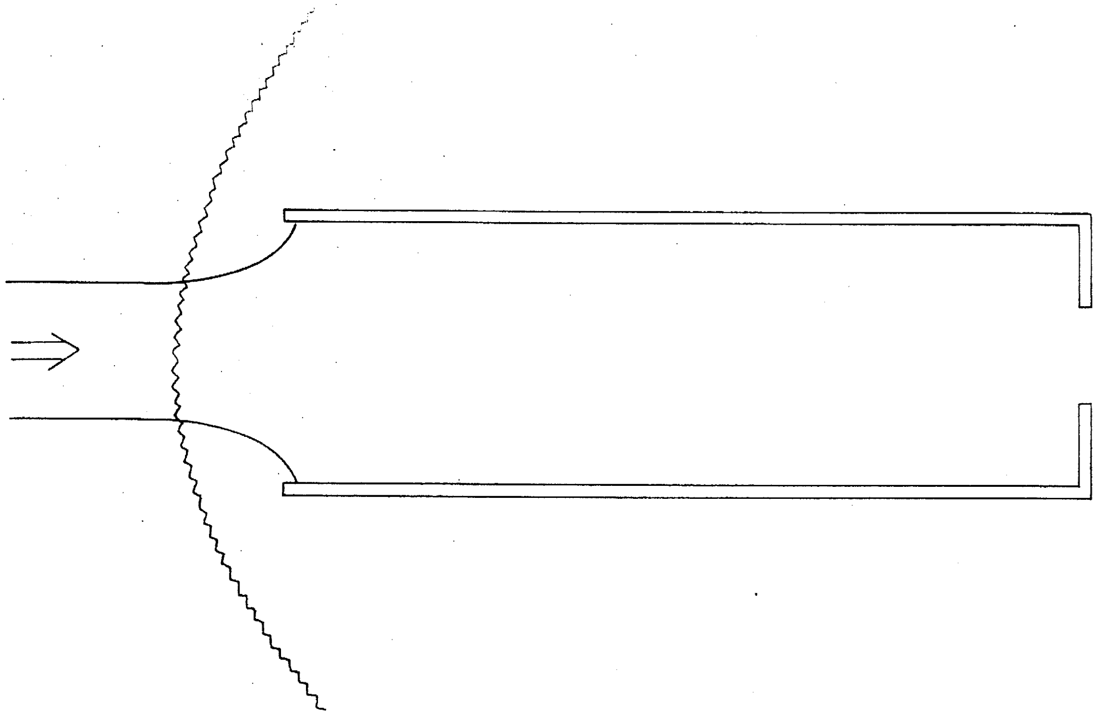


FIG. III - 11 OPEN ENDED TUBE

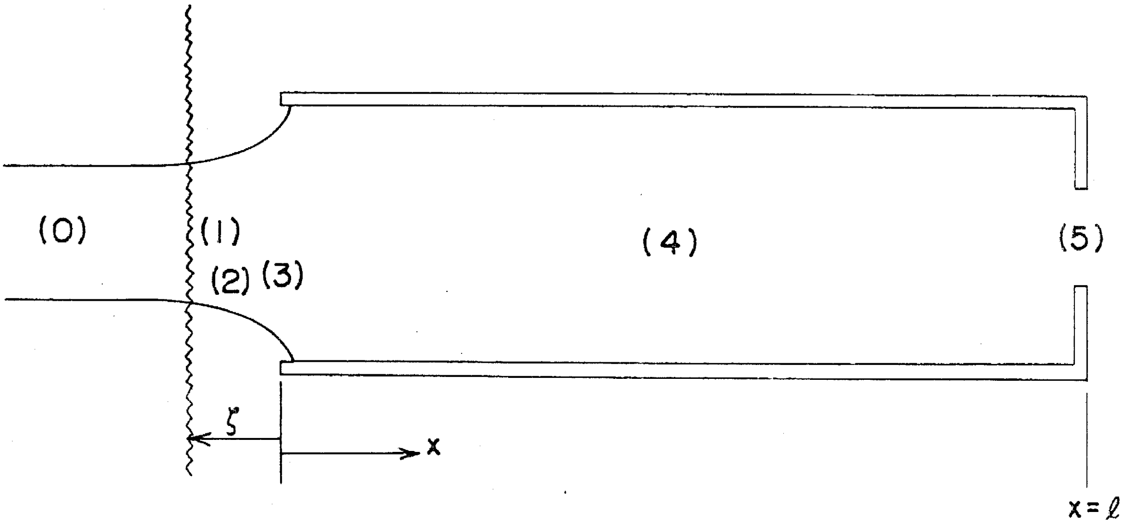


FIG. III - 12 MODEL OF OPEN ENDED TUBE

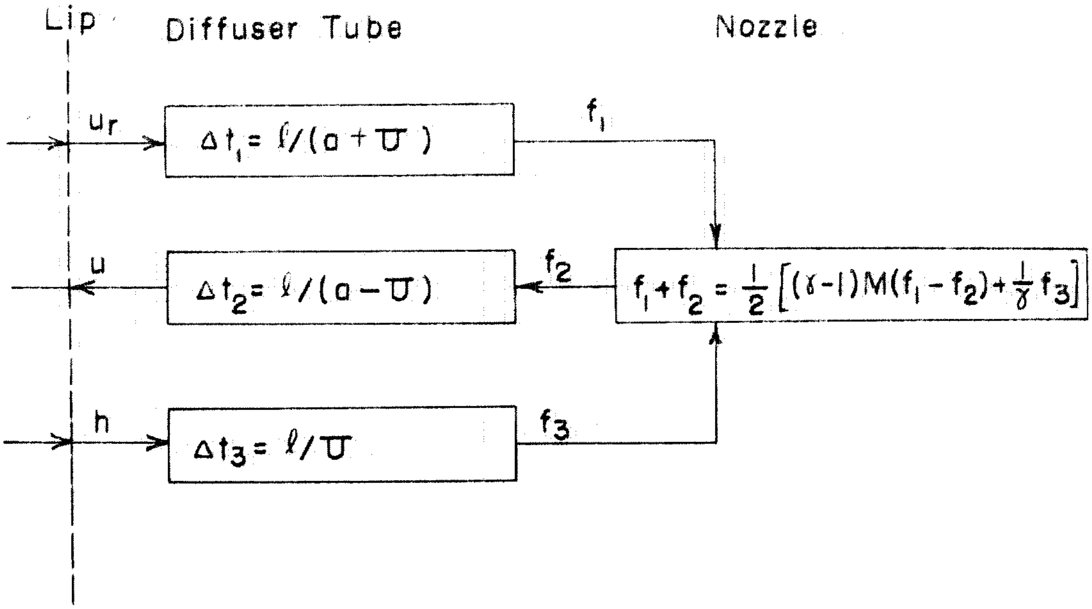

Lip
Diffuser Tube
Nozzle
u_r
$\Delta t_1 = \ell/(a+\overline{U})$
f_1
u
$\Delta t_2 = \ell/(a-\overline{U})$
f_2
$f_1+f_2 = \frac{1}{2}\left[(\gamma-1)M(f_1-f_2)+\frac{1}{\gamma}f_3\right]$
h
$\Delta t_3 = \ell/\overline{U}$
f_3

FIG. III - 13 DIFFUSER TUBE ANALOG

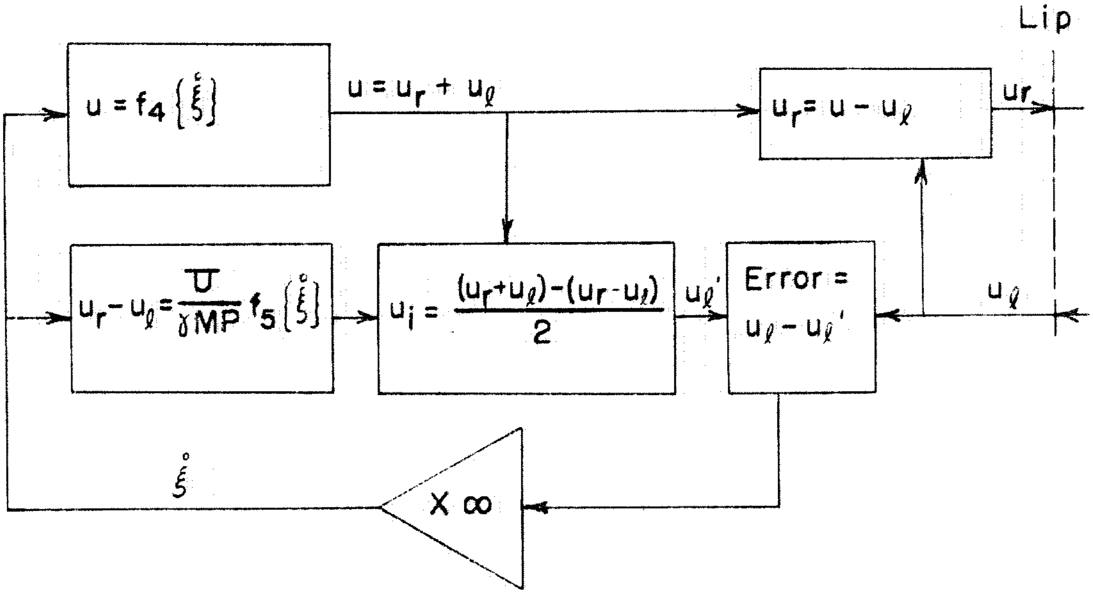

Lip
$u = f_4\{\mathring{\xi}\}$
$u = u_r + u_\ell$
$u_r = u - u_\ell$
u_r
$u_r - u_\ell = \frac{\overline{U}}{\gamma MP} f_5\{\mathring{\xi}\}$
$u_i = \frac{(u_r+u_\ell)-(u_r-u_\ell)}{2}$
u_ℓ'
Error = $u_\ell - u_\ell'$
u_ℓ
$\mathring{\xi}$
X ∞

FIG. III - 14 IDEALIZED EXTERNAL DIFFUSER ANALOG

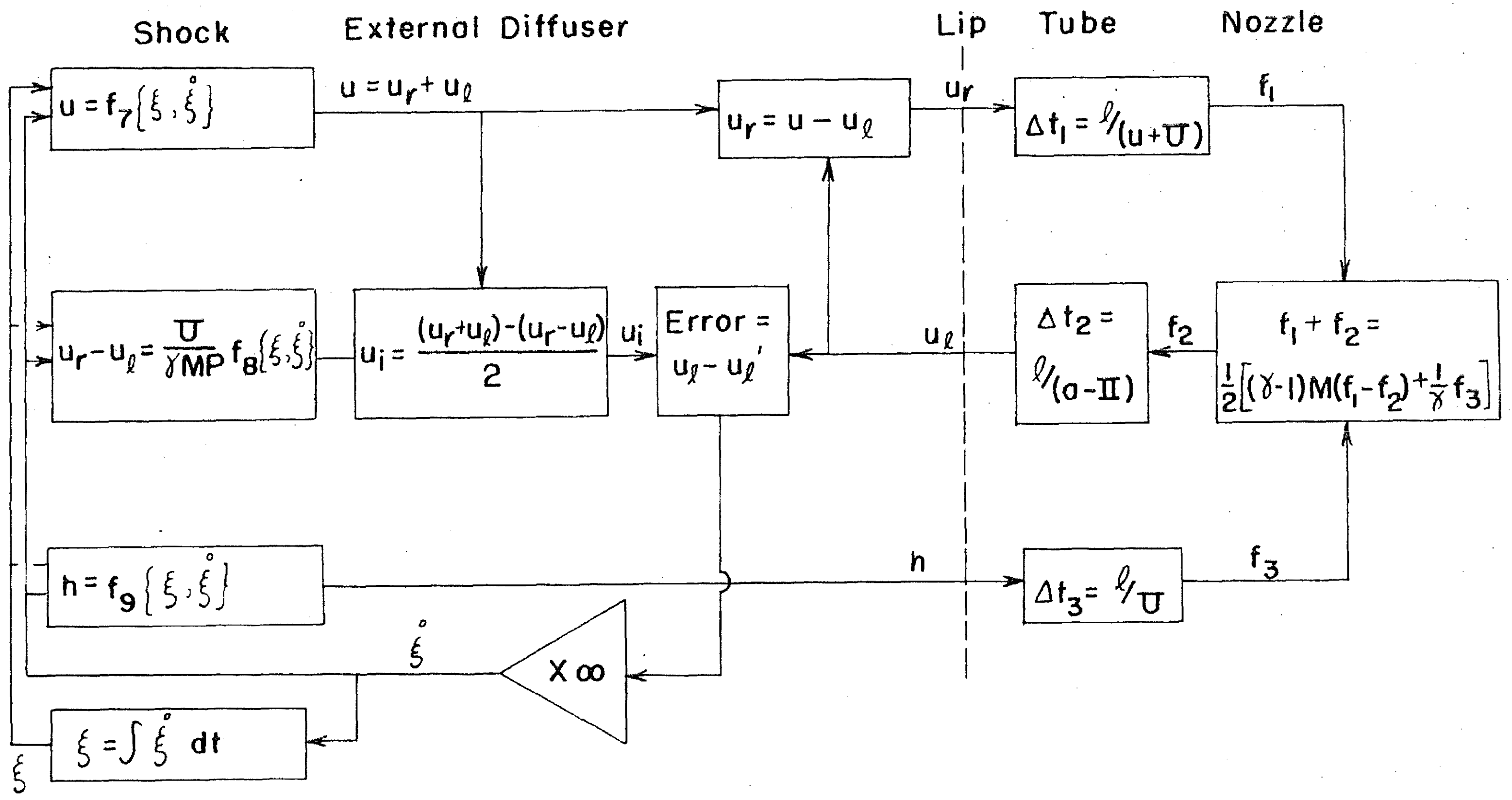

Shock
External Diffuser
Lip
Tube
Nozzle
$u = f_7\{\xi, \mathring{\xi}\}$
$u = u_r + u_\ell$
$u_r = u - u_\ell$
u_r
$\Delta t_1 = \ell/(u+U)$
f_1
$u_r - u_\ell = \frac{U}{\gamma MP} f_8\{\xi, \mathring{\xi}\}$
$u_i = \frac{(u_r+u_\ell)-(u_r-u_\ell)}{2}$
u_i
Error = $u_\ell - u_\ell'$
u_ℓ
$\Delta t_2 = \ell/(a-U)$
f_2
$f_1 + f_2 = \frac{1}{2}[(\gamma-1)M(f_1-f_2)+\frac{1}{\gamma}f_3]$
$h = f_9\{\xi, \mathring{\xi}\}$
h
$\Delta t_3 = \ell/U$
f_3
$\mathring{\xi}$
X ∞
$\xi = \int \mathring{\xi}\, dt$
ξ

FIG. III-15 COMPLETE ANALOG

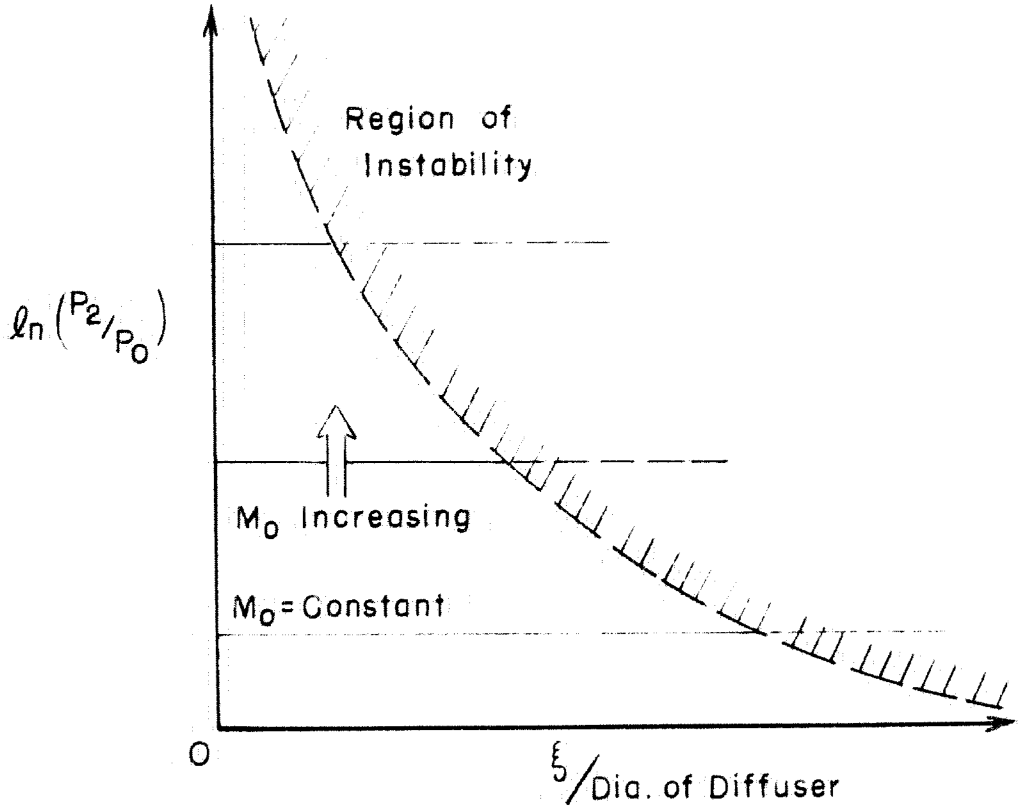


FIG. III - 16 STABILITY BOUNDARY, VARIABLE CHOKING

IV. PROCEDURES FOR IMPROVING SERVOMECHANISM RESPONSES

A. Introduction

The intent of this section[*] is to demonstrate simple experimental methods of treating servomechanisms. The given portion of the system (the amplifier, actuator, and load) is assumed to have *a defined set of characteristics and limitations.* The methods concentrate principally upon the inputs and outputs of this "forward" portion of the system, and other than altering the amplifier static gain function, no attention is paid to internal workings. The basic system dynamics may be equivalent to a third order or a fifteenth order differential equation that may be linear up to saturations or may include nonlinearities. Saturations (limiting) need not be perfect. Unconcern of the methods for system complexity should not be construed as a claim of complete universality; one could concoct any number of nonlinear systems for which the methods would be poor or not applicable. However, the techniques appear valid for many of the commonly encountered servomechanisms.

The reason for treating a servomechanism is to improve its performance, immediately requiring two basic performance definitions: a definition of the input and a definition of deviation or error compared to an ideal system response. More or less general inputs can be defined as functions of time by Fourier series, Taylor series, power series, etc. Limited Fourier series

[*] This section reviews and extends the method described in Ref. 34.

and limited Taylor series approaches have both been applied to nonlinear control systems. By assuming that the input contains only one frequency, the fundamental, and by further assuming that the system's output components other than the fundamental are all negligible, the present Fourier methods have reduced themselves to extended forms of linear theory. As mentioned in Section I, an input containing simultaneously two or more frequencies requires, if extended linear theory is to apply, superposition of the fundamental solutions. Furthermore, the infinite response speed of the ideal system of either the linear or the extended linear theory is unrealistic. Like the Fourier techniques, most of the current Taylor series methods consider the simplest form of the series, the displacement step input. Despite its crude manner of approximating a general input, it more than compensates by offering a definition of an ideal system having substance. For a step input, one can say, "the ideal system is one that, having the same physical limitations as the real system, moves most quickly from position a to position b." The importance of this quality cannot be overemphasized. For once, a pr ctical goal is created recognizing the fact any real servo system is in essence an accumulation of physical limitations.

In an effort to define the operating regions of a servomechanism, a task that must be performed if nonlinear treatments are to have meaning, we designate three zones:

Null region: the region in which small motions occur without excursions exceeding the tolerated error $|\delta|$ of the system. In itself, the character of the motions is generally unimportant except for the requirement that $|\delta|$ and possibly $|\dot{\delta}|$ limits should not be exceeded. Nonlinearities such as backlash and coulomb stiction are often prominent.

Optional unsaturated region: the region extending from the null region to the operating boundary at which saturations are first evidenced. The useful operations of a servomechanism treated as a linear system are entirely restricted to this region. Contactor (relay) control systems eliminate the region. Within the region, the system may be nearly linear or may have intentional and unintentional nonlinearities.

Saturated region: the zone extending as far as desired beyond the saturation boundary. Operation in this region complies with the performance concept of utilizing the maximum physical capabilities of the system.

It follows that the ideal system has no optional region and has an appropriately small null region. If, due to technique limitations, one cannot eliminate the optional region without unduly enlarging the null error, $|\delta|$, the compromise takes the form of minimizing this anti-ideal region.

For clarity's sake, the techniques of treating the specific regions are discussed in reverse order, beginning with the fundamental, saturated region. Throughout this discussion, the defined input is a displacement step. At the end of the sections, techniques of extending the methods to include more terms of the Taylor series are touched upon.

B. Saturated Region of Operation

The postulated system has its physical limitations, static gain curve, and whatever devices are needed for the inner two regions of performance defined. It is further assumed that either rate feedback or static error is available for modification in the saturated region of operation. All known time lags, time delays, and nonlinearities of the amplifier, actuator, coupling, and load should be included in the model, unless fortuitously the real system is available for tests.

First, certain interpretations of the worksheet, the output displacement vs. velocity plot, need be considered. The plot is viewed as an experimental data record, ignoring mathematical implications as a phase-space cross section. The following set of examples demonstrate work-plot pictures of quickest response to a step function input.

The system is initially at rest at point x_1 (Fig. IV-1), and at time $t = 0$ is requested by the input to move instantaneously to point x_i. For our purposes, the final point is surrounded by a null region, and this possibly surrounded by an unsaturated region. We wish to direct the response as displayed by a path or trajectory on the plot most rapidly from x_1 to the midst of the end regions about x_i.

If the system has almost unlimited available force, its quickest route to the final point is by means of a nearly infinite acceleration to some midpoint and then a nearly infinite deceleration to the final point. The corresponding response trajectory, Figure IV-1a, rises vertically from the initial point x_1 to an almost infinite velocity and then descends vertically down to x_i. The relation between the trajectory and the response time is apparent:

$$\text{Response time} = T = \int_0^T dt = \int_{x_i}^{x_1} \frac{dx}{dx/dt}$$

Thus the response time is the area under the reciprocal of the plotted trajectory (as a function of x). For the ideal response of Figure IV-1a, the travel time is practically zero.

Using this representation, the effects of physical limitations can be easily shown. Velocity saturation prevents the trajectories from rising above a velocity limit $(dx/dt)_{max}$. A control having no other limitations best performs the task of moving from x_1 to x_i by an infinite acceleration up to the bounding velocity and then a constant maximum velocity to the endpoint where an infinite deceleration stops the motion. The resultant trajectory, Figure IV-1b, now has a finite response time.

In all practical systems the output acceleration must be finite, limited by inertia of the load and available force. For example, a system having a pure inertia load and limited available force but no other restrictions has as its best response: 1) maximum acceleration to the midpoint and 2) maximum deceleration to the endpoint. The trajectory is that of Figure IV-1c.

If the control force is not able to reverse itself instantaneously at the midpoint, the apex of the trajectory is rounded as shown in Figure IV-1d.

Implied in all of the examples is the rule that minimum transit time trajectories are obtained with a single reversal. Some investigators feel this point is not obvious, and a rigorous mathematical proof has been attempted for second order systems with discontinuous forcing functions (Ref. 2). The same conclusion using a little physical intuition can be deduced for real systems. The rising trajectory, say of Figure IV-1d, forms the left-hand boundary of possible trajectories rising from x_1, for by definition this trajectory has the maximum acceleration sponsored by the maximum available force. Similarly, the descending trajectory to x_i forms the right-hand boundary of all possible trajectories terminating at x_i. (Obviously a trajectory to the right of the boundary could reach x_i by recycling, but this back door approach always adds a little time for the go around and can be neglected here.) The saturated reversal at the peak is the fastest possible means of transferring from the ascending boundary to the descending boundary; therefore it occurs closer to the apex of the boundary intersection (noting that the apex point represents infinitely fast reversal) than any other reversal. Therefore, any trajectory between x_1 and x_i made up of unsaturated or multiple reversal segments must lie inside the boundary set up by the single reversal saturated trajectory. Since transit time is the area under the reciprocal of the velocity curve, it is apparent that the minimum transit time corresponds to the bounding trajectory.

Now let us consider the switching function in more detail. The typical instantaneous reversal curves of Figure IV-2a have been shifted such that the endpoint is always the origin by the simple transformation $x' = x - x_1$. The abscissa is now the static error. For various starting points, x'_1, x'_2, - -, there exists perfect switching (or reversing) points, s_1, s_2, - -, which properly deflect trajectories toward the origin. If we could supply a switching function consisting of all of the "s" points, the saturated response of the system would always be ideal.

Admitting a maximum rate of change of controller force, $(dF/dt)_{max}$, the family of optimum responses become those of Figure IV-2b. Note the switching line $s_1\ s_2\ s_3$ - - initiating reversal has been advanced to account for the reversal time; the intervals s_1t_1, s_2t_2, - - are the reversal portion of the trajectories.

If some absolute velocity barrier is added to the system, the perfect switching line may take the odd form shown in Figure IV-2c.

In short, a system switched at the proper instant from saturated acceleration to saturated deceleration fulfills the performance ideal. Suppose now, that the amplifier is a perfect relay reversing instantaneously when the error signal reverses polarity. Looking at Figure IV-2, we note that the switching line $s_1\ s_2$ - - can be considered either as a function of velocity plotted versus the ordinate, dx/dt, or as a function of the static error plotted versus the abscissa, x'. With the first consideration the switching line is $S_1\left\{dx/dt\right\}$, and we wish the relay to reverse when a trajectory crosses this line. (See Figure IV-3a) In other words, when the static error, x', equals the switching function $S_1\left\{dx/dt\right\}$

the relay control signal should reverse polarity from plus (acceleration) to minus (deceleration).

$$\eta = -x' + S_1 \{dx/dt\} \qquad \text{IV-1}$$

The similarity between this desired relation and the actual error signal obtained by the usual linear position plus linear rate feedback is apparent; for the linear control,

$$\eta_L = x_1 - x - a_1 \cdot dx/dt = -x' - a_1 \cdot dx/dt \quad .$$

When the rate feedback is intentionally nonlinear, $A \{dx/dt\}$, the error signal is

$$\eta = -x' - A \{dx/dt\} \qquad \text{IV-2}$$

and comparison with the sought relation IV-1 shows that the ideal rate feedback function, $A \{dx/dt\}$, is precisely the negative of the data plot switching line, $S_1 \{dx/dt\}$.

If we had assumed the switching line to be a function of x', $S = S_2 \{x'\}$, by the same argument the error should reverse when the trajectory meets the switching line (Fig. IV-3b). In this case the error signal should reverse polarity when the output velocity, dx/dt, reaches the static error function $S_2 \{x'\}$; that is,

$$\eta = -dx/dt + S_2 \{x'\} \qquad \text{IV-3}$$

Suppose in the physical servo system we combine a function of static error, $A_2 \{x'\}$, with a linear rate feedback to form the total error signal, η (Fig. IV-4). For this arrangement, the relay control signal is

$$\eta = -dx/dt + A_2 \{-x'\} \qquad \text{IV-4}$$

and again comparing the desirable function (eq. IV-3) with the obtainable function (eq. IV-4), we note the nonlinear function of negative static error should be identical to the data plot switching line, $S_2\{x'\}$.

In actual systems, switching is not instantaneous for various reasons. The error signal must have a small but finite positive value to sponsor saturated accelerations and a small but finite negative value for saturated decelerations. Therefore, in passing through zero, a time delay or time lag is evidenced before saturated deceleration is observed. Time lags in the system and physical limitations restricting reversal rates enhance this prolongation of reversal. The switching line is actually a sort of switching band on the work plot. However, for the experimental techniques discussed here, the point is unimportant.

In brief, the features of the control system of present interest are: 1) either linear static error combined with functional rate feedback or functional static error combined with linear rate feedback produces a switching control which can be readily interpreted on the velocity vs. static error response plots; and 2) this switching function should be molded to deflect saturated rising trajectories downward directly toward the origin of the data plot. These features lead to a simple, direct, experimental method of determining the best function of rate feedback, or alternatively of static error.

Experimental Method of Obtaining "Ideal" Saturated Responses

1. The complete servomechanism including linear rate feedback and linear static error is set up either as a model or as an actual

system.

2. Velocity vs. static error responses are recorded for various displacement step inputs.

3. In general, the trajectories will not be perfect, but will miss the origin by certain increments, $\Delta x'_n$ (for the n-th trajectory), dependent upon initial starting points (Fig. IV-5a). If these trajectories are adjusted left or right by $-\Delta x'_n$, they become perfect (Fig. IV-5b). As previously shown, the shift corresponds directly to a modification of the rate feedback, $(dx/dt)_n$, immediately generating the correction to apply to the initial linear rate feedback. It is, of course, possible to interpret this new rate feedback function in terms of its fraternal static error function.

4. Test of the new system may still, for reasons of experimental errors, undue nonlinearities in switching, etc., evidence slight misses, which if unacceptable can be extracted by iteration.

Systems having terms which vary with output position x (such as those introduced by an aerodynamic flipper subjected to nonlinear static hinge moments) will have different families of responses for each final value, x_1. This in turn produces a family of "ideal" switching functions; i.e. the switching function is now a three dimensional surface rather than a line.

A simple example illustrates the method. The given system (Fig. IV-6) consists of an inertial load, a linear actuator having one time lag, an amplifier having one time lag, direct position feedback, and rate feedback sensed by a device having two time lags.

The amplifier statically saturates at some output level $|v_s|$. Treating the system first in its linear operating region ($|v|<|v_s|$), we provide a linear rate feedback and thus a feedback

$$-\left[1+\frac{a_1 p}{(1+\tau_3 p)(1+\tau_4 p)}\right] x$$

The forward transfer function is

$$x = K_1 \frac{1}{p^2(1+\tau_1 p)(1+\tau_2 p)} \eta$$

The resultant sixth order system displayed the instabilities usually encountered with rate feedback stabilization. For a fixed low gain (K_1), decreasing the rate feedback (decreasing a_1) below some value produced underdamped responses and finally the "displacement mode" hunting. Increasing rate feedback above some other higher value produced underdamped responses and finally a higher frequency "velocity mode" hunting. As static gain was increased, the two values of a_1 giving critical damping approached each other, and at approximately optimum gain these values merged.

With this experimental procedure, a linear static gain, K_1, and a linear rate feedback, $a_1 \dot{x}$, were selected. The resultant static characteristic of the amplifier is shown in Figure IV-7. If the amplifier output does not exceed $|v_s|$, which corresponds to an error signal $|\eta_s|$, the responses are linear and slightly less than critically damped. These responses are discussed in more detail in Part C below. For the present, we are interested only in the saturated and far saturated responses of the system $(|n| > |n_s|)$.

As an expedient for getting saturated responses rapidly, two response tests were run, each with a single large step input. The first employed the linear rate feedback prescribed by unsaturated tests, and the second used half this rate feedback. As shown in the resulting responses (Figs. IV-8,9), several saturated reversals come out of one test. By measuring the several misses, the correction rate feedback function was determined (taking into account the reduced linear damping in the second case), and this function (Fig. IV-10) was produced in the feedback by the nonlinear function generator discussed in Section II. Resultant responses to various step inputs are shown in Fig. IV-11. A comparison of saturated response times of the system with nonlinear rate feedback and the system with linear rate feedback is shown in Figure IV-12. This is, of course, a different manner of presenting the information shown in Figure IV-7.

C. Optional Unsaturated Region

If we accept the postulate, "best response is saturated response," this is the unwanted region. The preceding analysis assumed, for convenience, that this unsaturated region was approximately linear since the end zone was of no great consequence to performance in the far saturated regions (except possibly to furnish a reason for selecting the initial linear rate feedback).

In looking more carefully at a magnified view of the endpoint region outside the null, we are immediately confronted with a provocative dilemma. Linear tradition states per se that the system should be linear and operation remain within the saturation limits. In practical systems, the amplifier is set at some maximum gain as related to the choice of stabilizing elements. Suppose we have such a system, linear except for amplifier limiting with gain and rate feedback adjusted for approximately best linear response. We further restrict the displacement step inputs such that the saturation limits of the amplifier, $\pm \eta_s$, (Fig. IV-13) are not exceeded. This is about as ideal a situation as could be expected for the linear system. But comparison with the postulated "ideal saturated" system indicates performance can be improved by revising the static gain curve as shown by the dotted line in Fig. IV-13, extending saturation toward $\eta = 0$.

This super-performance, if you will, as compared to the linear performance has many subtleties attached, and the present investigation makes no pretense of optimization. It does show that even a rather primitively selected nonlinear system is superior both in step response and in frequency response to the equivalent linear system (keeping the same rules of the game, i.e. that the η_s as determined by the linear system is not exceeded).

Haphazard selection of the new, narrower static gain characteristic invites underdamped responses or instabilities; for example, simply steepening the central portion linearly (as dotted in Fig.

IV-13a) tends toward instability, since the original static gain slope was the maximum compatible with properly damped responses. It might be inferred from extended linear theory that any narrower static gain curve tends toward instability, but this inference is not valid. It is clear that the present method deals with "best performance" of systems: the linear system is assumed to have highest gain compatible with critically damped response; the saturated system has its response trajectories properly directed toward the null region. On the other hand some of the extended linear techniques deal with systems which are marginally stable or unstable. For example, a result of extended frequency analysis is that limiting, such as that produced by amplifier saturation, is stabilizing; i.e. the mean effective gain of the system is lowered by saturations. A servo system which is unstable in its linear region is "stabilized" at some limit cycle oscillation if saturation occurs. The present method suggests precisely the reverse. A servo system which produces critically damped responses in its linear region of operation evidences overshooting when saturation is introduced. There is, of course, no discrepancy in these results; the two approaches are concerned with quite different operating conditions of the servo system.

Here, attempting to improve upon linear (unsaturated) performance of the system, we intentionally saturate part of the operation by narrowing the unsaturated region. Between the points where static limiting occurs, the average gain of the amplifier is substantially increased. When this average gain exceeds the linear gain sponsoring neutral stability, extended linear theory tells us that a limit cycle is to be expected. To demonstrate that this is not a basic limitation of the proposed method, we can consider the following rather extreme case. Assume initially that we have a system working within its linear limits and producing critically damped responses. Now we collapse the central unsaturated region of the static gain to nearly zero width (infinite average gain), making sure that the small unsaturated section remaining across the null is maintained at low gain (Fig. IV-13b). Extended frequency analysis tells us that if the nonlinear static gain exceeds some linear limit, shown as a dotted line, the system becomes unstable. This leads to the conclusion that a stable limit cycle exists at point a, an unstable limit cycle or barrier exists at point b, and the null region is stable. Excursions of the error signal within the inner limit barrier result in stable responses. If the error signal exceeds the inner limit barrier, the responses "lock into" periodic stable limit cycle oscillations. There is a delay in reversal due to the time lags in the system ahead of the saturation. By the procedure basic to the proposed method, we can determine the correction to apply to the rate feedback function (directly related to the miss of a trajectory) and reduce the limit cycle. By a series of such tests, the limit cycle can be reduced to any arbitrary small maximum amplitude external to the unsaturated region. If the limit cycle is reduced to confinement within the unsaturated region, it disappears by virtue of the well damped responses of the low static gain characteristic. In other words, the limit cycle predicted by extended linear analysis may be reduced to an arbitrarily small amplitude or eliminated by adjustment of the rate feedback function and the shape of the unsaturated static characteristic.

The actual nonlinear static gain characteristic used was obtained by trial and error and was justified solely by the fact that it produced stable responses. It was characterized (Fig. IV-19) by new saturation limits, η'_s , substantially closer to $\eta = 0$ and by low gain through $\eta = 0$. The shape was obtained by a pair of diodes, back-to-back, shorting the internal feedback of the amplifier. One would suspect that such a static characteristic is far from optimum; unfortunately the interesting problem of optimum static characteristic is of a magnitude which excludes it from this research.

Once the new static gain curve is established, the experimental technique developed in Section IV, Part B, is applied. Here, the linearly determined rate feedback is more than adequate, switching too early (predicting a slower deceleration), and model tests with "best" linear feedback result in trajectories attempting to undershoot the origin (Fig. IV-15 noting initial position has been shifted to clarify details of the final approach).

Just as in treating the far saturated responses, by reading off the miss distance for each trajectory, the shift in x' required to bring the trajectories down on the endpoint is measured, thus providing the correction to the original rate feedback. In contrast to the preceding saturated analysis, the correction here is negative, decreasing the rate feedback.

A comparison of responses to various displacement step inputs was made between the system incorporating nonlinear static gain and the newly determined rate feedback function and the linear system optimized as previously described. Responses of both systems to three different steps are shown in Figures IV-14, 17, 18; the first step corresponded to an error signal just under η_s, (just under saturation of the linear system), the second to approximately 50% of η_s, and the last to approximately 15% of η_s. In all three instances, the inner trajectories were those of the linear system. Remembering that transit time is the integral of the inverse curve, we conclude that the nonlinear system is superior to the linear system for displacement step inputs, and particularly so at smaller requested inputs where the linear system is constrained to low forces.

Both systems were tested for frequency responses to an input with peak amplitude corresponding to about 25% η_s. The results are shown for 1/2, 1, and 3 cps respectively on Figure IV-20. Perfect response is a 45° line. The unfortunate pair of "teats" on the input vs. output traces is due to the odd waveform of the signal generator (Fig. IV-21). For each of the frequencies tested, the nonlinear system had less attenuation and less phase shift than the linear system.

D. Null Region

The character of the null region is intimately linked to the choice of static amplifier gain function and rate feedback function for the adjacent operating region. Its only feature of interest is size, determined by dead space effects (multivalued null), or possibly by small limit cycles. The simplest limit cycle tests are for complete instabilities, for which no matter how the response enters the null region, the final state is a limit cycle. Partial instabilities are more difficult, for only certain kinds of entry conditions excite the cycle. If a system has prominent nonlinearities around the null, one must always make sure the range of entry conditions (inner boundary exit conditions of region 2) do not provoke some unacceptable limit cycle. For model studies of nonlinear null effects, representative electrical analog functions are available: for example, stiction as shown in Figure II-14 and dead space as shown in Figure II-12.

E. Extension of the Taylor Series

Only the first term of the Taylor series approximation of the input was used in developing the technique described above for treating saturated responses. The more general approximation of the input introduces the velocity step, the acceleration step, etc. It will be shown that for a given system, the inclusion of the next higher order Taylor series term results in extending the switching function by one dimension. For example, if the optimum response to an arbitrary displacement step input can be obtained by a switching line, $A\left\{\dot{x}\right\}$, then the optimum response to an input consisting of an arbitrary displacement step plus an arbitrary velocity step can be obtained by a switching surface, $A\left\{\dot{x},\dot{z}\right\}$, where $\dot{z}$ is the velocity error $\dot{x} - \dot{x}_1$.

Admission of the second term of the Taylor series approximation of the input introduces the velocity step or ramp function. There have been some investigations of saturated system responses to ramps alone, but essentially none on treating systems subjected to the more general arbitrary displacement step plus arbitrary ramp input. On the data plot, the endpoint is now not the origin but a target which at initiation of the problem begins to move with constant velocity (see Fig. IV-22). The perfect system must be able to intercept this target with saturated motions and only one reversal. If the target velocity is $\dot{x}_1$, we can run the problem backwards in time, defining the collision point as occurring at $x' = 0$ (here, x' is not the static error) and obtaining a family of saturated trajectories such as those shown in Figure IV-23. This, of course, can also be produced by shifting X-wise the set of saturated trajectories obtained for step displacement inputs (for example, Figure IV-2) such that they all intercept point $0, \dot{x}_1$. Note the curious shape of the switching line. From the facts 1) that the function is an inconvenient shape to duplicate, 2) that the shape varies depending upon target velocity, $\dot{x}_1$, and 3) that the x' variable requires that we know the impact point when the problem is initiated, we are tempted to look for another approach.

At the risk of confusion, we fall back upon a mathematical representation of the system, assuming the forward portion of the system can be described by

$$L\{x,t\} = \eta\{t\}$$

where L is an ordinary differential operator, t the independent time variable appearing implicitly in L, and η is the error signal forcing the complete forward system. Saturations are thus included in L. Furthermore, we assume the error signal η, is the sum of an input, $x_i\{t\}$, and a function of the output, $-x - f\{x,t\}$. The feedback, as before, transfers- x to the input summer along with some intentionally added function of time derivatives of x, $f\{x,t\}$. Here the input is a position step plus a ramp, so

$$L\{x,t\} = x_{i_o} + \dot{x}_{i_o} \cdot t - x - f\{x,t\}$$

Since f contains t implicitly only, it can be combined with L, giving

$$L'\{x,t\} + x = x_{i_o} + x_{i_o} \cdot t \qquad t > 0 \qquad \text{IV-5}$$

Our objective is to select a substitution such that the right hand forcing function drops out, thus reducing the equation to autonomous form. The reduced equation produces solutions having a certain amount of generality. For example, in the treatment of systems subjected to single displacement step inputs, equation IV-5 can be written

$$L'\{x,t\} + x = x_{i_o} \qquad t > 0$$

and the substitution $x' = x - x_{i_o}$ (introducing static error) produces

$$L'\{x',t\} + x' = 0$$

providing L' contains only derivatives of x. The governing differential equation is unchanged by the substitution. If we utilize x' instead of x, the resultant responses are in a sense independent of x and of x_{i_o}. Instead of considering all possible x and all possible x_{i_o}, we need take only all possible x'. Paralleling this, we can attempt a substitution in IV-5,

$$x = x - x_{i_o} - \dot{x}_{i_o} t \qquad \text{(static error)}$$

$$\dot{z} = \dot{x} - \dot{x}_{i_o} \qquad \text{(velocity error)}$$

First, we note that terms like $g\{x\}$ in L' complicate the substitution by bringing in explicit time and consequently reaffirm the previous assumption that L' contains only derivatives of x. The substitution produces

$$L'\left\{(\dot{z} + \dot{x}_{i_o}),\ t\right\} + z = 0$$

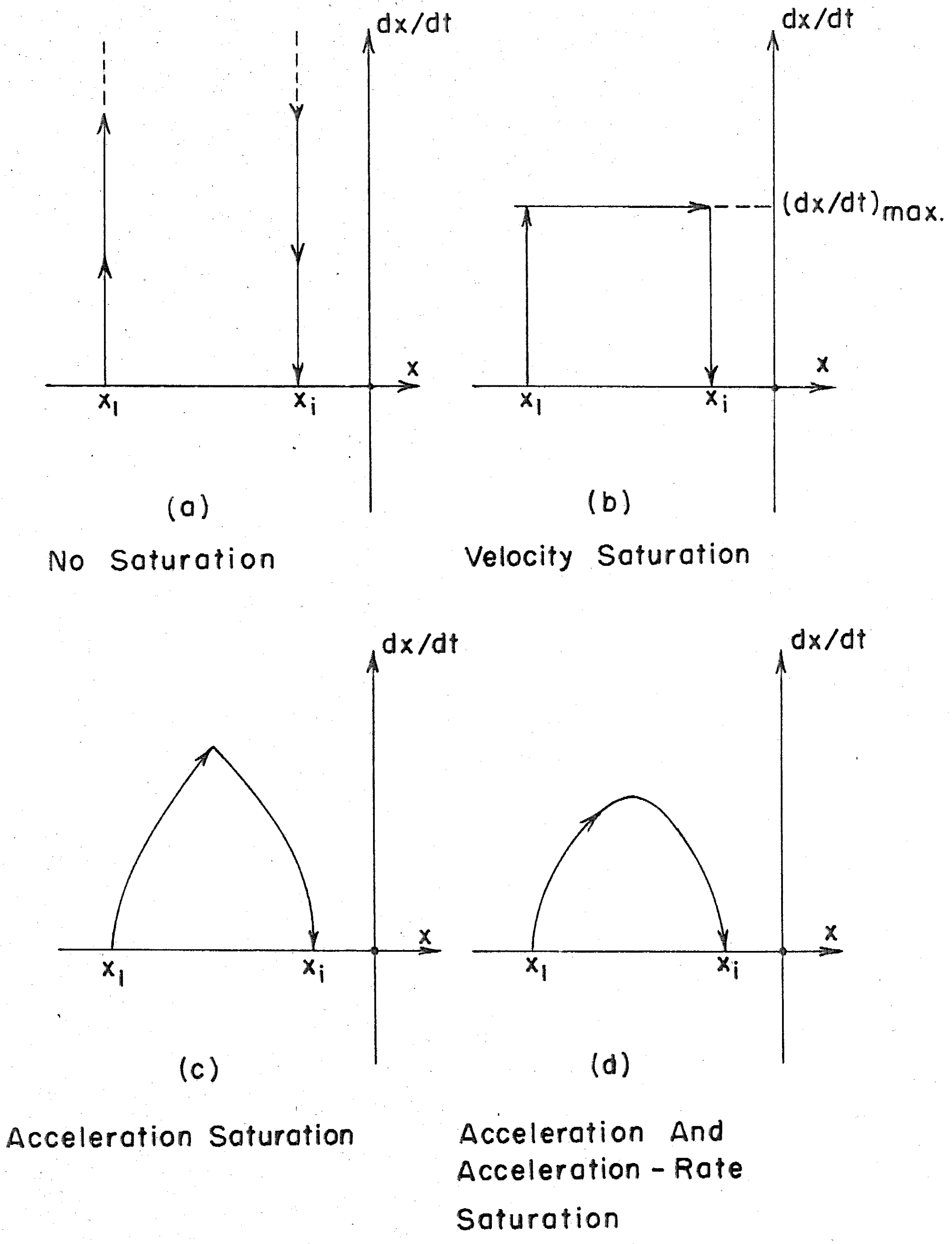
where L' includes terms such as $(\dot{z} + \dot{x}_{i_o})$, $\ddot{z}$, $\dddot{z}$ - - -. For systems having only second order or higher terms (inertia loads with no damping or springs), the $(\dot{z} + \dot{x}_{i_o})$ factors drop out and we get

$$L'\left\{z, t\right\} + z = 0 \qquad \text{IV-6}$$

The equation exhibits the desirable qualities of an analog. It says that physically the responses $\dot{x}$ vs. x' of a system forced by a step position have an exact correspondence with responses $\dot{z}$ vs. z of a system forced by a step position plus a step velocity. If we have obtained by the technique previously discussed the optimum responses, ideal switching line, and desired rate feedback function for a system exposed to displacement step inputs, we simply relabel the axes $\dot{z}$ and z, reading off the desired rate feedback now as a function of $\dot{z}$, the velocity error. Alternatively, the switching function can be thought of as a function of static error, leading to a system with functional static error and linear velocity error feedback. For such treatment the system must know velocity error as well as the natural measurement, static error. The approach gives an inkling of what a really intelligent servomechanism must do: continuously compare all pertinent features of the input with its own output and make the proper compensations within itself to minimize the importance of these errors.

The existence of damping in the system inserts $(\dot{z} + x_{i_o})$ terms in L', and the corresponding optimum feedback functions are different for different $\dot{x}_{i_o}$, a procedure identical in all respects to the position step treatment of systems having springs. When this more elite treatment is applied to systems containing springs, the switching function is four dimensional, x_{i_o} and $\dot{x}_{i_o}$ both influencing the selection of the proper rate feedback function.

Extension of such an approach to include third order terms adds one dimension of complexity to the feedback function.



(a)

No Saturation

(b)

Velocity Saturation

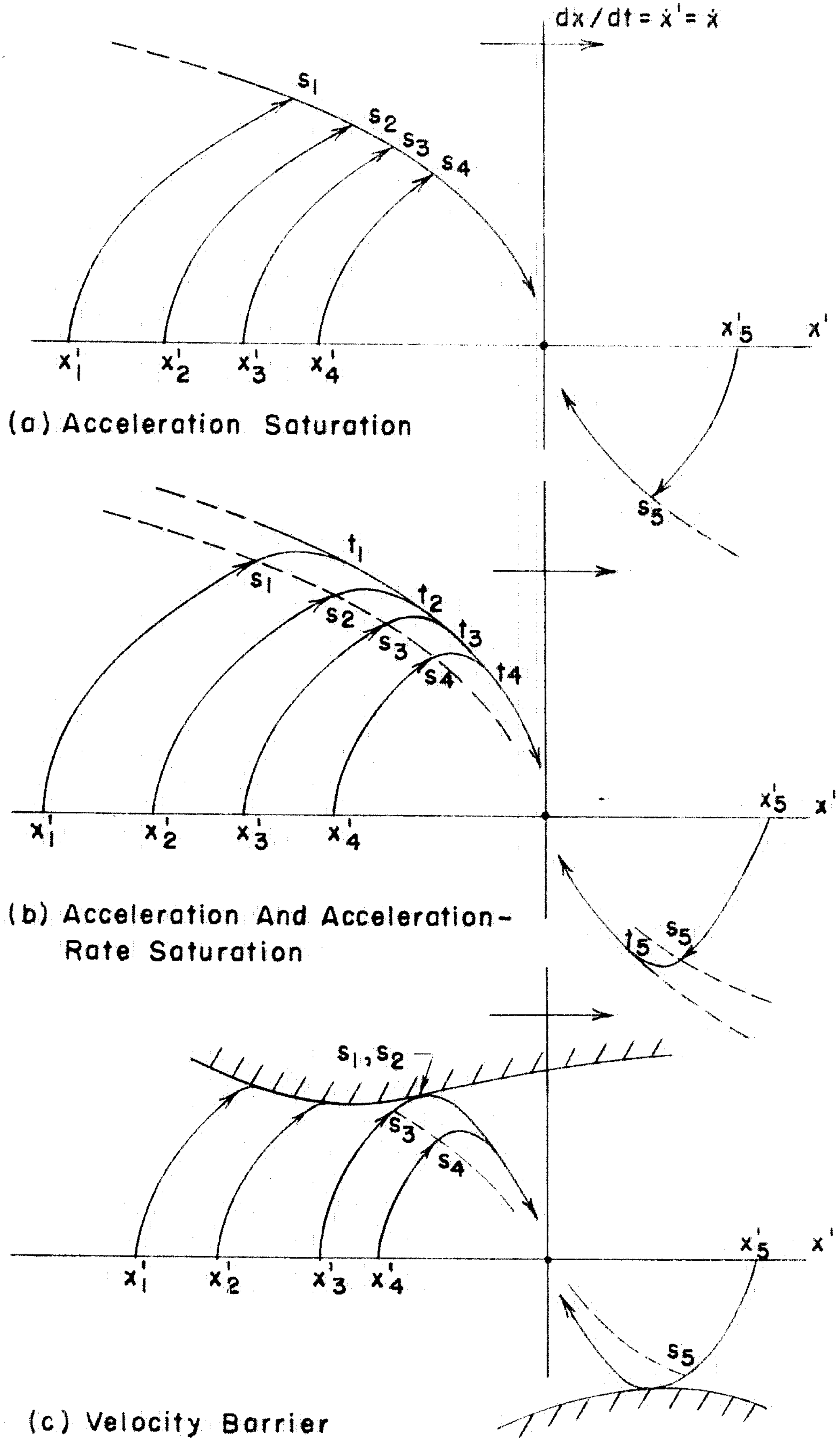
(c)

Acceleration Saturation

(d)

Acceleration And Acceleration - Rate Saturation

FIG. IV-1 OPTIMUM SATURATED RESPONSES

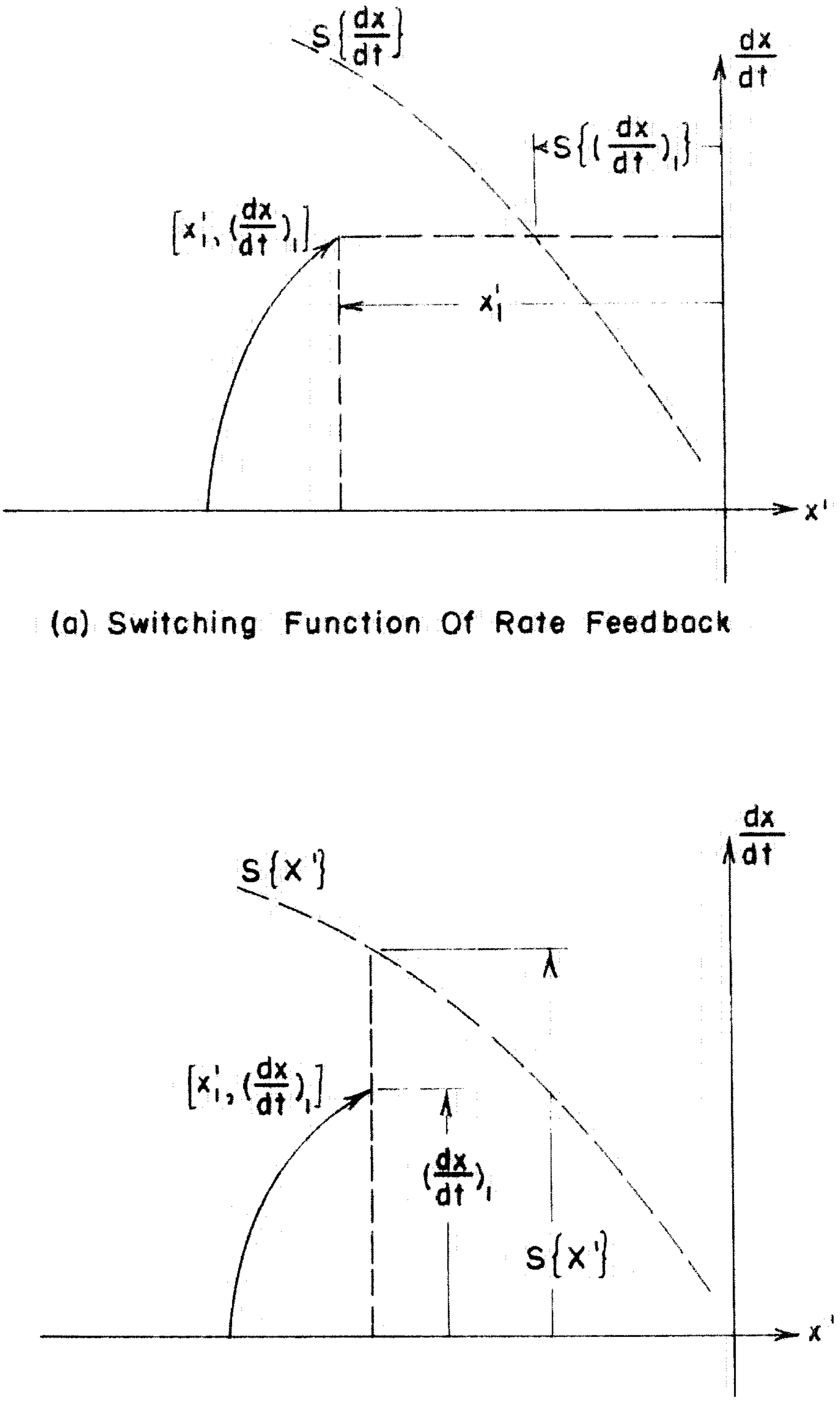


(a) Acceleration Saturation

(b) Acceleration And Acceleration-Rate Saturation

(c) Velocity Barrier

FIG. IV - 2 OPTIMUM SWITCHING LINES



$S\left\{\frac{dx}{dt}\right\}$
$\frac{dx}{dt}$
$S\left\{\left(\frac{dx}{dt}\right)_1\right\}$
$\left[x'_1, \left(\frac{dx}{dt}\right)_1\right]$
x'_1
x'
$S\{X'\}$
$\frac{dx}{dt}$
$\left[x'_1, \left(\frac{dx}{dt}\right)_1\right]$
$\left(\frac{dx}{dt}\right)_1$
$S\{X'\}$
x'

(a) Switching Function Of Rate Feedback

(b) Switching Function Of Static Error

FIG. IV - 3

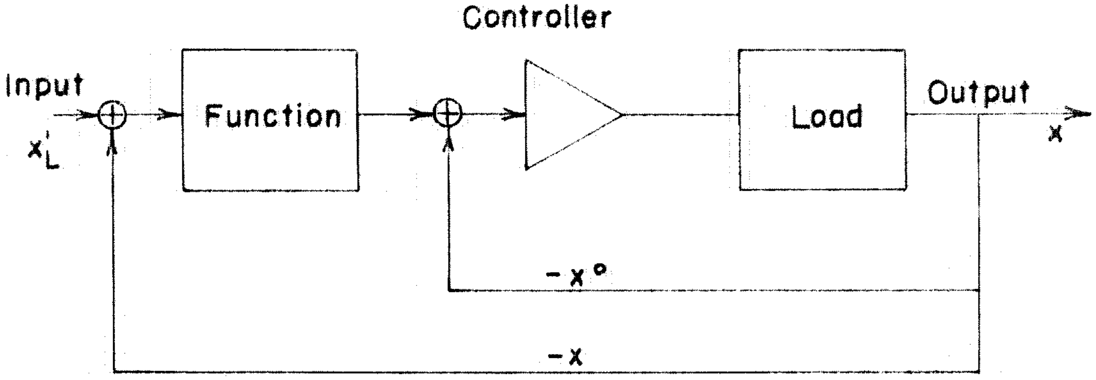


FIG. IV - 4 SERVO WITH STATIC ERROR FUNCTION

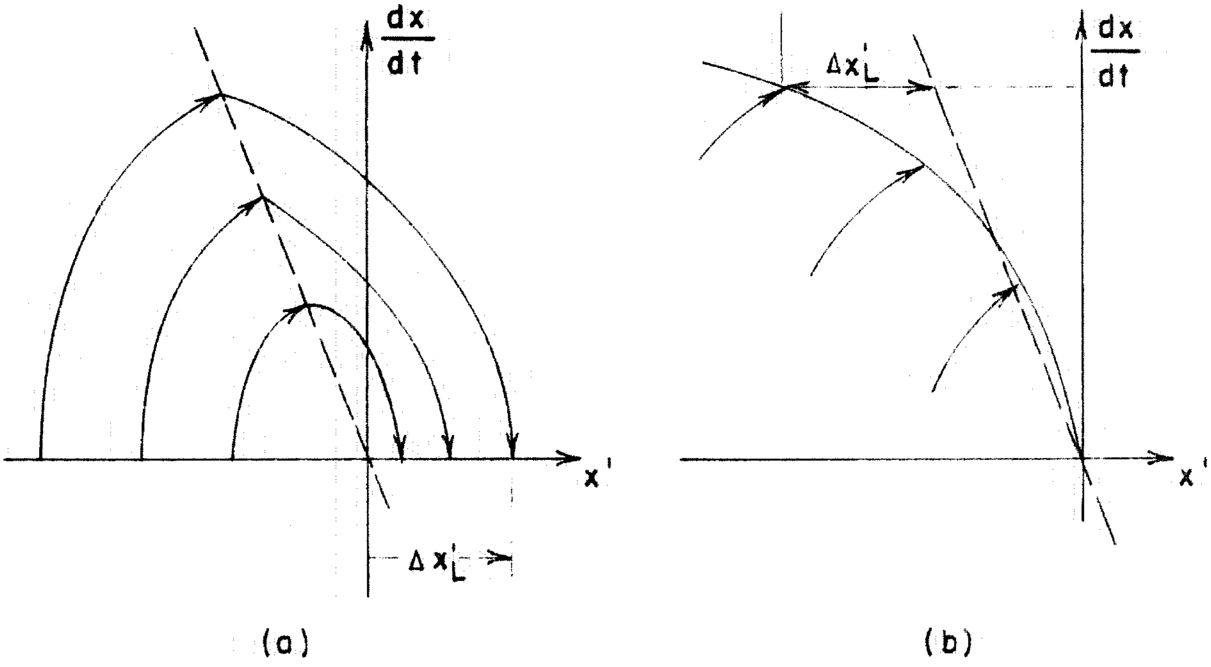
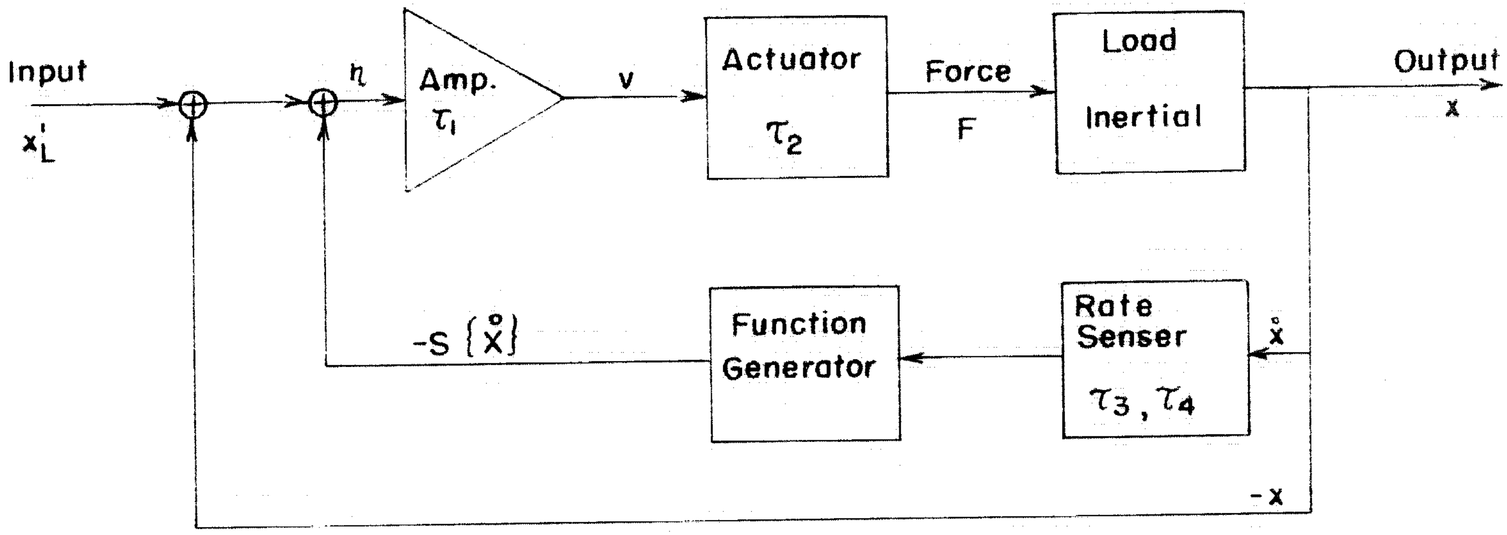


FIG. IV-5 METHOD OF FINDING RATE FEEDBACK FUNCTION



$\tau_1 = 0.023$ secs. $\tau_3 = 0.005$ secs.

$\tau_2 = 0.005$ secs. $\tau_4 = 0.0025$ secs.

FIG. IV-6 EXPERIMENTAL SERVO SYSTEM

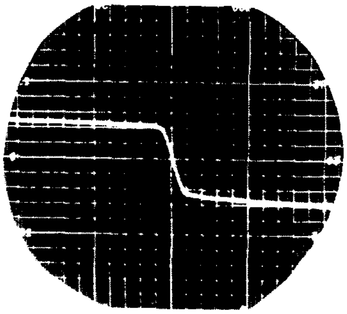


Figure IV-7

Amplifier Static Gain

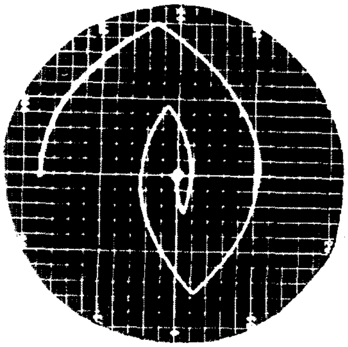


Figure IV-8

Responses with "best"

Linear Rate Feedback

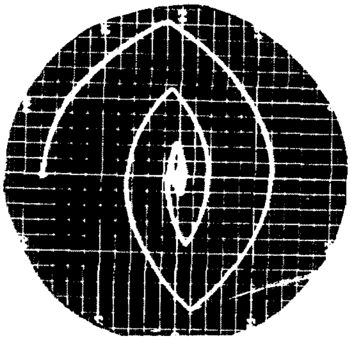


Figure IV-9

Responses with half of

"best" Linear Rate Feedback

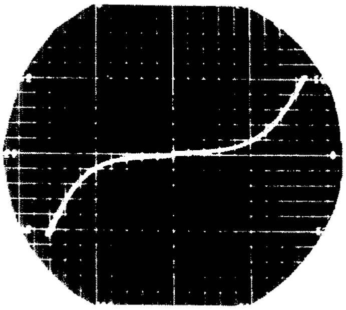


Figure IV-10

New Nonlinear Rate

Feedback Function

Figure IV-11

Responses to New

Rate Feedback

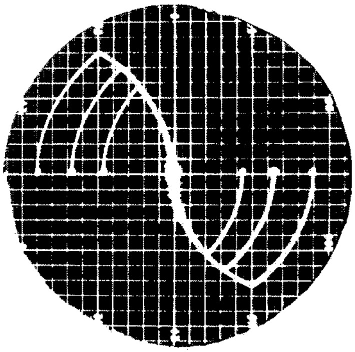
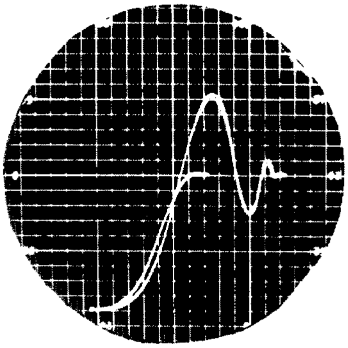
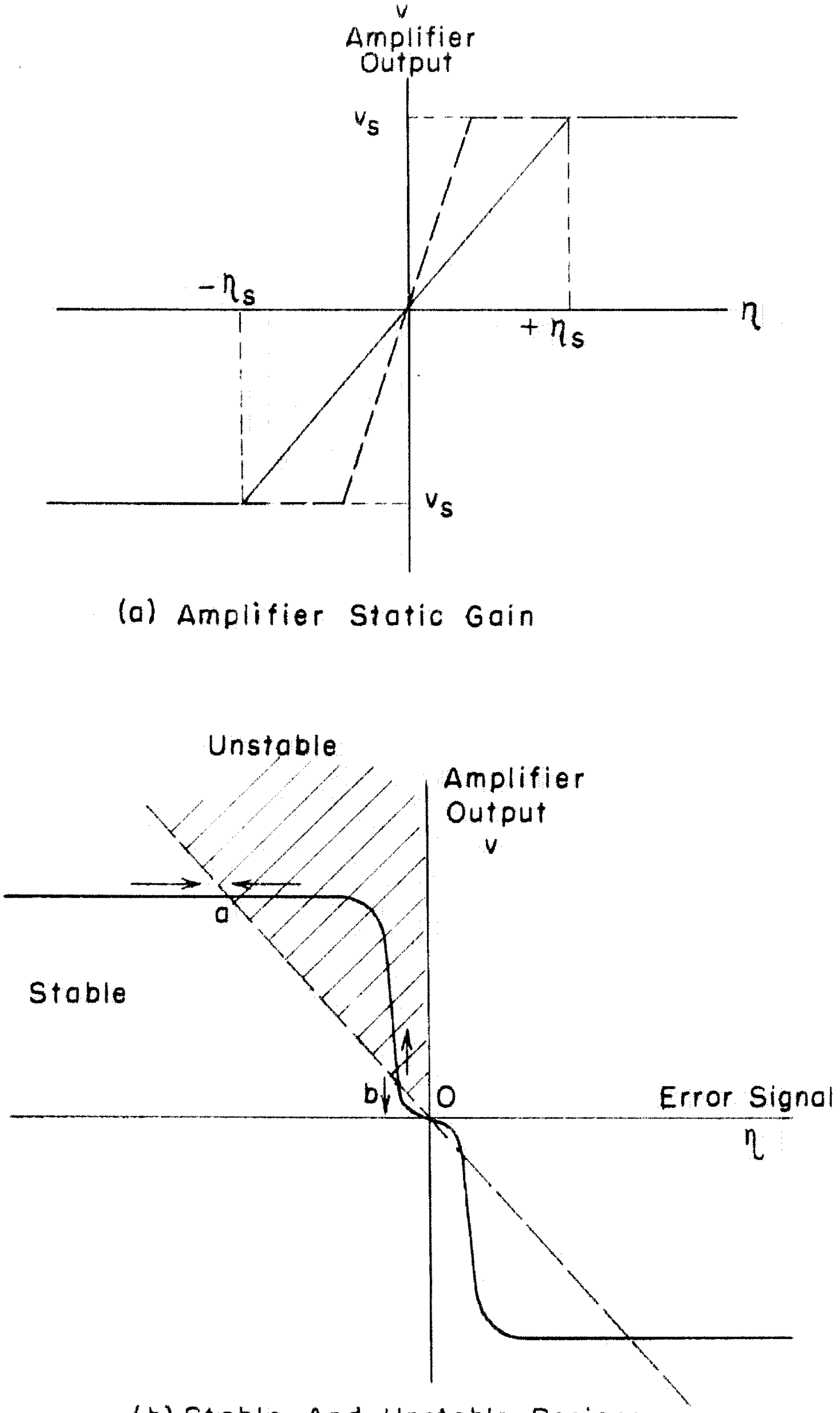


Figure IV-12

Comparison of Response Time

of New and Linear Systems





(a) Amplifier Static Gain

(b) Stable And Unstable Regions

FIG. IV - 13

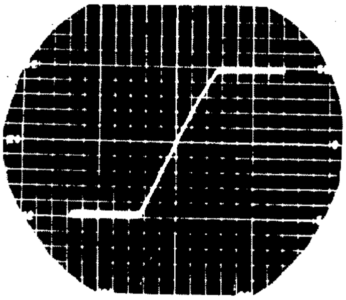


Figure IV-14

Amplifier Static Gain

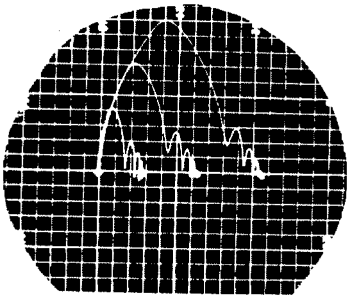


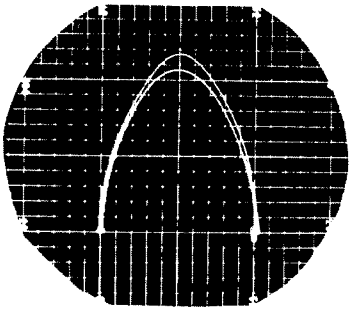
Figure IV-15

Responses with Linear

Rate Feedback

Figure IV-16

Large Step Input



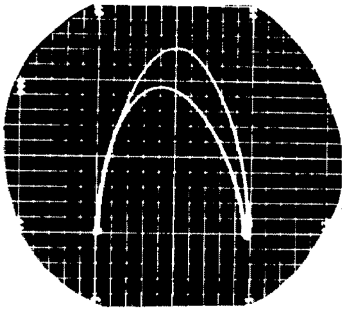


Figure IV-17

Medium Step Input

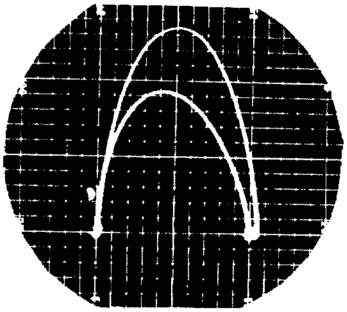


Figure IV-18

Small Step Input

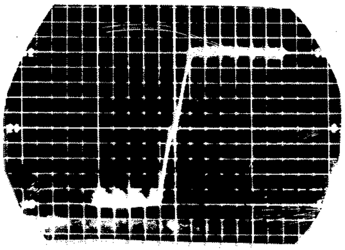
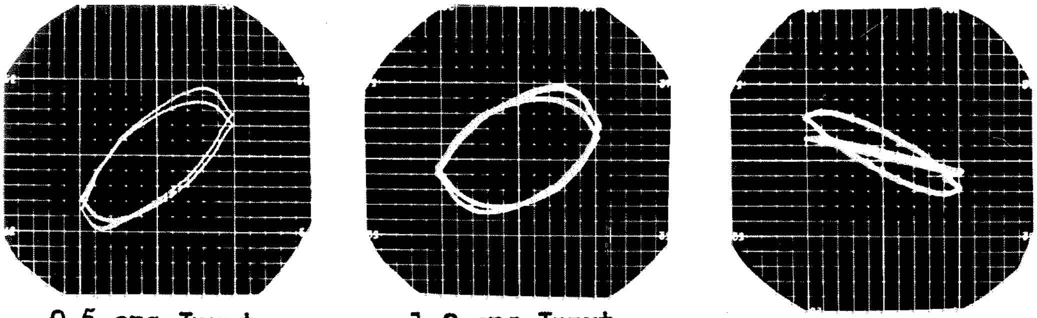


Figure IV-19

New Amplifier Static Gain

Figure IV-20

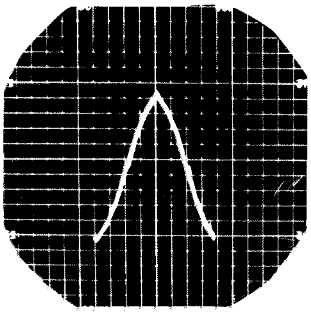


0.5 cps Input

1.0 cps Input

3.0 cps Input

Figure IV-21



Distortion in Input Waveform

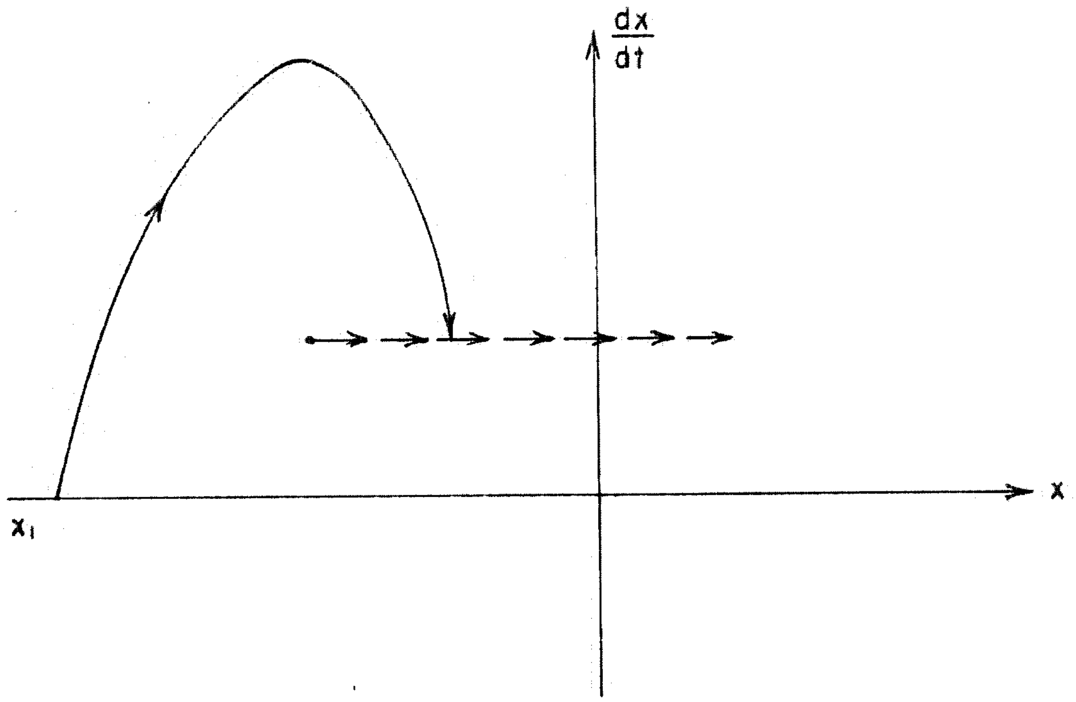


FIG. IV - 22 INTERCEPTING A CONSTANT VELOCITY INPUT

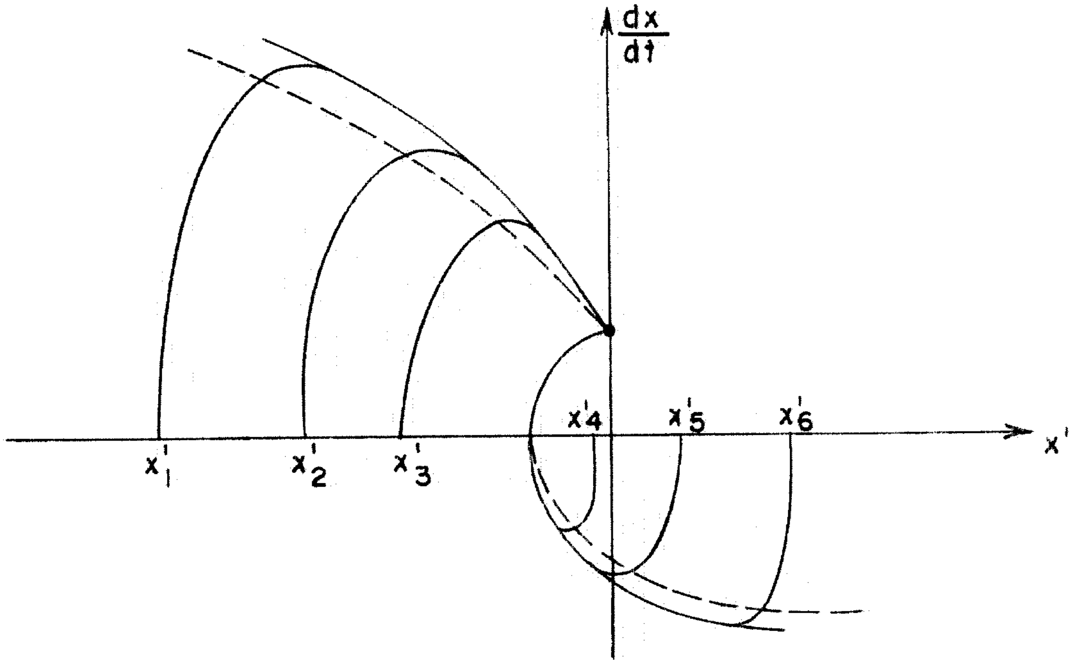


FIG. IV - 23 TRANSFORMED TRAJECTORIES

V. CONCLUSIONS

A. Internal methods of treating closed loop systems, using the word "internal" as it is defined in Section I, confine themselves to systems which can be described either by second order, quasi-linear, autonomous, ordinary differential equations or by higher order, constant coefficient, linear, ordinary differential equations having discontinuous forcing functions ($F = \pm$ constant). Neither descriptions apply to the majority of real physical problems.

B. External methods, again as defined in Section I, are experimental in their nature; these methods must accept full responsibility not merely for producing results but for providing means of interpreting the results. A complicated nonlinear system might for example, be modeled by an electrical analog and experimental tests run producing a set of responses. Unless the method (including the model) is so designed that it provides the problem director with insight of how to control the results or why the results occurred, the experimental solution is somewhat trivial.

C. If a variety of experimental tests by modeling are sought, first one looks for a versatile model. The electrical analog is useful and reasonably general in its applications. However, there are several gaps in the list of desirable analog elements, in particular the logarithm type of function generator answers the need for a flexible exploratory nonlinear element. Compared to other function generators, this unit performs rapidly, accepts three independent variables, and effects functions of each variable by a pair of controls.

D. Having selected the form of the model, one next attempts to translate the essential features of the real system into model terms. This is the most difficult and most deceptive step of the procedure. However, if the translation is done properly (including a fortuitous selection of the model form), the model offers a powerful means of analysis, that of paralleling experiments on the actual system and experiments on the flexible model. The hydraulic servomechanism is extraordinarily deceptive. The extremely complex ramjet system offers an example in which extraction and modeling of a relatively simple component instability can be accomplished.

E. The final step is model testing and interpreting test results. If the system is a servomechanism, one wishes to exercise some control over the results, manipulating available parameters such that the system performs well. This implies that good performance is defined: the system input is defined and the departure of the servomechanism responses from "ideal responses" is defined.

1. Approximation of the input by a limited Taylor series has the advantage of admitting into performance an "ideal system" which has realistic limitations. The ideal system always utilizes its full physical capabilities (i.e., is saturated).

2. If the system is operating in its saturated region, the "ideal control" can be found directly through experiment. This control may be either a nonlinear rate feedback or a nonlinear static error function. Each additional term of the Taylor series

(velocity step, acceleration step, etc.) included in the definition of the input adds one dimension to this control function.

3. Customarily, the linearly treated servomechanisms are confined to operation within their saturation limits. If all but the null region of such a servomechanism is saturated by, say, nonlinearly shaping the static gain characteristic, the system response to step inputs can always be improved. Once the static gain function has been selected, the procedure is the same experimental procedure applicable to the far saturated responses (conclusion E-2, above). Certain systems coincidentally improve their frequency responses; the latter result has not at present been fully explained.

REFERENCES

1. Flugge-Lotz, I.: "Discontinuous Automatic Control of Missiles," Stanford Division of Engineering Mechanics Technical Report 14 (1950), Part I.

2. Bushaw, D. W.: "Differential Equations with a Discontinuous Forcing Term," Stevens Experimental Towing Tank Report No. 469 (Jan. 1953), Ph.D. dissertation in Mathematics for Princeton University.

3. McDonald, D.: "Nonlinear Techniques for Improving Servo Performances," Cook Research Laboratories, Bulletin S-2 (1950).

4. Cook Research Laboratories: "Basic Research in Non-Linear Mechanics as Applied to Servomechanisms," Technical Report T16-1 (Jan. 1952).

5. Rose, N. J.: "Theoretical Aspects of Limit Control," Stevens Experimental Towing Tank Report No. 459 (Nov. 1953).

6. MacColl, L. A.: "Fundamental Theory of Servomechanisms," (book) Van Nostrand (1945), p.p. 107-125.

7. Weiss, H. K.: "Analysis of Relay Servomechanisms," Journal of Aeronautical Sciences, N. Y. (July, 1946), Vol. 13, p. 364.

8. Hopkin, A. M.: "A Phase-Plane Approach to the Compensation of Saturating Servomechanisms," AIEE Transactions (1951), Vol. 70, Part I, p.p. 631-639.

9. Uttley, A. M., and Hammond, P. H.: "The Stabilization of On-Off Controlled Servo-Mechanisms," Automatic and Manual Control (book), London, Butterworths Scientific Publications (1951), p.p. 285-299.

10. Hazen, H. L.: "Theory of Servo-Mechanisms," Journal, Franklin Institute (Sept. 1933), Vol. 218, No. 3, p.p. 279-331.

11. Minorsky, N.: "Introduction to Non-Linear Mechanics," J. W. Edwards, Ann Arbor (1947), Chap. VIII.

12. Minorsky, N.: "Introduction to Non-Linear Mechanics," J.W. Edwards, Ann Arbor (1947), Chap. IX.

13. Kryloff, N. and Bogoliuboff, N.: "Introduction to Non-Linear Mechanics," Princeton Univ. Press (1947), p.p. 55-63.

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

15. Minorsky, N.: "Introduction to Non-Linear Mechanics," J. W. Edwards, Ann Arbor (1947), Part I.

16. Levinson, N. and Smith, O. K.: "A General Equation for Relaxation Oscillations," Duke Mathematical Journal (1942), Vol. 9, p.p. 382-403.

17. Rauch, L. L.: "Oscillation of a Third Order Nonlinear Autonomous System," Contributions to the Theory of Nonlinear Oscillations, Princeton Univ. Press (1950), p.p. 39-88.

18. Cypser, R. J.: "Optimizing Performance of Servomechanisms by Continuous Variation of System Parameters," M.I.T. Servomechanisms Lab. Eng. Memorandum No. 20 (Nov. 1950).

19. Loeb, J.M.: "A General Linearizing Process for Non-Linear Control Systems," Automatic and Manual Control, London, Butterworths Scientific Publications (1952), p.p. 275-284.

20. McCann, G. D., Lindvall, F. C., and Wilts, C. H.: "Effect of Coulomb Friction on the Performance of Servomechanisms," AIEE Transactions (1948), Vol. 67, Part I, p. 540.

21. Cosgriff, R. L.: "Open Loop Frequency Response Method for Nonlinear Servomechanisms," AIEE Technical Paper 53-253 (1953).

22. Goldfarb, L. C.: "On Some Non-Linear Phenomena in Regulating Systems," (translation), Nat'l. Bur. of Standards Report 1691 (May 29, 1952).

23. Greenwood, I. A., Holdam, J. V., and MacCrae, D.: "Electronic Instruments" (book), M.I.T. Rad. Lab. Vol. 21 (1948), McGraw Hill, p.p. 55, 125.

24. Korn, G. A. and Korn, T. M.: "Electronic Analog Computers" (book), McGraw Hill (1952), p. 213.

25. Howard, R. C., Savant, C. J., and Neiswander, R. S.: "Linear-to-Logarithmic Voltage Converter," Electronics (July 1953), p.p. 156-157.

26. Howard, R. C.: "A Special-Purpose Electric Analog Computer," Ph.D. Thesis (1953), California Institute of Technology.

27. Savant, C. J.: "A Nonlinear Computer for the Solution of Servomechanism Problems," Ph.D. Thesis (1953), California Institute of Technology.

28. Canning, T. N.: "A Simple Mechanical Analogue for Studying the Dynamic Stability of Aircraft Having Nonlinear Moment Characteristics," NACA-TN-3125.

29. Chestnut, H., and Mayer, R. W.: "Servomechanisms and Regulating System Design, Vol. 1," (book), Wiley (1951), p.p. 179-186.

30. Baeriswyl, L.: "An Analysis of an Electrohydraulic Closed-loop Servomechanism," Bumblebee Series Report No. 155 (May 1951), APL, John Hopkins Univ.

31. Millstone, S. D.: "Electro-Hydraulic Analogies," Machine Design, Part I (Dec. 1952), p.p. 185-190; Part II (Jan. 1953), p.p. 166-170; Part III (Feb. 1953), p.p. 131-135.

32. Stoolman, L.: "Investigation of an Instability Phenomenon Occurring in Supersonic Diffusors," Ph.D. Thesis (1953), California Institute of Technology.

33. Dailey, C. L.: "Supersonic Diffuser Instability," Ph.D. Thesis (1954), California Institute of Technology.

34. Neiswander, R. S. and MacNeal, R. H.: "Optimization of Nonlinear Control Systems," AIEE Applications and Industry (Sept. 1953), No. 8, p.p. 262-272.

APPENDIX A

Phase Space Construction

An ordinary differential equation of the form $L\left\{x,t\right\} = 0$ and of order n can, if autonomous, be written as a set of n simultaneous equations:

$$L\left\{x_n, x_{n-1}, --- x_1 --- x\right\} = 0 \qquad \text{A-1}$$

$$\left.\begin{aligned} x_{n-1} &= \frac{dx_{n-2}}{dt} = \frac{d^{n-1}x}{dt^{n-1}} \\ x_1 &= \frac{dx}{dt} \end{aligned}\right\} \qquad \text{A-2}$$

The variables x_i (where $n \geq i > 0$) can be plotted against x, producing a set of phase space cross sections such as those shown in Figure A-1. For usual plotting techniques, one must restrict x_i where $n > i > 0$ to continuous functions, and x_n to piecewise continuity. However, the lower order terms may also be of piecewise continuity if they happen to be sufficiently well defined at discontinuities to establish complete new sets of initial conditions.

The construction technique is as follows. From initial conditions, the starting point on each of the planes is located (designated point 0). The slope of the trajectory (here the intersection of the multidimensional response function with a particular plane) can be linked to the subsequent plane trajectory by:

$$\frac{dx_i}{dx} = \frac{dx_i}{dt}\frac{1}{dx/dt} = \frac{x_{i+1}}{x_1}$$

We know the direction to begin motion in plane 1, because trajectories run to the right for $+x_1$ and to the left for $-x_1$. Therefore the slopes of trajectories for all but the n-th plane are determined immediately by rotating the initial x_1 value to the x axis and carrying this value through all of the planes (see Fig. A-1). The slopes x_{i+1}/x_1 can be directly picked off the i+1 planes and transferred to the corresponding points of the preceding i-th planes. Having completed all of the slopes for the incremental advance Δx on the first n-1 planes, we resort to the basic equation A-1 in order to determine the new x_n, knowing the new x_{n-1} through x. This last step may rarely be done geometrically.

The subsequent iteration procedure needed to obtain a complete phase space response is essentially identical to that commonly used for second order phase plane constructions.

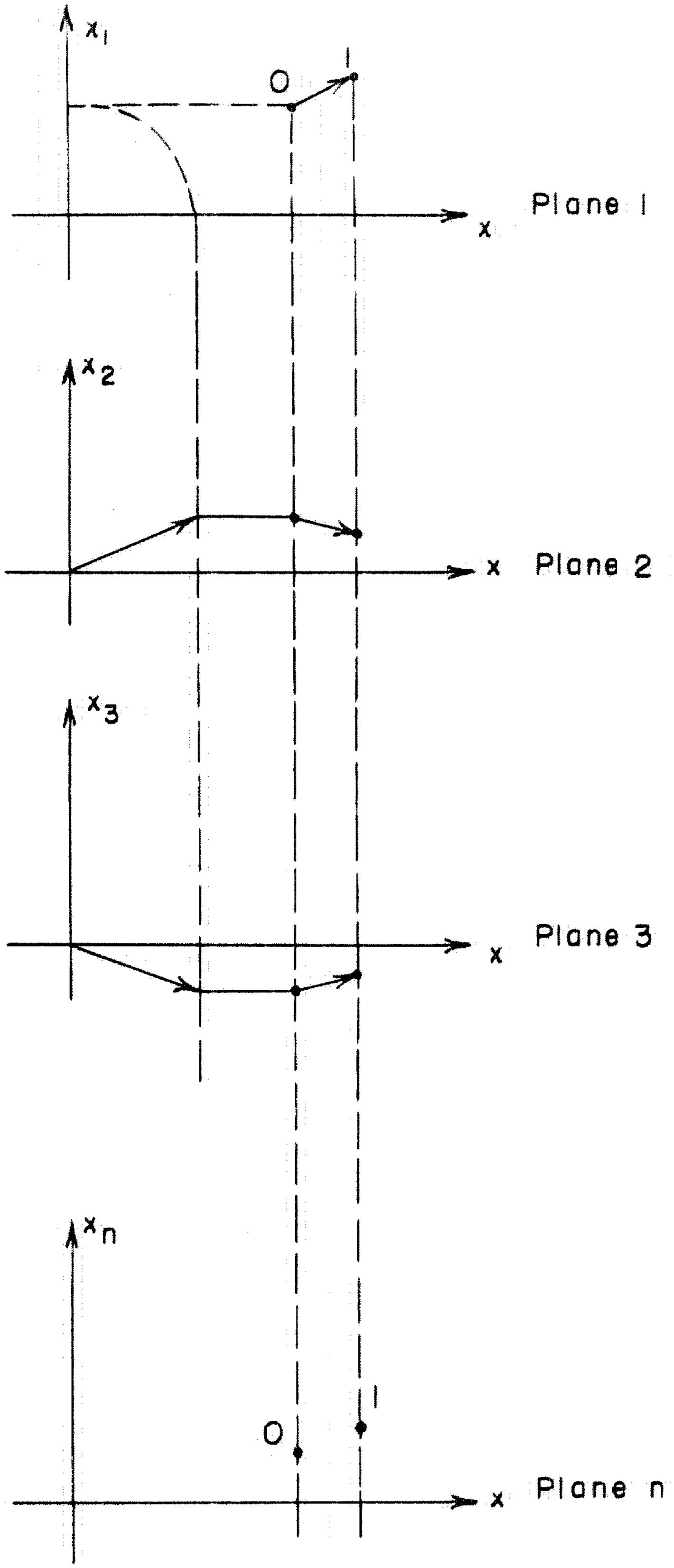


FIG. A-1 PHASE SPACE CONSTRUCTION

APPENDIX B

Details of the Function Generator

A brief discussion of the circuit details of each component of the function generator follows:

1. Input Splitter

The circuit diagram, Figure B-1, is self-explanatory. By a simple cascade of germanium diode voltage dividers, a pass gain of 1.0 for inputs up to 100 volts and a blocking gain of 0.002 are obtained. The input impedance is the back resistance of the diode.

2. Differential Amplifier

As shown in Figure B-2, the differential input is effected by means of common cathode, balanced triodes. Since the required output of the amplifier is always positive, the final stage (the cathode follower) is biased such that its output range is 350 volts maximum to -10 or -20 volts minimum. Its closed loop gain is +6, linear within about 1/2% over an output range of 0 to +300 volts. Having an open loop gain of about 1500, the amplifier is reasonably insensitive to power supply fluctuations. To minimize hum, the first stage employs a d.c. filament supply.

3. Logarithmic Converter

Except for rearrangement of the chassis layout, the unit is that described in Ref. 25. Its circuit diagram is shown in Figure B-3. The nonlinearities at the triode are sensitive to filament, plate, and bias voltage variations; consequently, the complete supply must be well regulated. A small unintentional change in the voltage representing $\log_a x$ when x is, say, 300 (volts) is highly magnified when converted to an inverse logarithm.

The triodes used as nonlinear elements must be selected (estimated rejection about 60-70%) and aged 500 hours. No precise tests have been made concerning the upkeep of a converter. A triode which has been warmed and then has the proper circuit adjustments made holds its conversion reasonably well for 100 hours continuous operation. Intermittent operation is less reliable.

4. Inverse Logarithm Converter

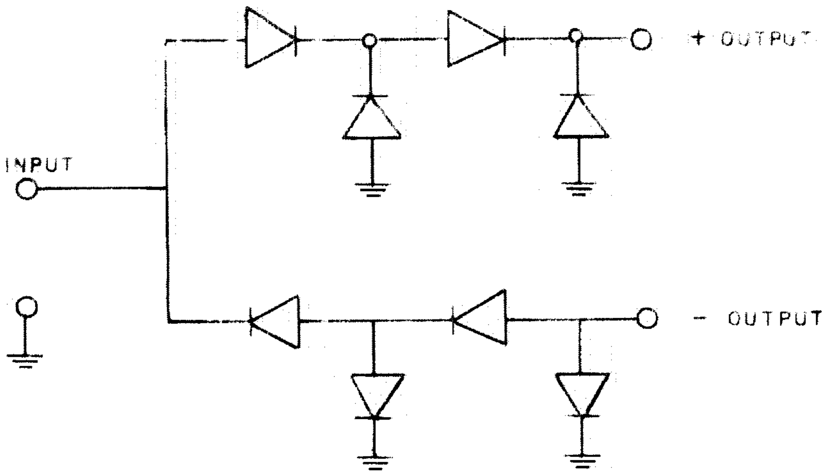
The feedback of this unit (diagram in Fig. B-4) is the logarithm converter just discussed, and all of the features mentioned also apply to this unit. The forward gain of the amplifier is about +1500. Excluding the idiosyncrasies of the feedback, the complete converter is fairly drift free and is easily balanced. Its output range is approximately 0 to +350 volts.

5. Polarity Control

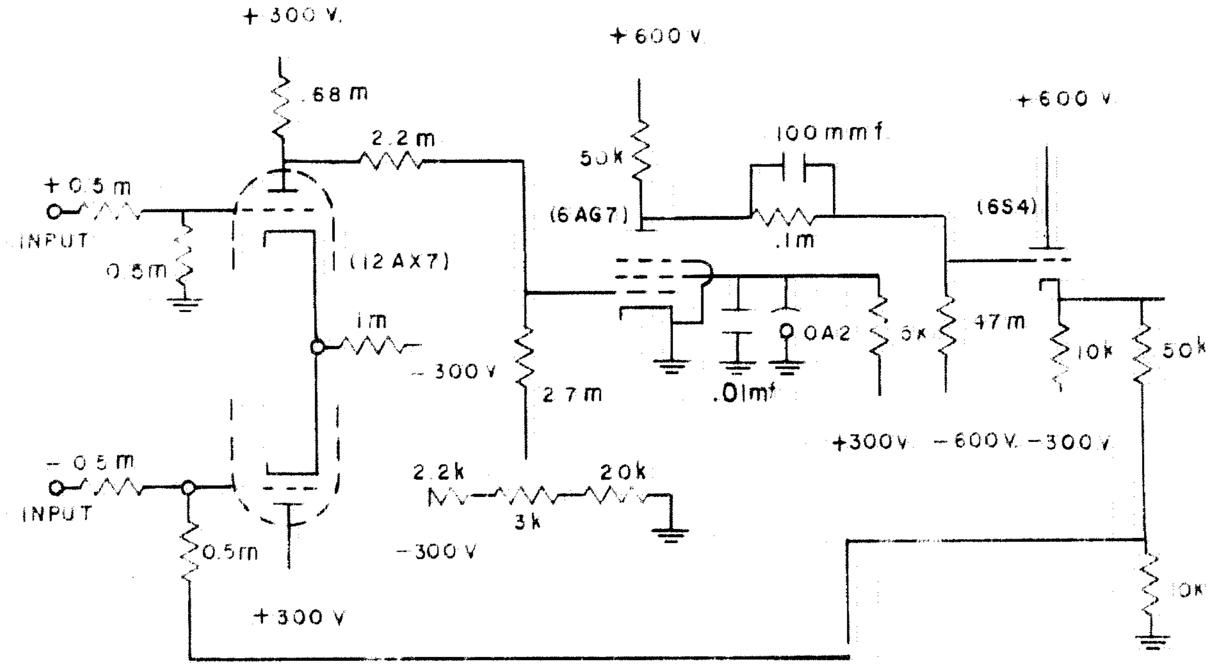
This is a logical circuit which always has a positive input. Its output is linear with its input, positive if the polarities of two control voltages are identical, and negative if the control voltages have opposite polarities. It is shown functionally in Figure B-5. As a matter of fact, the control could utilize relays; however, reliable, inexpensive relays are slow (2 milliseconds or so transfer time). Pulse relays having transfer times as low as 100 microseconds are quite expensive and are delicate. It was therefore decided to design the control as an all-electronic device.

One method of electronic control (Ref. 26) amplifies and limits each of the control voltages. By adding the two limited voltages, a switching voltage results which is ± v_0 if the control inputs are of the same polarity and which is 0 if the inputs differ in polarity. Satisfactory operation demands quite precise limiting, and elements working at various voltage levels. As a result , the unit has several required balancing operations and is subject to drift.

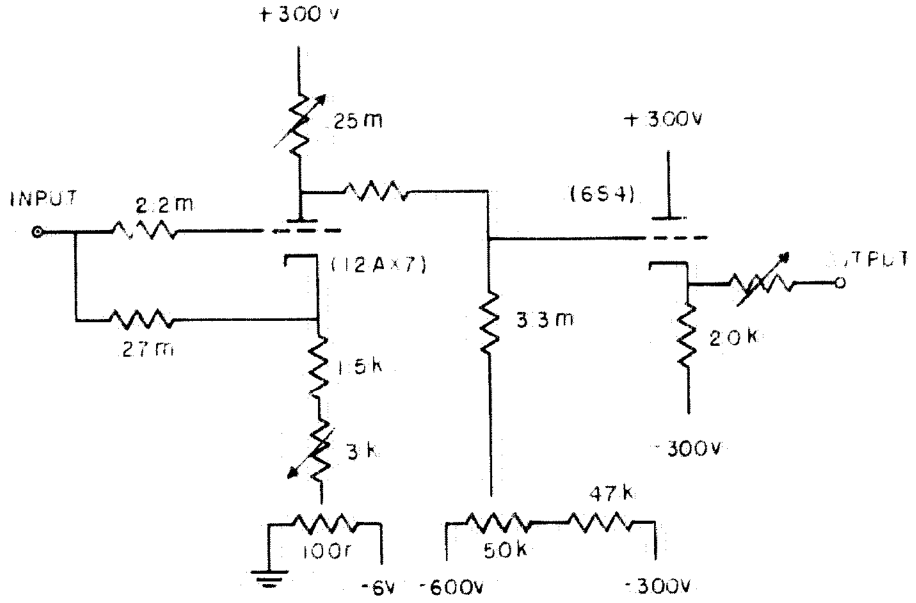
The approach used here directly treated the logical diagram, replacing relay switches with electronic switches. The most reliable electronic switching of the versions tested was by use of triodes clamping to ground. The clamp was quite simple; a single or double section of a triode such as a 12AX7 had its cathode grounded and its place tied to the input through a resistance of about 1/2 megohm. With a grid voltage of +4 volts (using paralleled sections of the twin triode), the tube clamped voltages up to 300 volts producing less than 1/2 volt output. With a grid voltage of -2 volts, the tube blocked current, allowing full output voltage. The complete circuit diagram utilizing this triode clamp is shown in Figure B-6. Simple amplifiers boost the control voltages such that these may be as small as ± 1/4 volts and effect polarity control. Isolation cathode followers drive the final clamps. The differential amplifier controlling polarity of the signal is a simplified version of the amplifier discussed in Section 2. The unit can switch an input voltage of +300 volts cleanly at frequencies up to 500 cps.



POLARITY SPLITTER
FIGURE B-1

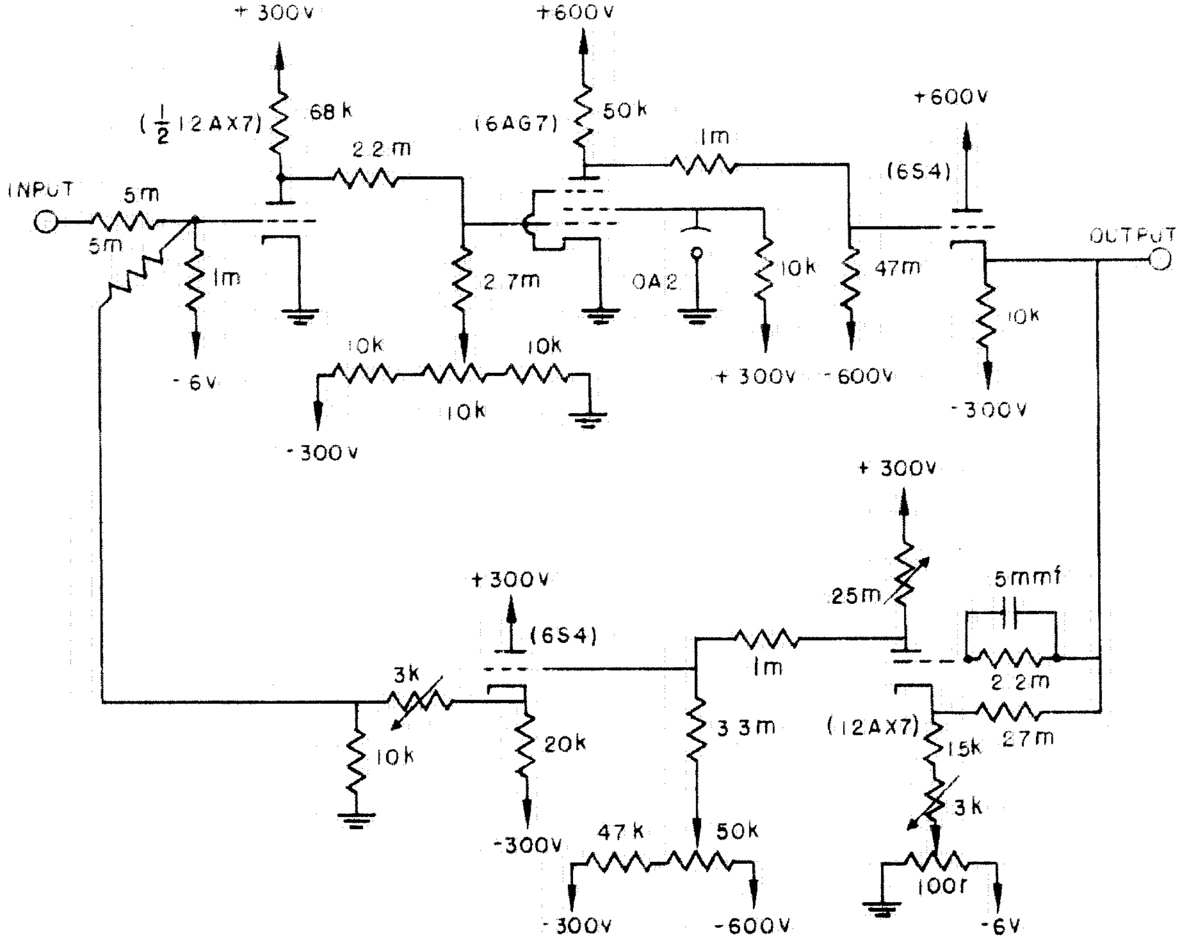


DIFFERENTIAL AMPLIFIER
FIGURE B-2



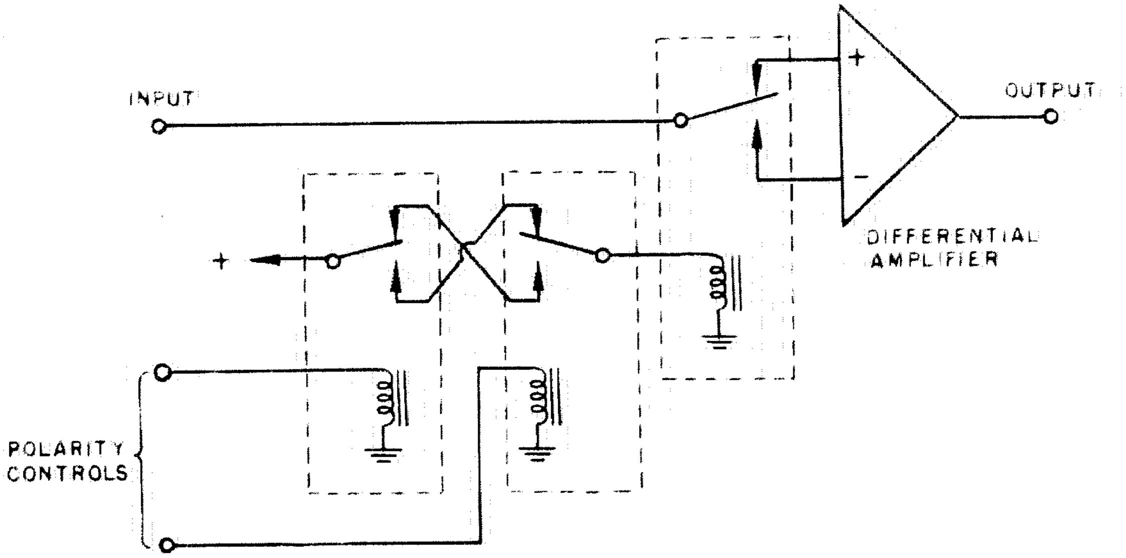
LOGARITHM CONVERTER

FIGURE B-3



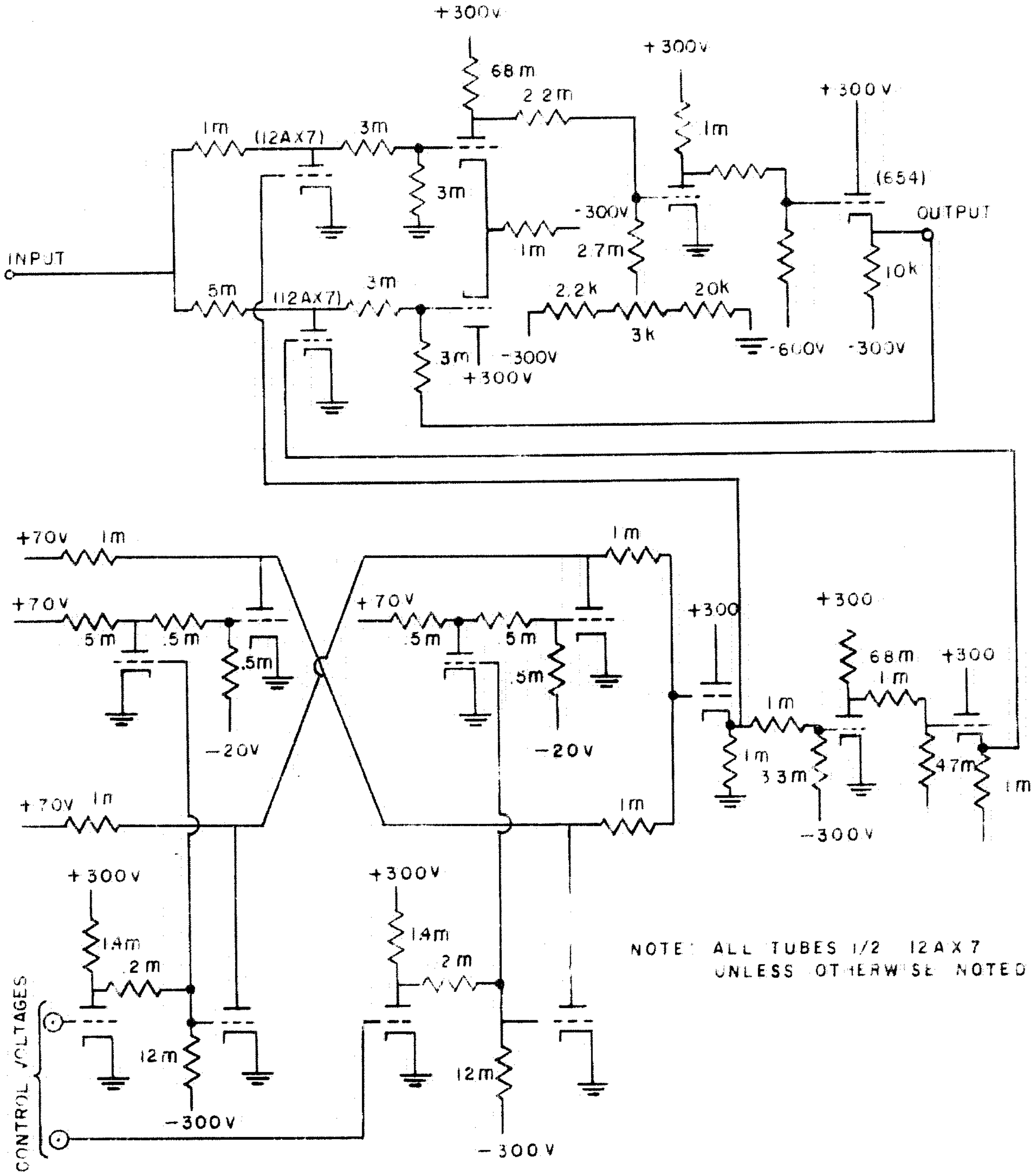
INVERSE LOGARITHM CONVERTER

FIGURE B-4



NOTE: RELAY CONTACTS SHOWN DE-ENERGIZED. RELAY ENERGIZES WHEN POSITIVE VOLTAGE APPLIED.

LOGICAL DIAGRAM OF POLARITY CONTROL
FIGURE — B5



NOTE: ALL TUBES 1/2 12AX7
UNLESS OTHERWISE NOTED

POLARITY CONTROL
FIGURE — B 6